



UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS MÉTODOS DE APRENDIZAJE PROFUNDO PARA LA EXTRACCIÓN DE GRAFOS A PARTIR DE IMÁGENES

Rafael Nicolás Gaggion Zulpo

Tesis remitida al Comité Académico del Doctorado como
parte de los requisitos para la obtención del grado de
DOCTOR EN INGENIERÍA
Mención en Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2023

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000,
Santa Fe, Argentina.



UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS MÉTODOS DE APRENDIZAJE PROFUNDO PARA LA EXTRACCIÓN DE GRAFOS A PARTIR DE IMÁGENES

Rafael Nicolás Gaggion Zulpo

Lugar de Trabajo:

$\text{sinc}(i)$

Instituto de Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:

Dr. Enzo Ferrante

$\text{sinc}(i)$, CONICET-UNL

Co-director:

Dr. Federico Ariel

IAL, CONICET-UNL

(1994-
2024)

30 años de la
Consagración Constitucional
de la Autonomía y Autarquía
Universitaria en Argentina.



ACTA DE EVALUACIÓN DE TESIS DE DOCTORADO

En la sede de la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, a los catorce días del mes de febrero del año dos mil veinticuatro, se reúnen en forma híbrida los miembros del Jurado designado para la evaluación de la Tesis de Doctorado en Ingeniería titulada “*Nuevos métodos de aprendizaje profundo para la extracción de grafos a partir de imágenes*”, desarrollada por el Ing. Rafael Nicolás GAGGIÓN ZULPO, DNI N° 38.570.067. Ellos son: Dr. Pablo Granitto, Dr. Martín Belzunce, y Dr. Federico Lecumberry,--
La Presentación oral y defensa de la Tesis se efectúan bajo la modalidad híbrida.

Luego de escuchar la Defensa Pública y de evaluar la Tesis, el Jurado considera:

Que la tesis incluye una cantidad considerable de aportes al estado del arte y de conocimientos innovadores en el campo del procesamiento de imágenes, de aprendizaje profundo y sus aplicaciones.

El documento de la tesis está muy bien organizado y redactado. La defensa fue sólida mostrando un gran dominio del área y respondiendo las preguntas del jurado con solvencia.

Por lo tanto, el Jurado resuelve aprobar la tesis con calificación 10 (diez) Sobresaliente, felicitando al tesista y sus directores por el trabajo realizado.

Sin más, se da por finalizado el Acto Académico con la firma de los miembros del Jurado al pie de la presente. -----

Dr. Pablo Granitto

Dr. Martín Belzunce

Dr. Federico Lecumberry



Dependencia

Dirección postal

Código postal, Ciudad, Argentina

Teléfono

Correo electrónico

(1994-
2024)

30 años de la
Consagración Constitucional
de la Autonomía y Autarquía
Universitaria en Argentina.



UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

• Santa Fe, 14 de febrero de 2024.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada *“Nuevos métodos de aprendizaje profundo para la extracción de grafos a partir de imágenes”*, desarrollada por el Ing. Rafael Nicolás GAGGIÓN ZULPO, en el marco de la Mención “Inteligencia Computacional, Señales y Sistemas”, certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de la copia digital de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Pablo Granitto

Dr. Martín Belzunce

Dr. Federico Lecumberry

Santa Fe, 14 de febrero de 2024.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención “Inteligencia Computacional, Señales y Sistemas” y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

.....
Dr. Federico Ariel
Codirector de Tesis

.....
Dr. Enzo Ferrante
Director de Tesis



Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Dependencia

Dirección postal

Código postal, Ciudad, Argentina

Teléfono

Correo electrónico

TESIS POR COMPILACIÓN

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución No 255/17 (Expte. No 888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

“En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión.”

AGRADECIMIENTOS

A mis directores, el Dr. Enzo Ferrante y el Dr. Federico Ariel, les agradezco por la confianza que depositaron en mí y por brindarme la oportunidad de formar parte de sus proyectos. Al Dr. Diego Milone, quien aunque no figura en los documentos como co-director, desempeñó un papel fundamental en este recorrido. A los tres, por guiarme en esta etapa de aprendizaje para formarme como mejor profesional y mejor persona.

A mis padres, hermanos, mis abuelos, quienes están y quienes siempre estuvieron, por su compañía y amor incondicional.

A las siguientes instituciones:

- sinc(i): Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional.
- Facultad de Ciencias Hídricas de la Universidad Nacional del Litoral.
- CONICET: Consejo Nacional de Investigaciones Científicas y Técnicas.

Por último, remarco particularmente mi agradecimiento a la universidad pública, que me permitió estudiar, y a la ciencia pública, que me llevó por el camino de hacer un doctorado.

Rafael Nicolás Gaggion Zulpo
Santa Fe, Febrero de 2024.

Índice general

1. Introducción	1
1.1. Descripción del problema	1
1.1.1. Método desacoplado para la extracción de grafos a partir de series temporales de imágenes de raíces de plantas	2
1.1.2. Método unificado para la segmentación anatómica basada en landmarks de imágenes de rayos X	3
1.1.3. Método unificado para la extracción de mallas de corazón a partir de imágenes de resonancia magnética cardíaca	4
1.1.4. Generación de anotaciones para grandes bases de datos a partir de datos multi-céntricos y métodos de validación automática de la calidad	4
1.2. Objetivos	5
1.2.1. Objetivos particulares	5
1.3. Organización de la tesis	5
2. Método desacoplado	7
2.1. Antecedentes	7
2.2. Métodos propuestos	8
2.2.1. Descripción del problema	8
2.2.2. Segmentación automática de raíces de plantas mediante redes convolucionales profundas	9
2.2.3. Generación del grafo a partir de secuencias de segmentaciones binarias	9
2.3. Experimentos y resultados	11
2.3.1. Base de datos	11
2.3.2. Resultados	11
3. Método unificado 2D	15
3.1. Antecedentes	15
3.2. Métodos propuestos	16
3.2.1. Descripción del problema	16
3.2.2. Convoluciones espectrales en grafos	16
3.2.3. Autocodificadores	16
3.2.4. Modelos híbridos	17
3.2.5. Aprendiendo en múltiples resoluciones	17
3.2.6. Conexiones de salto entre grafos e imágenes	18
3.2.7. Modelo HybridGNet completo	18
3.3. Experimentos y Resultados	19
3.3.1. Base de Datos	19
3.3.2. Modelos para comparación	19
3.3.3. Evaluando la segmentación anatómica	19
3.3.4. Evaluación en contexto de cambio de dominio	19
3.3.5. Generación de nuevas bases de datos mediante HybridGNet	20
3.3.6. Evaluación de la robustez ante oclusiones	21

4. Método unificado 3D	23
4.1. Antecedentes	23
4.2. Métodos propuestos	24
4.2.1. Descripción del problema	24
4.2.2. Evaluando diferentes términos de regularización	25
4.3. Experimentos y resultados	26
4.3.1. Base de datos	26
4.3.2. Modelo para comparación	26
4.3.3. Resultados en mallas de superficie	26
4.3.4. Resultados en mallas volumétricas	28
5. Grandes bases de datos multi-céntricas	31
5.1. Antecedentes	31
5.2. Métodos propuestos	32
5.2.1. Descripción del problema	32
5.2.2. Entrenando modelos basados en landmark con datos heterogéneos	33
5.2.3. Evaluación automática de la calidad mediante <i>reverse classification accuracy</i>	33
5.3. Experimentos y Resultados	34
5.3.1. Bases de Datos	34
5.3.2. Entrenando modelos con datos heterogéneos	35
5.3.3. Validación del método de evaluación de calidad mediante <i>reverse classification accuracy</i>	37
5.3.4. Segmentación masiva de rayos X utilizando HybridGNet	37
6. Conclusiones	41
7. Publicaciones	43
Apéndices	53
A. Contribuciones	55
B. ChronoRoot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture	56
C. Hybrid graph convolutional neural networks for landmark-based anatomical segmentation	90
D. Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis	104
E. A multi-view hybrid graph convolutional architecture for Volume-to-Mesh extraction in cardiovascular MR images	128
F. Multi-center anatomical segmentation with heterogeneous labels via landmark-based models	146
G. CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images	156

Índice de tablas

2.1. Evaluación cuantitativa de modelos de segmentación en raíces de plantas	12
3.1. Resultados de la segmentación anatómica basada en landmarks para el conjunto de datos JSRT.	19
3.2. Resultados de cambio de dominio para la segmentación anatómica basada en landmarks	20
3.3. Resultados de la generación de anotaciones de landmark a partir de segmentaciones densas	20
4.1. Resultados cuantitativos sobre de segmentación en mallas de superficie.	27
4.2. Resultados cuantitativos de segmentación en mallas volumétricas.	28
4.3. Resultados cuantitativos del análisis de calidad de los tetraedros en mallas volumétricas. .	28
5.1. Resultados de la validación manual clínica, con desvíos estandar	38
5.2. Análisis estadístico del DSC estimado por RCA en las distintas bases de datos segmentadas	39
5.3. Estimación a nivel dataset obtenida por el reentrenamiento de la HybridGNet en los nuevos datos generados	39

Índice de figuras

1.1. Diagrama ilustrativo	3
2.1. Pipeline para la extracción de grafos a partir de imágenes temporales	8
2.2. Salida del modelo y ejemplos de parámetros fenotípicos extraíbles.	10
2.3. Parámetros fenotípicos observados a través del tiempo	13
3.1. Modelo Híbrido	17
3.2. Arquitectura HybridGNet Completa	18
3.3. Estudio de oclusiones artificiales.	21
3.4. Estudio de oclusiones reales.	22
4.1. Arquitectura del modelo propuesto para la extracción de mallas cardíacas a partir de imágenes de resonancia magnética cardíaca.	24
4.2. Análisis del suavizado laplaciano en mallas de superficie	27
5.1. Bases de datos multi-céntricas con anotaciones heterogéneas	32
5.2. Metodología propuesta para la evaluación automática de calidad mediante RCA.	34
5.3. Evaluación cualitativa	35
5.4. Inspección del espacio latente mediante UMAP	36
5.5. Correlación de los valores de RCA contra los valores reales de DSC	37
5.6. Histograma de DSC estimado por RCA para la base de datos CheXmask completa	38

Resumen

En el ámbito de la informática, los grafos desempeñan un papel fundamental al representar problemas que exhiben una estructura subyacente en forma de red. Con frecuencia, surge la necesidad de extraer estos grafos a partir de imágenes para obtener representaciones precisas de dicha estructura. Estas representaciones se convierten en componentes esenciales para abordar problemas específicos que exigen una extracción fiable de estas estructuras, y este desafío se extiende a una amplia variedad de dominios en el procesamiento de imágenes.

La presente tesis aborda el desafío de extraer grafos a partir de imágenes en diversas aplicaciones, incluyendo la representación de la arquitectura de raíces de plantas, la segmentación anatómica en imágenes de rayos X y la extracción de mallas cardíacas en imágenes de resonancia magnética. Para alcanzar este propósito, se proponen distintos enfoques basados en redes neuronales convolucionales con el objetivo de desarrollar métodos precisos y eficientes que permitan la extracción automática de grafos desde imágenes.

Uno de los enfoques presentados se dirige a la representación detallada de la arquitectura de las raíces de plantas. Se propone un método que se enfoca en la dinámica temporal de las raíces de *Arabidopsis thaliana*, una planta modelo ampliamente utilizada. Este método utiliza redes neuronales convolucionales para realizar una segmentación inicial de las imágenes y luego aplica técnicas de esqueletización y extracción de grafos para permitir un seguimiento preciso de la evolución de las raíces a lo largo del tiempo. Esto no solo facilita la observación de cambios y crecimiento en las raíces, sino que también permite la extracción de nuevos parámetros fenotípicos previamente inobservables.

En segundo lugar, se presenta un enfoque unificado que combina redes neuronales convolucionales con redes neuronales convolucionales en grafos. Este modelo permite obtener grafos directamente a partir de imágenes y entrenarlos de manera integral. Se aplica con éxito a la segmentación anatómica en imágenes de rayos X, donde la representación basada en grafos mejora la plausibilidad anatómica en la segmentación al incorporar de manera directa información topológica mediante la matriz de adyacencia del grafo. Este enfoque unificado luego se extiende para abordar la extracción de mallas cardíacas a partir de imágenes de resonancia magnética, permitiendo la extracción directa de mallas a partir de imágenes, lo que es esencial para aplicaciones de simulación. Además, proporciona la capacidad de imponer criterios de regularización para garantizar la suavidad de las superficies o la calidad de los elementos en las mallas, lo que las hace adecuadas para su uso en simulaciones numéricas.

Por último, se aborda la generación automática de bases de datos de anotaciones en grafos con el objetivo de aumentar la disponibilidad de conjuntos de datos anotados. Aplicado en un contexto de segmentación anatómica en imágenes de rayos X, se trabaja en base a datos multi-céntricos con anotaciones heterogéneas, demostrando la capacidad de los modelos propuestos para lidiar con estos problemas. Este proceso se complementa con el desarrollo de métodos para la evaluación automática de la calidad, con el objetivo de obtener una métrica que permita evaluar cuantitativamente la confiabilidad de cada una de las anotaciones generadas.

Sección 1

Introducción

1.1. Descripción del problema

Los grafos, como estructuras de datos, desempeñan un papel esencial en el ámbito de la informática al proporcionar una forma poderosa de modelar problemas que exhiben una estructura subyacente, generalmente en forma de red. En numerosas ocasiones, se hace necesaria la extracción de estos grafos a partir de imágenes con el propósito de obtener una representación precisa de dicha estructura subyacente. Esta representación se convierte en un componente fundamental para afrontar problemas específicos que demandan un enfoque centrado en la extracción de tales estructuras de manera directa y confiable, abarcando esto una amplia variedad de dominios de aplicación en el contexto del procesamiento de imágenes.

La tarea de extraer grafos a partir de imágenes bidimensionales (2D) es un problema que encuentra aplicación en una gran variedad de dominios, como pueden ser la extracción de calles a partir de imágenes satelitales [92] para generar mapas, representación de estructuras vasculares a partir de imágenes de fondo de ojo [68], obtención de líneas centrales correspondientes a la vasculatura en angiografías [89, 96] o el modelado de raíces de plantas en imágenes biológicas [76, 12]. En situaciones en las que la imagen es un volumen tridimensional (3D), la extracción de grafos se convierte en una herramienta poderosa para representar estructuras aún más complejas, como mallas de superficies o volumétricas. Esto se puede apreciar en aplicaciones como la segmentación de órganos en imágenes de resonancia magnética cardíaca [97], o en la segmentación de la corteza cerebral en el caso de resonancias magnéticas del cerebro [38].

Comúnmente, la extracción de estructuras en imágenes es abordada como un problema de segmentación, donde uno busca reconocer que píxeles corresponden a cada estructura subyacente dentro del espacio de la imagen. Sin embargo, es importante resaltar que a diferencia del problema clásico de segmentación de imágenes, la generación de un grafo implica necesariamente una transformación en la estructura de los datos de salida con respecto a la imagen de entrada. En la segmentación convencional el resultado consiste en una imagen de la misma resolución que la original, en la cual cada píxel (o vóxel en el caso 3D) se etiqueta con un valor entero correspondiente a alguna de las clases de interés. Sin embargo, en la extracción de un grafo a partir de una imagen, la salida se configura como una estructura de datos completamente diferente, compuesta por nodos, aristas y, posiblemente, funciones definidas en ambos.

En los últimos años, las redes neuronales profundas han experimentado una auténtica revolución en el campo de la visión computacional, el análisis de imágenes, el reconocimiento de patrones y, de particular interés para este trabajo, en las tareas de segmentación. Estos modelos son capaces de aprender representaciones de los datos de entrenamiento en múltiples niveles de abstracción, mediante la composición de módulos simples (pero generalmente no lineales), que transforman progresivamente estas representaciones en otras de mayor nivel de abstracción [53]. Un tipo específico de redes neuronales, conocido como redes neuronales convolucionales (CNN), desempeñará un papel central en esta tesis debido a su capacidad innata para el análisis de imágenes.

Dado que el problema central abordado en esta tesis doctoral se refiere a la extracción de grafos

a partir de imágenes, se explorarán distintos enfoques para la extracción de grafos que puedan hacer uso de las CNN, utilizadas en el dominio de las imágenes. Para ello, se plantea la idea de combinar las CNN con redes convoluciones definidas en estructuras de grafos. Si bien las CNN son altamente eficaces en tareas relacionadas con imágenes o señales de audio, su aplicabilidad directa se ve limitada cuando se trata de problemas que involucran señales muestreadas de manera irregular o definidas en dominios con una estructura geométrica específica, como las redes o los propios grafos. Este desafío ha motivado el surgimiento de un subcampo en el aprendizaje profundo conocido como aprendizaje profundo geométrico (o *geometric deep learning* en inglés) [10].

Esta tesis tiene como objetivo principal proponer nuevos métodos para la extracción automática de grafos a partir de imágenes, considerando tanto enfoques clásicos basados en técnicas de procesamiento de imágenes, como enfoques híbridos que combinen distintos tipos de redes neuronales, con el propósito final de desarrollar modelos unificados capaces de realizar dicha extracción de manera automática y que puedan ser entrenados de punta a punta. Esto permitirá superar las limitaciones tradicionales de las CNN en situaciones con estructuras geométricas específicas, como las que se encuentran en los grafos. Como se puede observar en la Figura 1.1, la aplicación de este enfoque se centrará en tres tareas objetivo particulares: la extracción de grafos a partir de secuencias de imágenes 2D de raíces de plantas en crecimiento mediante un enfoque desacoplado (con etapas independientes de segmentación y construcción del grafo), la segmentación basada en *landmarks*, o puntos de interés, a partir de imágenes 2D de rayos X mediante un enfoque unificado, y la extracción de mallas de corazón en imágenes volumétricas de resonancia magnética cardíaca, con el mismo enfoque. Por último, y dada la escasez de bases de datos disponibles con anotaciones en grafo, se abordará la tarea de generar grandes bases de datos de manera automática utilizando los enfoques propuestos, y diseñar métodos para el control de calidad automático de dichas anotaciones en situaciones donde no hay anotaciones disponibles.

Se procede entonces a explicar el contexto de cada aplicación, y la motivación para resolver los problemas utilizando métodos de extracción de grafos.

1.1.1. Método desacoplado para la extracción de grafos a partir de series temporales de imágenes de raíces de plantas

En el ámbito de la investigación en biología vegetal, comprender cómo crecen las raíces de las plantas es una tarea fundamental. Las plantas, debido a su inmovilidad, no pueden desplazarse en busca de las condiciones ambientales óptimas para su desarrollo y supervivencia. Sin embargo, poseen una sorprendente capacidad de adaptación que les permite completar su ciclo de vida incluso en entornos en constante cambio [71]. Esta adaptación se refleja en la arquitectura de sus sistemas de raíces, que varían en respuesta a factores como la disponibilidad de nutrientes y agua en el suelo, la presencia de obstáculos y otros estímulos ambientales.

Para comprender esta adaptación y sus implicaciones genéticas, es crucial caracterizar la arquitectura del sistema de raíces (RSA, por las siglas del inglés *root system architecture*) de las plantas en crecimiento. Esto involucra el fenotipado (es decir, la extracción de parámetros de interés) de una planta a lo largo del tiempo y depende de variables clave, como la longitud de la raíz principal (MR o *main root*) o la densidad y longitud de las raíces laterales (LRs o *lateral roots*) [40]. Aunque existen herramientas semiautomáticas que asisten en el fenotipado de las raíces en momentos específicos [63, 98], el seguimiento temporal suele estar limitado por restricciones tecnológicas. A menudo, se pasan por alto parámetros fenotípicos potencialmente valiosos que podrían estar relacionados con la dinámica temporal del crecimiento de las raíces.

En este contexto, colaborando con investigadores del Institute of Plant Sciences Paris-Saclay (IPS2), se planteó el desafío de desarrollar una estrategia para abordar este problema de manera más efectiva. Utilizamos un sistema de bajo costo desarrollado en el IPS2 que permite capturar videos de raíces de *Arabidopsis thaliana* en crecimiento en condiciones controladas. Proponemos una estrategia desacoplada de cuatro pasos que, mediante el uso de un grafo para la representación de la RSA, nos permita obtener las características fenotípicas deseadas: una segmentación inicial obtenida mediante el uso de CNNs, una etapa de esqueletización, un paso de extracción del grafo que represente la arquitectura de las raíces y, por primera vez, un seguimiento temporal que nos permitirá obtener las características

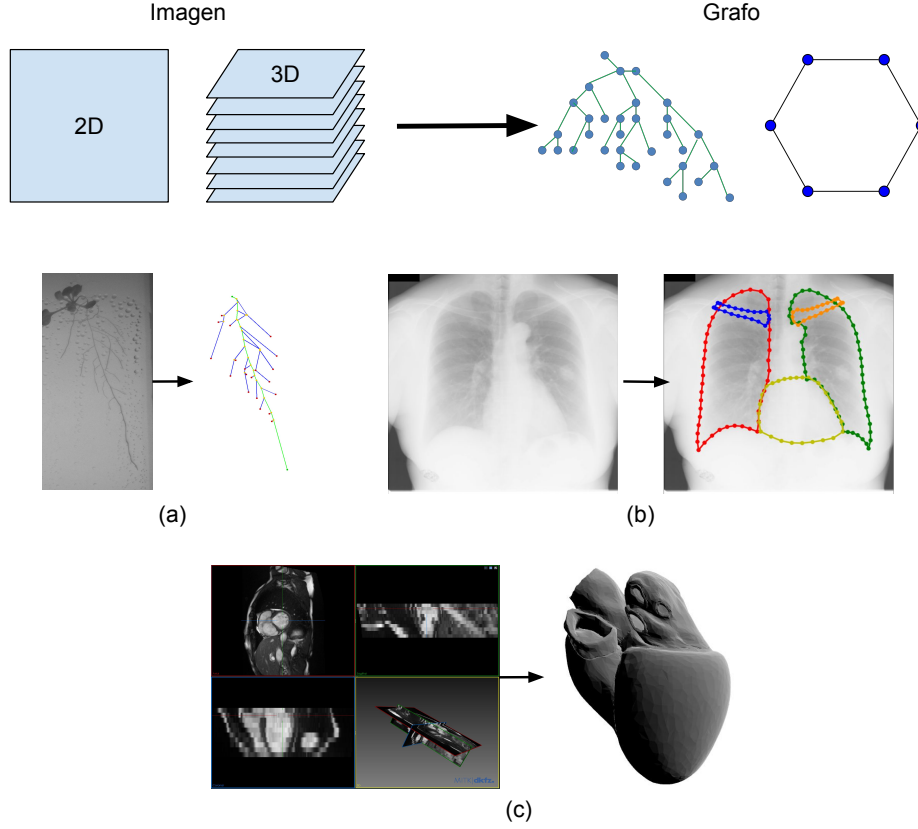


Figura 1.1: **Diagrama ilustrativo.** (a) Extracción de la RSA a partir de imágenes de raíces de plantas. (b) Segmentación basada en landmarks en imágenes de rayos-x. (c) Extracción de mallas de superficie a partir de imágenes de resonancia magnética cardíaca.

fenotípicas de la RSA a lo largo del tiempo en condiciones controladas.

1.1.2. Método unificado para la segmentación anatómica basada en landmarks de imágenes de rayos X

En el contexto de la segmentación de imágenes médicas, las CNN han demostrado un rendimiento sobresaliente. En general, el enfoque más usado es emplear arquitecturas estándar de CNN de tipo codificador-decodificador, basadas en modelos U-Net [82], que predicen la segmentación deseada a nivel de píxeles mediante el aprendizaje de características jerárquicas a partir de conjuntos de datos anotados. En situaciones donde la topología y la ubicación tienden a no conservarse entre individuos, como la segmentación de lesiones, es deseable abordar la segmentación de imágenes como un problema de etiquetado de píxeles. Sin embargo, los órganos y las estructuras anatómicas suelen presentar una topología característica que tiende a ser regular.

En este contexto, es fundamental señalar que las arquitecturas de segmentación, que se centran en asignar píxeles de manera precisa a categorías específicas, pueden no garantizar una segmentación anatómica coherente desde una perspectiva topológica. Esto ocurre debido a la limitada capacidad de las métricas utilizadas en las funciones de pérdida a nivel de píxeles (empleadas para entrenar estos modelos), como la entropía cruzada o el índice de coincidencia Soft-Dice [59], para captar la forma y la topología global de las estructuras. Como resultado, es común observar artefactos en las segmentaciones, como fragmentación de estructuras, inconsistencias topológicas e islas de píxeles [8], especialmente en entornos clínicos desafiantes que involucran oclusiones en imágenes y variaciones de

dominio entre diferentes centros médicos.

Como alternativa a las estrategias tradicionales de segmentación a nivel de píxeles, la representación basada en grafos ofrece una forma natural de incorporar restricciones topológicas en el proceso de segmentación. Esto es especialmente valioso para establecer correspondencias de puntos de referencia entre individuos, una consideración crucial en el análisis estadístico de formas. Con esta representación, los grafos se convierten en una herramienta natural para describir puntos de referencia, contornos y superficies, permitiendo codificar la estructura anatómica a través de su matriz de adyacencia y, así, limitar el espacio de soluciones, promoviendo correcciones topológicas y aumentando la coherencia anatómica.

Para abordar estas limitaciones y mejorar la representación de las estructuras anatómicas en imágenes de rayos X, se plantea un enfoque basado en grafos de contorno. Este enfoque se distingue al considerar los contornos de los órganos y estructuras como grafos con una estructura de conectividad predefinida, en la cual cada máscara de segmentación se representa mediante un número constante de nodos, compartiendo todas las máscaras la misma matriz de adyacencia. En esta tesis, se explorará en detalle cómo aprovechar esta representación basada en grafos para mejorar la segmentación anatómica en imágenes de rayos X. Para lograrlo, combinaremos un codificador CNN, eficaz para extraer características de las imágenes, con un decodificador basado en convoluciones espectrales en grafo (GCNN, del inglés *graph convolutional, neural networks*). Esta combinación nos permitirá extraer directamente los contornos de las estructuras a partir de la imagen, superando las limitaciones de las estrategias convencionales.

1.1.3. Método unificado para la extracción de mallas de corazón a partir de imágenes de resonancia magnética cardíaca

La imagen cardíaca ha experimentado avances significativos en los últimos años, convirtiéndose en una herramienta indispensable en el diagnóstico, la planificación del tratamiento y la gestión de diversas enfermedades cardiovasculares. Uno de los componentes clave para estos avances es la extracción de mallas tridimensionales precisas a partir de imágenes de resonancia magnética cardíaca (CMR, del inglés *cardiovascular magnetic resonance*), que sirven como base para simulaciones computacionales [24], el descubrimiento de biomarcadores [9] y el análisis de la deformación y la dinámica del corazón.

Sin embargo, la extracción de estas mallas cardíacas presenta desafíos significativos. Los métodos tradicionales, como los modelos de forma activa (*active shape models* en inglés) [67] o la segmentación multi-atlas [5], a menudo requieren importantes recursos computacionales y pueden llevar mucho tiempo. La variabilidad en las formas, tamaños y patologías cardíacas complica aún más el proceso de extracción, lo que exige un método que sea robusto y adaptable al mismo tiempo.

Además, la transición de cortes de imágenes en 2D a una representación cohesiva en 3D exige una integración fluida de la información espacial, especialmente al modelar mallas volumétricas. Las metodologías actuales a menudo requieren pasos de pos-procesamiento complicados para refinar las mallas [24, 64], lo que las hace adecuadas para simulaciones pero puede introducir fuentes adicionales de error y prolongar el tiempo de procesamiento en general.

Ante estos desafíos, existe una necesidad de tener enfoques innovadores que puedan agilizar el proceso de extracción de mallas cardíacas, mejorar la precisión y reducir el tiempo y la carga computacional. En esta tesis se propondrán entonces modelos híbridos convolucionales en imágenes 3D, combinados con convoluciones en grafos, que permitirán modelar mallas a partir de imágenes de manera directa, tanto en el caso de mallas de superficie como volumétricas, representado volúmenes mediante el uso de elementos tetraédricos. Esta investigación busca abordar las limitaciones mencionadas y avanzar en la extracción de mallas cardíacas de manera más eficiente y precisa.

1.1.4. Generación de anotaciones para grandes bases de datos a partir de datos multi-céntricos y métodos de validación automática de la calidad

Una de las mayores limitaciones que se encuentran al querer aplicar los métodos propuestos en esta tesis es la falta de bases de datos de gran escala. En el aprendizaje profundo, contar con grandes bases

de datos anotadas es un requerimiento crucial para entrenar modelos robustos y generalizables.

Por ello, exploraremos cómo el modelo unificado propuesto en esta tesis puede utilizarse para obtener anotaciones de manera automática en un contexto de grandes cantidades de imágenes sin anotar. Evaluaremos esto en la tarea de segmentación masiva de imágenes de rayos X, abordando desafíos relacionados con los cambios de dominio y el aprendizaje a partir de datos heterogéneos.

Al construir dichas bases de datos, uno de los requerimientos necesarios es obtener una medida de confianza que nos indique si las anotaciones generadas son correctas. Aquí se propone utilizar un enfoque de validación automática de la calidad de las anotaciones propuestas, basado en el marco teórico denominado Exactitud por Clasificación Inversa (RCA, del inglés *Reverse Classification Accuracy*) [91]. Dentro de este enfoque, encontramos que el proceso de validación automática puede aplicarse también para evaluar sesgos en los modelos, sin la necesidad de tener anotaciones manuales, descubriendo variaciones en el rendimiento de los modelos en subpoblaciones caracterizadas por atributos demográficos específicos, como sexo, edad o grupo étnico [80].

1.2. Objetivos

El objetivo general de este proyecto es contribuir a la construcción de nuevos métodos de aprendizaje profundo para la extracción de grafos a partir de imágenes, desarrollando enfoques que combinen redes neuronales convolucionales clásicas con redes neuronales convolucionales en grafos.

1.2.1. Objetivos particulares

- Desarrollar un método basado en redes neuronales convolucionales clásicas para la segmentación de imágenes de raíces de plantas y posterior extracción de grafos por medio de etapas desacopladas.
- Desarrollar nuevos enfoques unificados para la extracción de grafos a partir de imágenes por medio de un único modelo, entrenable de punta a punta, que combine convoluciones sobre imágenes y convoluciones sobre grafos.
- Validar los métodos propuestos para la extracción de estructuras curvilíneas en imágenes 2D, en la tarea de generación de grafos que representen la estructura de la arquitectura de raíces de plantas a partir de videos con secuencias temporales de imágenes.
- Validar los métodos propuestos para extracción de contornos en el contexto de la segmentación anatómica de órganos de interés en imágenes de rayos X (2D) y de resonancia magnética (3D).
- Explorar la capacidad de los modelos propuestos para la generación automática de anotaciones para grandes bases de datos, aplicando técnicas de evaluación de calidad sobre las anotaciones generadas.

1.3. Organización de la tesis

Esta tesis se encuentra organizada bajo el formato de tesis por compilación de la siguiente forma:

- En esta sección introductoria se proporcionó una visión general de la motivación y el contexto de las aplicaciones abordadas, así como los objetivos establecidos para guiar este trabajo de investigación.
- En la Sección 2, se aborda el método desacoplado clásico, centrado en la extracción de grafos que representan la estructura de las raíces de las plantas a partir de secuencias temporales de imágenes.
- En la Sección 3, se construye el modelo unificado, aplicado a la segmentación basada en landmarks en imágenes de rayos X.

- En la Sección 4, se expande el modelo unificado para abordar la extracción de mallas, tanto superficiales como volumétricas, a partir de imágenes de resonancia magnética cardíaca.
- En la Sección 5, se explora el problema de generar anotaciones para bases de datos extensas en el ámbito de la segmentación basada en landmarks, explorando las ventajas del modelo unificado en esta tarea. Se proponen métodos para realizar control de calidad automático de las segmentaciones, y se investiga el descubrimiento no supervisado de sesgos mediante este proceso de validación.
- Posteriormente, en la Sección 6, se presentan las conclusiones generales derivadas de la tesis.
- Por último, al ser esta una tesis por compilación, se recopilan en el Apéndice las publicaciones científicas que se realizaron a lo largo de esta tesis doctoral.

Sección 2

Método desacoplado: extracción de grafos a partir de imágenes de raíces de plantas

En este capítulo, abordaremos el desarrollo de un método desacoplado para la extracción de grafos que representen la estructura de la arquitectura del sistema de raíces de plantas (RSA, por sus siglas en inglés, *root system architecture*) a partir de secuencias temporales de imágenes. Este enfoque, diseñado para abordar los desafíos específicos de analizar el crecimiento de las raíces de *Arabidopsis thaliana*, se realizó aprovechando un sistema de adquisición innovador basado en una cámara infrarroja que captura imágenes a intervalos regulares.

2.1. Antecedentes

Tradicionalmente, la obtención de imágenes de raíces de *Arabidopsis thaliana*, creciendo en condiciones controladas, implicaba el escaneo manual de una placa de Petri (en la cual las plantas crecen verticalmente sobre un medio semisólido) en un instante de tiempo fijo. Sin embargo, en contraposición a este enfoque, el sistema de adquisición utilizado en esta investigación aprovecha una cámara infrarroja para capturar imágenes a intervalos regulares sin la necesidad de retirar la tapa de la caja de cultivo. Este enfoque innovador presenta desafíos únicos para el análisis de imágenes de raíces, dado que el análisis, habitualmente realizado en un solo instante de tiempo, ahora debe realizarse para cientos o miles de instantes para una sola planta en particular.

Al analizar la literatura y las herramientas de software disponibles, como se observa en Quantitative Plant [55, 78], podemos notar que se han desarrollado más de 40 programas para el análisis de la RSA. Sin embargo, la mayoría de estos programas implican procesos manuales, lo que requiere corrección de resultados o delineación manual por parte de biólogos expertos para obtener resultados precisos, incluso para imágenes obtenidas mediante el escaneo manual de un solo instante de tiempo.

Algunos métodos, como RootTrace [30, 62], emplean técnicas clásicas de procesamiento de imágenes y seguimiento, lo que permite obtener parámetros fenotípicos como la longitud de la raíz o el número de raíces laterales, pero no logran un detalle más profundo que el que buscamos en nuestro método. Otros programas como GiaRoots [31] y EZ Rhizo [1] permiten la segmentación de la raíz, pero utilizan técnicas clásicas de umbralización. Estos enfoques clásicos, incluyendo la segmentación por umbralización, no son adecuados para las imágenes infrarrojas de nuestro método de adquisición, ya que la condensación de gotas de agua producida al no retirar la tapa (como en la forma clásica, mediante el escaneo) puede causar obstrucciones, lo que resulta en cortes y errores en la detección de la raíz. Otros métodos como Win Rhizo [95], una herramienta comercial y propietaria, solo funcionan con imágenes de escáneres ópticos de alta calidad y no permiten la obtención de datos temporales.

En trabajos más recientes y relacionados con nuestro enfoque, se ha demostrado que las técnicas de aprendizaje profundo (deep learning) pueden mejorar la segmentación de raíces de plantas [15, 14].

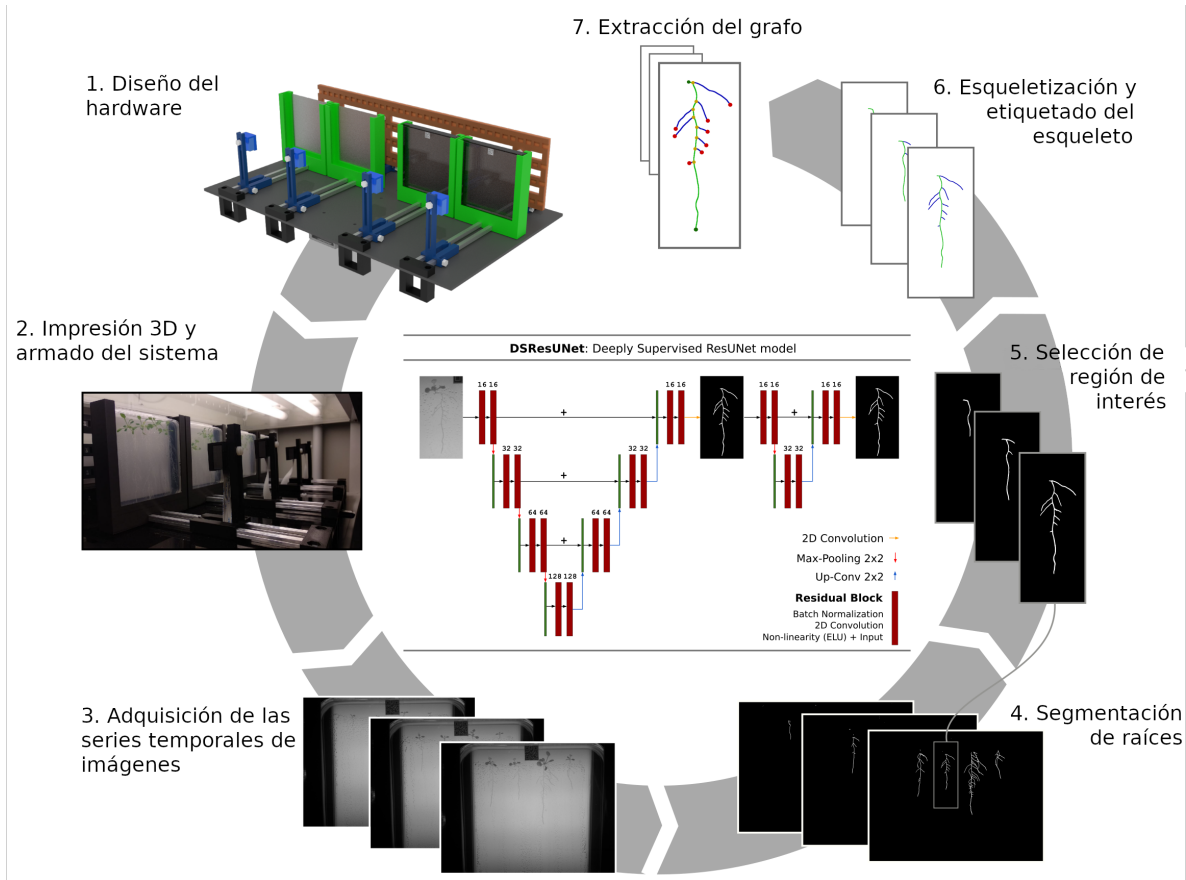


Figura 2.1: **Pipeline para la extracción de grafos a partir de imágenes temporales.** Se representa la serie de pasos completa para la extracción de grafos a partir de series temporales de raíces de *Arabidopsis thaliana*. A partir de las secuencias de imágenes, se procede a segmentar a nivel de píxel las raíces en las placas completas. Luego, a partir de cada raíz tomada de forma individual, se procede con la esqueletización, el etiquetado de las ramas del esqueleto y, finalmente, la construcción del grafo etiquetado correspondiente.

Esto incluye a RootNav 2 [98], que también se basa en modelos de aprendizaje profundo. RootNav 2 permite obtener parámetros fenotípicos detallados, pero sólo para un instante de tiempo específico, sin considerar la dinámica temporal, y su entrenamiento requiere múltiples tipos de anotaciones manuales por parte de biólogos, incluyendo máscaras de segmentación multiclase y la detección de las puntas de las raíces laterales. En cambio, como observaremos, nuestro método solo requiere el uso de máscaras de segmentación binaria para guiar el proceso de segmentación, realizando las siguientes etapas de manera no supervisada.

Para una descripción más detallada de los antecedentes, y las diferencias con nuestra metodología, ver la Sección *Discussion* en el Apéndice B.

2.2. Métodos propuestos

2.2.1. Descripción del problema

Dada una secuencia temporal de imágenes de *Arabidopsis thaliana* en crecimiento, se desea poder observar individualmente los parámetros fenotípicos que representan la RSA para cada planta en cada instante de tiempo. En la Figura 2.1, se presenta una representación simplificada del método propuesto, donde trabajando individualmente con una planta a la vez, se genera un grafo que posibilita

la extracción sencilla de los parámetros fenotípicos deseados en cada instante de tiempo.

Este método se divide en dos fases fundamentales: la obtención de una imagen de segmentación binaria del sistema de arquitectura de raíces y la construcción del grafo que representa la estructura subyacente del árbol en dicho sistema. Dado el carácter dinámico del crecimiento de las plantas, es esencial incorporar información temporal en ambas etapas para obtener con precisión los diversos parámetros fenotípicos.

2.2.2. Segmentación automática de raíces de plantas mediante redes convolucionales profundas

Para abordar la tarea de segmentación de las imágenes de raíces de plantas, se realizaron investigaciones exhaustivas utilizando seis arquitecturas de CNN con el propósito de identificar el enfoque más efectivo. Tres de estas arquitecturas son ampliamente reconocidas en la comunidad de visión por computadora: UNet [82], SegNet [4] y DeepLab v3 [16]. A estas le sumamos la variante de la UNet con bloques residuales [36], la ResUNet, que demostró ser mejor que la UNet para tareas de segmentación de estructuras finas, en el contexto de segmentación de calles en imágenes satelitales [100].

Además de estas arquitecturas de última generación, en esta tesis se proponen dos modelos de CNN adicionales:

- **Deeply Supervised ResUNet:** Como se observa en el centro de la Figura 2.1, este modelo se basa en ResUNet y combina conexiones residuales con supervisión profunda [54], implicando el cálculo de un término de pérdida adicional en una capa intermedia de la red. La salida de ResUNet se concatena con la imagen de entrada original y se procesa a través de dos capas convolucionales adicionales, las cuales actúan como un post-procesamiento automático de la segmentación, resultando en una cascada de dos redes entrenadas conjuntamente.
- **Ensamble [46]:** El enfoque final propuesto consiste en combinar las predicciones de los cinco modelos anteriores, mediante el promedio de las salidas suaves de los modelos segmentación.

Para más detalles de los modelos utilizados y su implementación, ver la Sección *Deep learning models for root segmentation* del Apéndice B.

Incorporación de la dinámica temporal a la segmentación

Una vez elegido un modelo de segmentación, se toma la salida del mismo como segmentación suave, es decir, sin binarizar. Dado que procesamos secuencias temporales de crecimiento de raíces de plantas, proponemos aplicar un paso de postprocesamiento para mejorar la consistencia temporal mediante una variante del promedio ponderado acumulativo. La segmentación actual y la acumulación de las anteriores se promedian para evitar la pérdida de partes de la raíz debido a gotas u otras formas de obstrucción. Dada la segmentación actual s^t en el tiempo t y la máscara acumulada hasta el paso anterior a^{t-1} , calculamos el mapa actual $a^t = s^t + \alpha a^{t-1}$. El valor del peso α se elige según el tamaño del intervalo de tiempo (en nuestros experimentos, utilizamos $\alpha = 0,9$). El objetivo es utilizar las máscaras de segmentación de raíces obtenidas en pasos temporales anteriores para corregir posibles segmentos faltantes de la raíz.

2.2.3. Generación del grafo a partir de secuencias de segmentaciones binarias

La segmentación de las raíces se realiza en la imagen completa, incluso en una placa donde crecen varias plantas simultáneamente. Sin embargo, la generación del grafo se lleva a cabo de manera individual para cada planta, y esto se logra seleccionando una región de interés específica a lo largo del video.

Binarizado y esqueletización: En primer lugar, se binariza la máscara acumulada de segmentación mediante un umbral fijo y, posteriormente, se aplican operaciones morfológicas [33] de apertura y cierre para eliminar píxeles espurios en la máscara binaria. Luego, se identifica la raíz como la componente

conectada más grande dentro de la región de interés. A continuación, se obtiene el esqueleto, que representa la línea central de la segmentación, utilizando un algoritmo clásico de esqueletización [99]. A partir del esqueleto, se procede a generar el grafo.

Construcción del grafo: Para cada punto en el tiempo, el grafo se construye mediante un algoritmo de búsqueda en profundidad. El proceso comienza con un punto de inicio, que puede ser seleccionado manualmente o simplemente la punta superior del esqueleto en la imagen. Dado que el esqueleto tiene un ancho de un píxel, se puede recorrer pixel a píxel hasta encontrar una bifurcación o un punto final. Los nodos se registran como aquellos píxeles que representan bifurcaciones o puntos finales, y las aristas se conectan entre nodos utilizando las secciones del esqueleto como referencia. Además, se registra el largo real en términos de píxeles recorridos entre nodo y nodo como una característica de la arista. A pesar de obtener el grafo que representa la arquitectura del sistema de raíces en un momento dado, este enfoque por sí solo no es suficiente para medir los parámetros de interés, ya que no se sabe a qué secciones de la raíz pertenecen los nodos (si son de la raíz principal o de las raíces laterales).

Etiquetado del grafo: Para abordar este problema, se aprovecha la dinámica temporal del crecimiento. Inicialmente, se asume que solo hay una raíz principal en plantas de *Arabidopsis thaliana*. Esto permite etiquetar los dos primeros nodos en el grafo como el inicio y el final de la raíz principal. A medida que se construye el grafo en momentos temporales posteriores, se realiza un seguimiento basado en el vecino más cercano para propagar las etiquetas de los nodos de manera progresiva. Los nodos recién agregados se etiquetan como bifurcaciones (nodos con más de un vecino) o extremos de raíces laterales (nodos con solo un vecino), lo que permite distinguir la raíz principal del resto de los nodos en el grafo. Este procedimiento también permite identificar en el esqueleto cuál es la raíz principal, cuáles son las raíces laterales y el orden en que aparecen a lo largo del tiempo. Gracias a esta información, no solo se obtiene el grafo etiquetado en cada momento, sino que también se puede generar una máscara de segmentación multiclase, etiquetando las secciones del esqueleto que pertenecen a la raíz principal y las que pertenecen a las raíces laterales, y luego dilatando esta máscara para volver al grosor normal de las raíces.

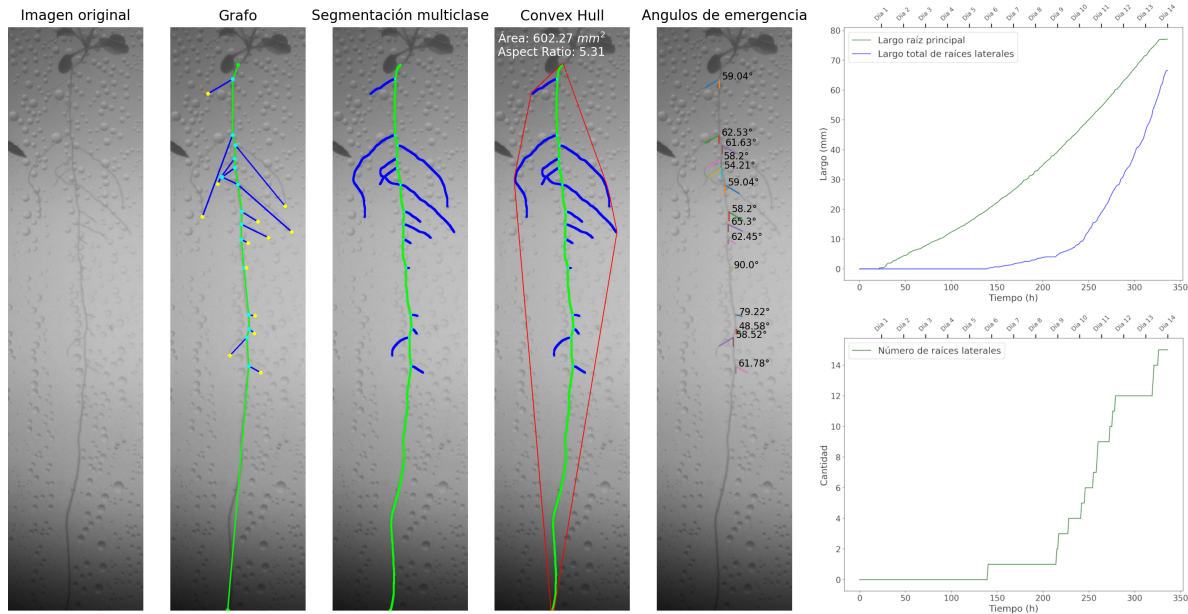


Figura 2.2: Salida del modelo y ejemplos de parámetros fenotípicos extraíbles.

Extracción de parámetros fenotípicos:

Una vez obtenido el grafo, la extracción de parámetros fenotípicos se realiza utilizando las distintas salidas del modelo. Se puede medir la longitud de la raíz principal siguiendo el camino más corto

entre el nodo de inicio y el nodo *tip*, que fueron rastreados temporalmente. Esto nos permite etiquetar las aristas que forman este camino como raíz principal, y, por lo tanto, definir el resto de las aristas como raíces laterales. Sumando los pesos de las aristas, que contienen la longitud real del esqueleto en términos de píxeles y multiplicando estos largos por el tamaño en milímetros de los píxeles (obtenible gracias a un código QR de tamaño fijo que se coloca en la placa), obtenemos la longitud de la raíz principal y la longitud de las raíces laterales, respectivamente. Para obtener la cantidad de raíces laterales, una aproximación inicial es contar los nodos de la raíz principal menos 2, pero dado que pueden existir *loops*, donde la raíz lateral vuelve a chocar con la principal, se resta 1 por cada uno de estos *loops* encontrados. La segmentación, a su vez, nos permite obtener parámetros a través del *convex hull* (el polígono convexo de menor tamaño que contiene toda la segmentación), como el área ocupada por la raíz o la relación entre altura y ancho. Partiendo desde los nodos de inicio de raíz lateral, y caminando sobre el esqueleto, también podemos obtener ángulos respecto al eje de gravedad a medidas fijas (obteniendo un ángulo de emergencia) o hasta el final de la raíz lateral (obteniendo el ángulo total de la raíz). A partir de todas estas métricas, también se pueden derivar métricas adicionales, como velocidades de crecimiento, porcentaje de raíz principal sobre el total, etc. Ejemplos de los parámetros primarios obtenidos se pueden observar en la Figura 2.2.

2.3. Experimentos y resultados

2.3.1. Base de datos

Generamos dos conjuntos de datos diferentes en este trabajo. El primero se utilizó para entrenar y evaluar el rendimiento de los modelos de segmentación, mientras que el segundo se empleó como un caso de uso exploratorio para evaluar el crecimiento de las raíces bajo fotoperíodos alternativos, específicamente bajo luz continua (denominado *continuous light*) y un ciclo de 16 horas de luz y 8 horas de oscuridad (denominado *long day*). Es importante destacar que al dividir las particiones de entrenamiento, validación y prueba, evitamos incluir imágenes correspondientes al mismo video en diferentes particiones para evitar evaluaciones sesgadas y excesivamente optimistas.

- **Conjunto de datos utilizado para entrenar y validar los modelos de aprendizaje profundo para la segmentación de raíces:**

- Este conjunto de datos de entrenamiento consta de 331 imágenes recopiladas de 55 videos diferentes. En promedio, se tomaron seis imágenes de la misma placa en diferentes estados de crecimiento.
- El conjunto de datos de prueba consta de 55 imágenes recopiladas de 11 videos diferentes, con 5 imágenes por video.

- **Conjunto de datos para el caso de uso:**

- En este caso de uso exploratorio, se utilizaron 12 videos para cada uno de los dos fotoperíodos analizados (luz continua y ciclo de 16 horas de luz / 8 horas de oscuridad).
- Las imágenes se capturaron a intervalos de 15 minutos a lo largo de los primeros 17 días, lo que resultó en un total de 1632 fotogramas por fotoperíodo.
- Después de procesar los videos, se descartaron los resultados correspondientes a los primeros 3 días previos a la germinación de las semillas para centrarse en el período de crecimiento de las raíces.
- Se seleccionaron 25 plantas de cada fotoperíodo para llevar a cabo los análisis de crecimiento temporal.

2.3.2. Resultados

Para la evaluación de los modelos de segmentación, medimos tres métricas diferentes:

Model	DSC		Recall		HD (mm)		Time	#	Params
	Umbralizado	CRF	Umbralizado	CRF	Umbralizado	CRF			
UNet	0.769±0.048	0.774±0.044	0.871±0.044	0.830±0.056	10.25±7.45	9.39±7.94	0.29s	488.212	
ResUNet	0.768±0.050	0.770±0.047	0.862±0.046	0.823±0.057	8.83±6.71	7.53±5.91	0.33s	505.046	
Deeply Supervised	0.769±0.048	0.772±0.045	0.861±0.044	0.815±0.057	8.14±7.34	6.95±5.42	0.49s	532.336	
ResUNet (propuesto)									
SegNet	0.768±0.043	0.773±0.040	0.862±0.044	0.824±0.053	7.42±6.40	6.81±5.65	1.49s	29.460.450	
DeepLab	0.666±0.055	0.609±0.079	0.763±0.077	0.600±0.113	7.58±7.79	7.56±7.52	1.86s	58.009.410	
Ensamble (propuesto)	0.772±0.048	0.774±0.044	0.864±0.044	0.804±0.061	6.68±5.08	6.45±4.98	4.5s		

Tabla 2.1: **Evaluación cuantitativa de las diferentes arquitecturas de CNN comparadas.** Se evaluaron el coeficiente Dice-Sørensen, la distancia de Hausdorff y la sensibilidad (*recall*), los tiempos de ejecución fueron medidos en un equipo Intel Core i7-8700, con 64 GB de RAM, utilizando una GPU NVIDIA Titan Xp.

1. El coeficiente Dice-Sørensen (DSC) cuantifica la superposición entre la predicción y la máscara de segmentación de referencia.
2. La distancia de Hausdorff (HD) indica la distancia máxima entre la predicción y la máscara de segmentación de referencia.
3. La sensibilidad representa la proporción de píxeles de raíz recuperados sobre el total de píxeles de raíz.

Los resultados cuantitativos se encuentran detallados en la Tabla 2.1, donde se comparan salidas sin preprocesar utilizando solo umbralizado, y salidas preprocesadas mediante *Conditional random fields* (CRF) [50], comúnmente utilizados para esta tarea. Basándonos en estos hallazgos, proponemos la adopción de dos modelos diferentes según el contexto de aplicación, considerando la búsqueda de velocidad o precisión. En el caso del método rápido, observamos que las variantes de UNet se destacan, siendo capaces de procesar una imagen de alta resolución en tan solo medio segundo utilizando una GPU estándar. Estos modelos presentan una menor complejidad de parámetros en comparación con arquitecturas como SegNet (Badrinarayanan et al., 2017) y DeepLab (Chen et al., 2018), lo que explica su tiempo de ejecución reducido. Entre los modelos UNet, resalta la arquitectura Deeply Supervised ResUNet propuesta en esta tesis, que muestra una distancia de Hausdorff significativamente menor mientras mantiene valores igualmente buenos para el coeficiente de DSC y la sensibilidad. Por otro lado, en el enfoque de método preciso, sugerimos la combinación de todas las arquitecturas implementadas en un ensamble. Aunque esto aumenta el tiempo de ejecución en un factor de 9, se logra el mejor rendimiento en todas las métricas, superando a modelos como SegNet y DeepLab.

La elección de usar una alternativa u otra dependerá de las capacidades computacionales de quien quiera utilizar esta herramienta. Analizar un video en crecimiento, como los propuestos para el caso de uso a presentar, requiere segmentar 1632 imágenes, lo que tomaría 13.6 minutos en el modelo rápido o 122.4 minutos con el modelo preciso. El caso de uso requirió segmentar 24 videos, lo que equivaldría a 6 horas de GPU con el modelo rápido o casi 49 horas con el modelo preciso, y posteriormente realizar la extracción de los grafos temporales planta por planta.

En la Figura 2.3, se puede observar la dinámica de los distintos parámetros fenotípicos extraídos a lo largo del tiempo. En particular, observamos que los valores de largo de la raíz principal (*MR Length*), largo de las raíces laterales (*LR Length*) y número de raíces laterales (*Number of LR*) coinciden con los valores obtenidos manualmente por los biólogos para instantes fijos de crecimiento. También se identificaron nuevos parámetros, como la velocidad de crecimiento de la raíz principal (*MR Growth Speed*) y su transformada de Fourier, que mostraron diferencias notables entre las condiciones de luz continua y las de día y noche, destacándose una diferencia importante en las frecuencias de 1/24hs y 1/12hs. Esto conllevó a reafirmar resultados biológicos conocidos respecto al ciclo circadiano de las raíces que sugieren que las plantas crecen más rápido durante el día que durante la noche, que no habían podido ser cuantificado hasta este momento.

Para una descripción detallada de los resultados obtenidos y su discusión, puede referirse a la Sección *Analyses* del Apéndice B.

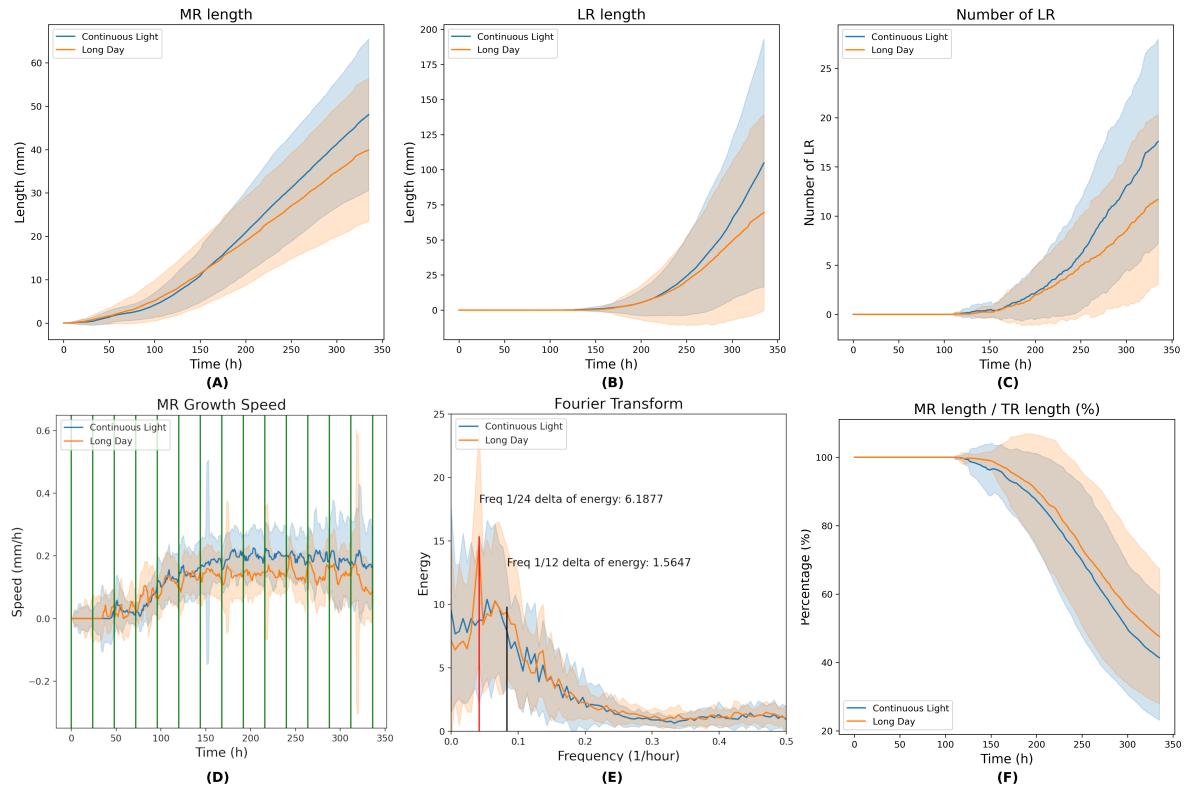


Figura 2.3: **Parámetros fenotípicos observados a través del tiempo.** (A) Muestra el largo de la raíz principal promedio por fotoperíodo a lo largo del tiempo, con el desvío estándar sombreado. (B) El largo de las raíces laterales (C) El número de raíces laterales (D) La velocidad de crecimiento de la raíz principal (E) La transformada de Fourier de la velocidad de crecimiento (F) La proporción de raíz principal en el total de raíces (TR)

Sección 3

Método unificado 2D: segmentación anatómica basada en landmarks para imágenes de rayos X

En el capítulo anterior se introdujo un método desacoplado para la extracción de grafos a partir de imágenes 2D. Tal como fue anticipado, dicho método presenta etapas independientes, que pueden introducir errores acumulables que disminuyan la performance de los modelos. En este capítulo, presentaremos una nueva arquitectura híbrida de redes neuronales convolucionales, diseñada para obtener estructuras de grafo directamente a partir de imágenes, sin la necesidad de contar con etapas desacopladas de procesamiento.

3.1. Antecedentes

Desde principios de la década de 1990, la tarea de segmentar estructuras anatómicas se ha abordado con el uso de *landmarks*, utilizando distintas variantes de lo que se conoce como modelo de distribución de puntos (PDM, por sus siglas en inglés, *point distribution model*) [19]. Los PDM son plantillas de forma flexibles que describen cómo pueden variar las ubicaciones relativas de los landmarks. Técnicas basadas en PDM, como los modelos de forma activa (ASM) [19, 87] y los modelos de apariencia activa (AAM) [18], fueron convirtiéndose en el estándar para la segmentación anatómica a fines del siglo pasado. Posteriormente, durante la siguiente década, el desarrollo de algoritmos de registración de imágenes más potentes y robustos [101] posicionó a los algoritmos de *deformable template matching* como la opción preferida para la segmentación anatómica y construcción de atlas [29, 37, 72].

Más recientemente, con la llegada de las CNN al mundo de la segmentación de imágenes médicas [84, 82], se realizaron grandes esfuerzos para incorporar restricciones anatómicas en tales modelos [45, 66]. Sin embargo, la mayoría de estos métodos trabajan directamente en el espacio de píxeles, produciendo máscaras de segmentación densas, pero sin proveer ni incorporar información sobre la estructura anatómica subyacente. Por el contrario, modelos estructurados como los grafos pueden representar fácilmente landmarks, contornos y superficies. En línea con esta idea, estudios recientes [60, 7, 6] han integrado CNN estándar con diferentes representaciones de estructuras de landmarks. Estos métodos emplean representaciones de forma de baja dimensionalidad, como la descomposición del espacio de los landmarks mediante Análisis de Componentes Principales (PCA) [60, 7] o modelos más sofisticados de distribución de partículas [6].

Es en este contexto donde proponemos una nueva dirección, inspirada en investigaciones sobre modelos generativos de grafos [79]. Nuestra propuesta es reemplazar las representaciones lineales tradicionales por representaciones no lineales basadas en convoluciones en grafo [10]. En este capítulo, exploraremos en detalle cómo este enfoque se traduce en una metodología para la segmentación anatómica basada en landmarks de imágenes de rayos-X.

3.2. Métodos propuestos

3.2.1. Descripción del problema

Supongamos que tenemos un dataset $\mathcal{D} = \{(\mathbf{I}, \mathbf{G})_n\}_{0 < n \leq N}$, compuesto de N imágenes $\mathbf{I}$ con su correspondiente segmentación basada en landmarks representada como un grafo $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$, donde V es el conjunto de vértices para M landmarks, $\mathbf{A} \in \{0, 1\}^{M \times M}$ es la matriz de adyacencia indicando la conectividad entre pares de vértices ($a_{ij} = 1$ indica que hay una arista que conecta los vértices i y j , mientras $a_{ij} = 0$ indica lo contrario) y $\mathbf{X} \in \mathbb{R}^{M \times s}$ es una función (representada como una matriz) asignando un vector de características a cada vértice. En este caso, asigna una coordenada espacial bidimensional (la posición del landmark en la imagen) a cada vértice ($s = 2$). En este contexto es común tener anotaciones manuales con el mismo número de nodos. Por lo tanto, asumiremos que V y $\mathbf{A}$ son iguales para todas las imágenes del dataset, siendo la única diferencia entre ellas la posición espacial de los landmarks definida en $\mathbf{X}$. Asumir esto nos permite seguir el trabajo de [22, 79] y usar convoluciones espectrales en grafos para aprender representaciones no lineales de la anatomía.

3.2.2. Convoluciones espectrales en grafos

Las convoluciones espectrales en grafos se construyen utilizando la descomposición en valores y vectores propios de la matriz Laplaciana del grafo, aprovechando la propiedad de que las convoluciones en el dominio de los vértices son equivalentes a multiplicaciones en el dominio espectral del grafo [86], de manera análoga a cómo las convoluciones en señales 2D son equivalentes a multiplicaciones en el dominio frecuencial.

Siguiendo el trabajo de Defferrard et al [22], restringimos la clase de filtros que utilizaremos para la convolución a filtros polinomiales, ya que estos tienen la propiedad de funcionar estrictamente localizados en el dominio de los vértices (un polinomio de orden k considera k saltos de vecindad a partir del vértice central) y reducen computacionalmente la complejidad del operador de convolución. Más aún, estos filtros pueden aproximarse utilizando una aproximación en términos de polinomios de Chebyshev, que se computan recursivamente. Siguiendo el trabajo de [22, 79], tomaremos esta aproximación para implementar las convoluciones espectrales, notando en particular que una convolución espectral tomará matrices de características $\mathbf{X}^\ell$ como entrada y producirá versiones filtradas $\mathbf{X}^{\ell+1}$ como salida, similar a cómo las convoluciones trabajan sobre mapas de características en el dominio de la imagen.

3.2.3. Autocodificadores

El modelo propuesto toma imágenes como entrada y genera un grafo como salida, combinando convoluciones estándar en imágenes con convoluciones en grafos, con la idea de que esta estructura se puede entrenar de manera unificada, mediante la retro propagación directa de una función de pérdida sobre el espacio de los nodos del grafo. Su diseño está influenciado por los autocodificadores variacionales [48], que modelan el espacio latente como una variable Gaussiana multivariada. Para comprender mejor este concepto, primero definamos lo que son los autocodificadores.

Los autocodificadores (o *autoencoders* en inglés) son redes neuronales diseñadas para reconstruir su entrada en la salida de la misma red. Siguen una estructura codificador-decodificador, similar a la arquitectura de la red neuronal U-Net, pero sin incorporar conexiones de salto. En un autocodificador, un codificador $z = f_{enc}(\mathbf{x})$ mapea la entrada $\mathbf{x}$ a un código latente $\mathbf{z}$ de menor dimensión, que luego es procesado por el decodificador $f_{dec}(\mathbf{z})$ para reconstruir la entrada original. La restricción impuesta por la baja dimensionalidad del código $\mathbf{z}$ obliga al modelo a capturar información relevante, lo que lleva al aprendizaje de representaciones efectivas de la distribución de datos.

Estos modelos se entrenan para minimizar una función de pérdida $\mathcal{L}_r(\mathbf{I}, f_{dec}(f_{enc}(\mathbf{I})))$, que considera la reconstrucción de la entrada en la salida. Para regularizar el espacio latente $\mathbf{z}$, se agrega un término a la función de costo, convirtiendo el autocodificador en un autocodificador variacional [48]. En este enfoque, se asume que los códigos en el espacio latente $\mathbf{z}$ se muestrean de una distribución $Q(\mathbf{z})$, que se elige como una Gaussiana multivariada con media 0 y desviación estándar 1. En la práctica, durante el entrenamiento, esto significa que los códigos en el espacio latente $\mathbf{z}$ se muestrean de una distribución

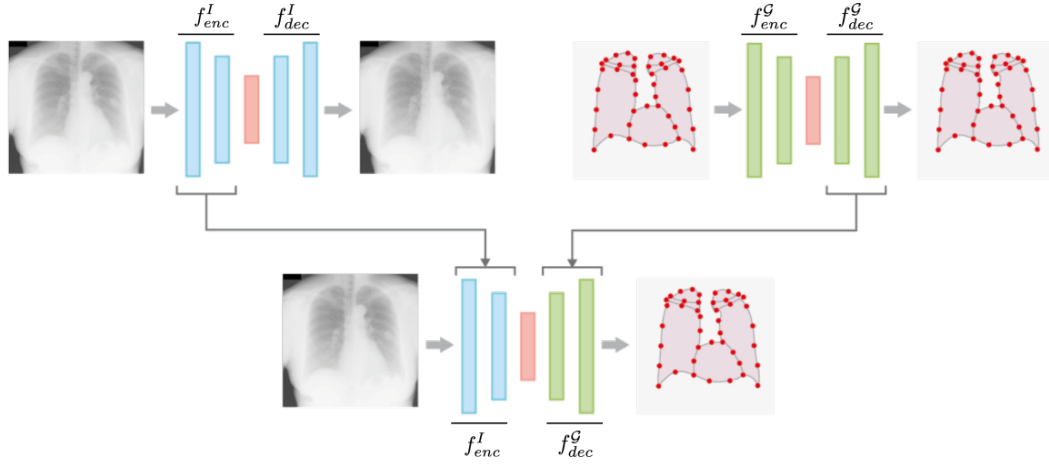


Figura 3.1: **Modelo Híbrido:** El modelo híbrido propuesto (abajo) es una arquitectura codificador-decodificador que combina convoluciones estándar para la codificación de características de la imagen (en azul), con convoluciones espectrales en grafos (en verde) para decodificar representaciones anatómicas plausibles.

$\mathcal{N}(\mu, \sigma)$ utilizando el truco de la reparametrización [48]. Aquí, μ y σ son parámetros determinísticos generados por el codificador $f_{enc}(\mathbf{I})$.

Dado un valor muestreado $\mathbf{z}$, podemos generar (reconstruir) el dato correspondiente utilizando el decodificador $f_{dec}(\mathbf{z})$. La función de entrenamiento del modelo se define como:

$$\mathcal{L}_a = \mathcal{L}_r(\mathbf{I}, f_d(\mathbf{z})) + w \text{KL}(\mathcal{N}(0, 1) || Q(\mathbf{z}|\mathbf{I})), \quad (3.1)$$

donde el primer término es la pérdida de reconstrucción, el segundo término impone una distribución Gaussiana $\mathcal{N}(0, 1)$ en el espacio latente a través de la divergencia de Kullback-Leibler, y w es un factor de escala.

Dado que los autocodificadores pueden aplicarse a todo tipo de datos estructurados, como imágenes y grafos, esto nos permite entrenar un autocodificador para aprender la distribución de ciertas imágenes de interés y otro para modelar la distribución de grafos asociados a esas imágenes.

3.2.4. Modelos híbridos

En base a estas ideas, se construyó un modelo híbrido que combina el codificador de un autocodificador entrenado para reconstruir imágenes f_{enc}^I con el decodificador de un autocodificador entrenado para reconstruir grafos f_{dec}^G , lo que permite obtener un grafo que represente la estructura de interés presente en la imagen. En la Figura 3.1 se puede ver ilustrada esta idea, aplicada a una imagen de rayos X junto con segmentaciones representadas mediante grafos que modelan el contorno de los órganos (pulmones, clavículas y corazón en el ejemplo). Para establecer la conexión entre el codificador y el decodificador, se requiere que ambos utilicen capas completamente conectadas llegando el codificador a un espacio latente $\mathbf{z}$ con un tamaño determinado, y el decodificador partiendo desde ese mismo tamaño.

3.2.5. Aprendiendo en múltiples resoluciones

Al decodificador convolucional en grafos se le agregó una capa fija de sobremuestreo (*unpooling*), provocando que durante la etapa de decodificación haya dos resoluciones. A partir del muestreo desde espacio latente (enz), se obtiene un grafo de menor resolución mediante una capa completamente conectada. Luego, se utilizan capas de convoluciones en grafo y posteriormente se sobremuestra para llegar a la resolución original. Esto permite aprender características en distintas resoluciones, siguiendo el trabajo en autocodificadores en grafos realizado en [79]. Adoptamos una estrategia simple, donde

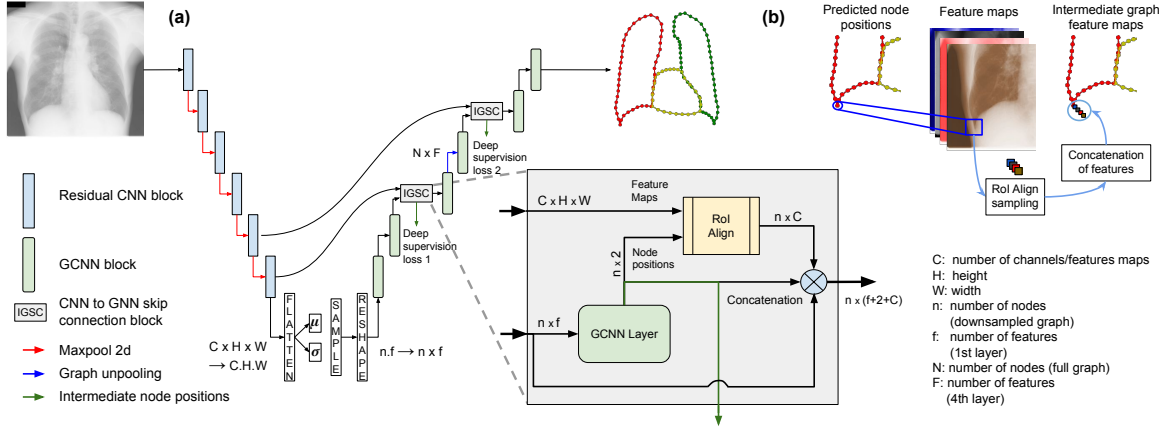


Figura 3.2: (a) La arquitectura HybridGNet propuesta combina convoluciones estándar en imágenes (celeste) con convoluciones espectrales en grafos (verde) para obtener segmentaciones anatómicamente plausibles mediante representaciones basadas en grafos. (b) Visualización ilustrativa del muestreo mediante RoIAlign y la concatenación de las características obtenidas al grafo, dentro del módulo IGSC.

todos los grafos $\mathbf{G}$ en nuestro dataset son preprocesados para obtener grafos de menor resolución $\mathbf{G}^k$ al reducir a la mitad el número de nodos k veces consecutivas, reemplazando un par de nodos por uno solo, computado como el promedio de estos mismos. Luego, la capa de sobremuestreo propuesta hace la operación inversa a esta tarea, duplicando los nodos y colocando los nuevos en posiciones interpoladas entre los de baja resolución.

3.2.6. Conexiones de salto entre grafos e imágenes

Bajo la hipótesis de que características locales pueden ayudar a producir mejores estimaciones de las posiciones finales de los landmarks, se diseñó una nueva capa de conexiones de salto entre imágenes y grafos, denominada *Image-to-Graph Skip Connection* (IGSC). IGSC utiliza el bien conocido módulo de extracción de características localizadas en modelos de detección llamado *RoIAlign* [35], con el fin de obtener características localizadas en una ventana alrededor de cada nodo desde una capa del codificador en particular. Para aprender desde qué posiciones obtener las características para cada nodo, se requiere la incorporación de una capa intermedia que aprenda las posiciones de los nodos mediante *deep supervision* [54] a partir de las características del grafo de cada nodo. Así, las características del codificador muestreadas en el nivel correspondiente se concatenan a las características del grafo (de forma similar a como sucede en los clásicos saltos de conexiones o *skin-connections* de la arquitectura UNet clásica) y a las posiciones desde las que se obtuvieron, obteniendo un grafo con características aumentadas.

3.2.7. Modelo HybridGNet completo

Combinando las ideas del modelo híbrido propuesto originalmente, con el aprendizaje a distintas resoluciones, la supervisión profunda y la extracción localizada de características para la aumentación del grafo, obtenemos la arquitectura final, denominada *HybridGNet*, que se puede observar en la Figura 3.2.a). El grafo aumentado se ilustra gráficamente en la subfigura (b). Para más detalles sobre la metodología, remitimos al Apéndice C, donde se introdujo por primera vez esta arquitectura, y al Apéndice D, donde se extiende la idea original mediante la incorporación de múltiples resoluciones y el uso de conexiones de salto entre imágenes y grafos.

Tabla 3.1: Resultados de la segmentación anatómica basada en landmarks para el conjunto de datos JSRT. Media (desviación estándar). HD en milímetros.

Tipo	Model	MSE	DSC Lungs	HD Lungs	DSC Heart	HD Heart
Métodos en Landmarks	PCA	340.024 (243.549)	0.945 (0.014)	17.445 (9.669)	0.906 (0.037)	14.602 (5.400)
	FC	332.197 (242.379)	0.945 (0.017)	17.535 (10.352)	0.910 (0.038)	15.020 (5.785)
	MultiAtlas	492.262 (298.138)	0.944 (0.013)	20.317 (9.344)	0.886 (0.056)	16.780 (6.839)
	HybridGNet (sin IGSC)	294.621 (274.497)	0.952 (0.013)	15.642 (10.922)	0.913 (0.038)	13.658 (5.548)
	HybridGNet (2 IGSC)	200.748 (211.080)	0.974 (0.007)	12.089 (9.344)	0.933 (0.031)	11.613 (5.581)
Métodos nivel pixel	UNet	—	0.981 (0.008)	21.839 (26.291)	0.942 (0.030)	25.176 (34.570)
	UNet + Post-DAE	—	0.965 (0.010)	17.969 (14.457)	0.935 (0.029)	15.444 (14.283)
	nnUNet	—	0.984 (0.005)	9.615 (7.874)	0.952 (0.023)	9.782 (5.006)

3.3. Experimentos y Resultados

3.3.1. Base de Datos

Utilizamos el conjunto de datos de la Asociación Japonesa de Tecnología Radiológica (JSRT) [85], que consta de 247 imágenes de alta resolución con anotaciones de puntos de referencia realizadas por expertos para los pulmones, el corazón y las clavículas [32]. Para estos experimentos, nos centramos en los puntos de referencia del pulmón y el corazón, que incluyen un número fijo de 44, 50 y 26 puntos de referencia para cada paciente, correspondientes al pulmón derecho, el izquierdo y el corazón, respectivamente. Para evaluar el rendimiento de los modelos en situaciones de cambio de dominio, también incorporamos las bases de datos Shenzhen [43] y Montgomery [13], que solo contienen máscaras de pulmón. Para evaluar el rendimiento frente a oclusiones reales causadas por marcapasos, incorporamos un pequeño subconjunto del conjunto de datos Padchest [11], con anotaciones manuales proporcionadas por radiólogos del Hospital Italiano de Buenos Aires para este trabajo. Para obtener más detalles sobre las bases de datos utilizadas, consulte el Apéndice D, sección IV-A.

3.3.2. Modelos para comparación

Para comparar, elegimos un conjunto de modelos basados en landmarks, empezando por un modelo de segmentación MultiAtlas [39] y dos modelos de forma: uno basado en un decodificador de PCA y otro cuyo decodificador es simplemente una capa completamente conectada, ambos utilizando el mismo codificador que la HybridGNet. A la vez, elegimos modelos muy utilizados en segmentación a nivel pixel para compararnos, tomando una UNet [82] a la que le aplicamos el método PostDAE [51] para obtener segmentaciones con mayor plausibilidad anatómica. Por último, utilizamos el modelo nnUNet [42], que incluye un proceso de autoconfiguración para obtener la mejor performance ante un conjunto de datos de entrenamiento. nnUNet representa actualmente el estado del en segmentación de imágenes médicas a nivel pixel.

3.3.3. Evaluando la segmentación anatómica

Como primer experimento, comparamos HybridGNet sin y con conexiones de salto con modelos del estado del arte basados en landmarks y píxeles. La Tabla 3.1 muestra cómo HybridGNet supera a los distintos modelos basados en landmarks en términos de todas las métricas observadas. La incorporación de las conexiones de salto produce una mejora significativa en los resultados y establece a HybridGNet como una opción altamente competitiva en comparación con el método de segmentación médica del estado del arte, nnUNet [42].

3.3.4. Evaluación en contexto de cambio de dominio

En un segundo experimento, evaluamos los mismos modelos, pero en un escenario más complejo, el de cambio de dominio. El cambio de dominio se refiere a una variación en el dominio en el que se evalúa el modelo, en comparación con el dominio en el que se entrenó [17]. En muchos casos, el cambio de dominio puede provocar una disminución significativa en el rendimiento de los métodos supervisados de aprendizaje automático. Este problema puede surgir debido a varios factores, como cambios en

Tabla 3.2: Resultados de cambio de dominio para la segmentación anatómica basada en landmarks del conjunto de datos JSRT a Montgomery y Shenzhen. Media (desviación estándar). HD en píxeles.

Model	Montgomery		Shenzhen	
	DSC Lungs	HD Lungs	DSC Lungs	HD Lungs
PCA	0.906 (0.082)	60.08 (36.89)	0.894 (0.054)	79.12 (47.73)
FC	0.897 (0.087)	60.02 (35.77)	0.895 (0.051)	77.11 (48.15)
Multi-alas	0.909 (0.080)	61.77 (31.62)	0.900 (0.054)	88.13 (48.94)
HybridGNet (sin IGSC)	0.909 (0.070)	55.97 (35.70)	0.901 (0.047)	72.13 (47.40)
HybridGNet 2 IGSC	0.954 (0.043)	45.50 (32.48)	0.935 (0.038)	64.46 (51.53)
UNet	0.944 (0.068)	127.72 (97.76)	0.933 (0.055)	220.89 (102.94)
UNet+PostDAE	0.907 (0.102)	119.53 (85.10)	0.906 (0.083)	135.05 (97.32)
nnUNet	0.955 (0.068)	44.79 (60.31)	0.949 (0.055)	61.31 (67.97)

Tabla 3.3: Resultados de la generación de anotaciones de landmark a partir de segmentaciones densas en el conjunto de datos JSRT. Promedio (desviación estándar). HD en milímetros.

Modelo	MSE	DSC Pulmones	HD Pulmones	DSC Corazón	HD Corazón
PCA	77.2 (133.7)	0.978 (0.009)	6.02 (3.46)	0.97 (0.007)	4.37 (1.61)
FC	105.3 (173.2)	0.970 (0.014)	7.82 (3.96)	0.96 (0.014)	5.78 (2.94)
Multi-atlas	236.3 (244.8)	0.991 (0.004)	10.98 (8.53)	0.99 (0.006)	4.64 (2.48)
HybridGNet (sin IGSC)	96.9 (145.0)	0.970 (0.009)	7.65 (3.75)	0.96 (0.013)	6.02 (2.77)
HybridGNet 2 IGSC	55.1 (113.4)	0.991 (0.003)	3.92 (4.42)	0.99 (0.005)	2.58 (1.59)

los métodos de adquisición, diferentes centros médicos o diferentes características demográficas de la población. En este contexto, evaluamos el rendimiento del modelo entrenado en datos de JSRT, obtenidos de centros clínicos en Japón, cuando se evalúa con imágenes de los conjuntos de datos Shenzhen (China) y Montgomery (Estados Unidos). La Tabla 3.2 muestra que el modelo de HybridGNet con 2 conexiones de salto experimenta una disminución mucho menor en el rendimiento en comparación con los otros modelos, lo que demuestra su capacidad para funcionar bien incluso en situaciones de cambio de dominio. Para obtener más detalles sobre estos resultados, consulte el Apéndice D, sección V-3.

3.3.5. Generación de nuevas bases de datos mediante HybridGNet

Como tercer experimento, y debido a la limitada disponibilidad de datos con anotaciones de landmark, desarrollamos un experimento en el cual, en lugar de utilizar imágenes como entrada, suministramos segmentaciones densas al modelo con el propósito de obtener landmarks con correspondencia uno a uno. Esta estrategia nos permitió entrenar nuestro modelo en nuevos conjuntos de datos.

Al contar con segmentaciones densas como referencia para los conjuntos de datos Montgomery, Shenzhen y un pequeño subconjunto de Padchest, decidimos entrenar un modelo que transformara estas segmentaciones en anotaciones basadas en landmarks. Inicialmente, entrenamos un modelo utilizando los mismos conjuntos de entrenamiento y prueba de JSRT para validar esta estrategia, cuyos resultados se encuentran en la Tabla 3.3. Es importante destacar que los resultados mostraron que, en términos del Coeficiente Dice-Sørensen (DSC), HybridGNet con 2 conexiones de salto alcanza valores en el orden de 0.99.

Con la confirmación de que nuestra hipótesis era acertada, procedimos a entrenar un modelo que utilizara exclusivamente las imágenes de JSRT para generar anotaciones de landmark para los conjuntos de datos de Montgomery y Shenzhen, así como para el subconjunto de Padchest, sin reservar ninguna imagen para pruebas, dado que el modelo ya había sido validado. Para asegurar que las anotaciones fueran idénticas a las de entrada, desplazamos los landmarks al punto más cercano en el contorno original. Este proceso culminó en la creación de una nueva base de datos que denominamos *Chest X-ray Landmark Dataset* y que se encuentra disponible de forma gratuita en GitHub¹. Para obtener detalles adicionales sobre este experimento, consulte la Sección D V-2.

¹<https://github.com/ngaggion/Chest-xray-landmark-dataset>

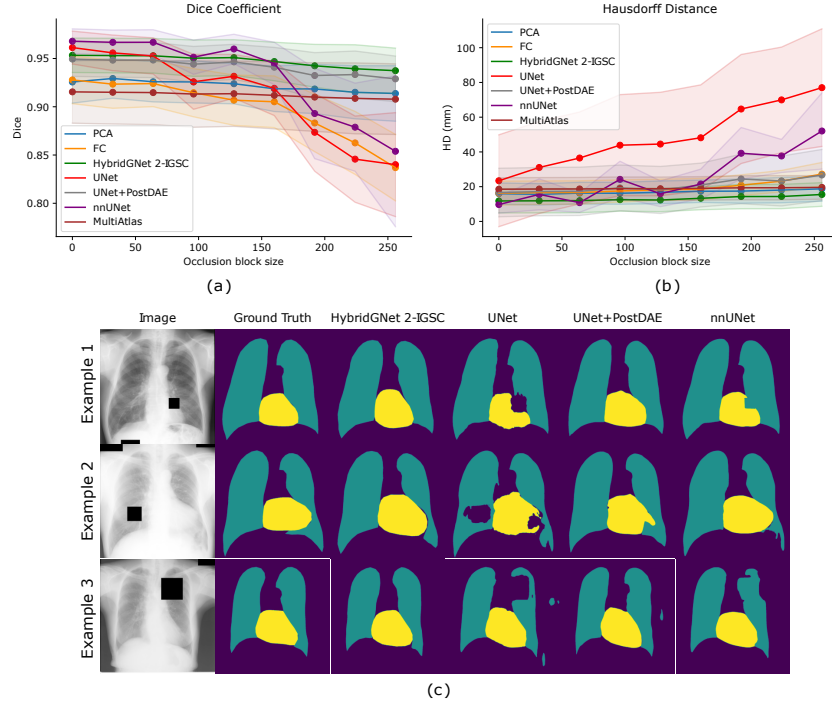


Figura 3.3: Estudio de occlusiones artificiales. (a) Coeficiente Sorensen-Dice y (b) distancia de Hausdorff para un aumento del tamaño del bloque de occlusión en occlusiones artificiales. (c) Resultados cualitativos.

3.3.6. Evaluación de la robustez ante occlusiones

Por último, desarrollamos dos experimentos para evaluar la robustez del modelo en casos más extremos, como lo son las occlusiones de imágenes. Las occlusiones de imágenes (IO) son una ocurrencia común en las radiografías de tórax, ya sea debido a la desidentificación del paciente, lo que implica la cobertura de información confidencial con parches negros, o a la presencia de dispositivos como marcapasos, tubos, cables o electrodos que pueden ocultar partes importantes de órganos u otras estructuras de interés.

En un primer experimento, simulamos occlusiones artificiales utilizando cuadros negros de diferentes tamaños en las imágenes de prueba de JSRT. Es importante notar que el modelo no fue entrenado con imágenes que posean dichas occlusiones (ver las Figuras 3.3 (a) y (b)). Observamos que HybridGNet demuestra ser más resistente a estas occlusiones en comparación con los modelos basados en píxeles, sobre todo al aumentar su tamaño.

En el segundo experimento (ver Figura 3.4), evaluamos occlusiones reales causadas por dispositivos médicos utilizando un conjunto de imágenes de pacientes con marcapasos del conjunto de datos Padchest. Para evitar problemas de cambio de dominio en este contexto, desarrollamos nuevos modelos incorporando en entrenamiento la gran mayoría de las imágenes de Padchest para las cuales obtuvimos anotaciones en landmark, excluyendo aquellas con marcapasos. Observamos que en esta configuración, HybridGNet supera a UNet, y aunque no se obtuvieron diferencias significativas en términos cuantitativos en comparación con nnUNet, visualmente se evidencia que las occlusiones no afectan la segmentación de HybridGNet. Para obtener información detallada sobre estos resultados, consulte el Apéndice D, Sección V-4.

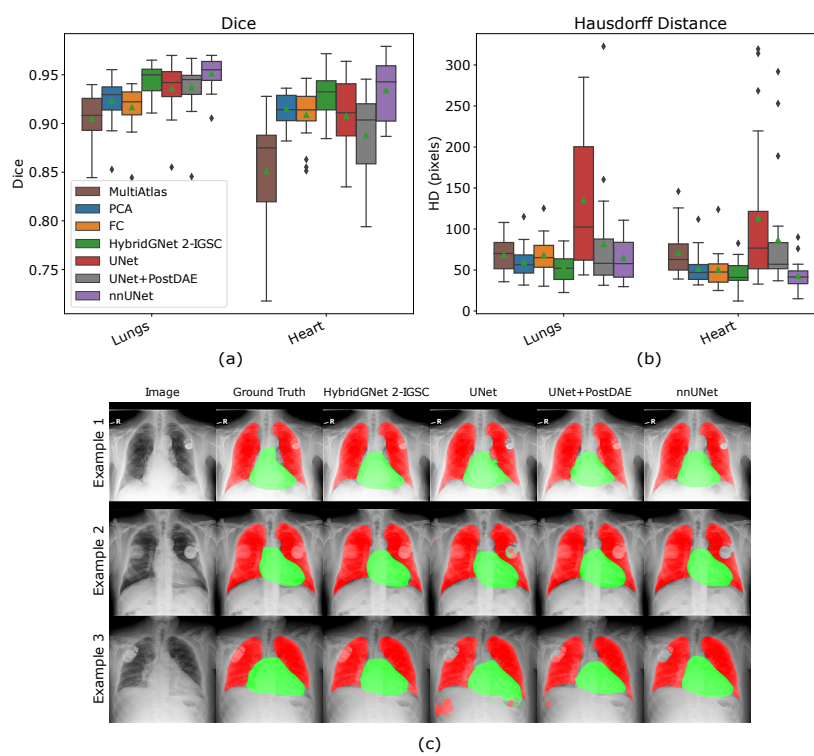


Figura 3.4: Estudio de oclusiones reales. (a) Coeficiente Sorensen-Dice y (b) distancia de Hausdorff para el subconjunto de marcapasos de Padchest. (c) Resultados cualitativos.

Sección 4

Método unificado 3D: extracción de mallas a partir de imágenes de resonancia magnética cardíaca

En el capítulo anterior se presentó un modelo para extracción de grafos a partir de imágenes 2D. En este capítulo, se presenta una extensión de dicha arquitectura pensada para obtener grafos que representen mallas de superficie y volumétricas a partir de imágenes 3D, con aplicación al problema de análisis de imágenes cardíacas.

4.1. Antecedentes

La obtención precisa de mallas tridimensionales del corazón a partir de imágenes de resonancia magnética cardíaca (CMR) desempeña un papel fundamental en el diagnóstico, tratamiento y estudio de enfermedades cardiovasculares. Estas mallas son esenciales para estudiar el corazón: realizar simulaciones, identificar biomarcadores o analizar su funcionamiento.

La generación detallada y anatómicamente plausible de mallas ventriculares a partir de CMR es un desafío que consume mucho tiempo si se realiza de manera manual. Los métodos automáticos para la extracción de mallas dependen en su mayoría de una serie de etapas, generalmente incluyendo la segmentación a nivel de voxel, la extracción de malla superficial y, finalmente, la generación de una malla volumétrica [24, 64]. Una de las principales limitaciones de estos enfoques es que las técnicas de segmentación a nivel de voxel (generalmente modelos de aprendizaje profundo como U-Net [82] o V-Net volumétrico [59]) son propensas a introducir errores debido al soporte local de los modelos convolucionales, lo que resulta en máscaras poco realistas con agujeros o segmentaciones espurias [52]. Otros enfoques se basan en deformar una plantilla inicial. En este sentido, estudios previos han propuesto estimar las deformaciones de los nodos de la malla [77, 69] o deformar directamente el espacio que encierra una plantilla de corazón completa lista para la simulación [49]. Sin embargo, la precisión de estos enfoques está limitada por la calidad de las deformaciones estimadas.

Más cerca de nuestro trabajo, en [90, 97] se entrena una red neuronal de extremo a extremo para estimar formas parametrizadas directamente a partir de imágenes. Estos métodos infieren los parámetros de un modelo de forma basado en Análisis de Componentes Principales (PCA) utilizando redes neuronales convolucionales. A pesar de la utilidad de los modelos basados en PCA en la extracción de mallas a partir de imágenes volumétricas, su expresividad está inherentemente limitada por la linealidad del PCA. Para abordar esta limitación, mejoramos la expresividad de las formas paramétricas aprovechando un decodificador no-lineal de grafos convolucionales, capaz de manejar tanto mallas superficiales como volumétricas.

4.2. Métodos propuestos

4.2.1. Descripción del problema

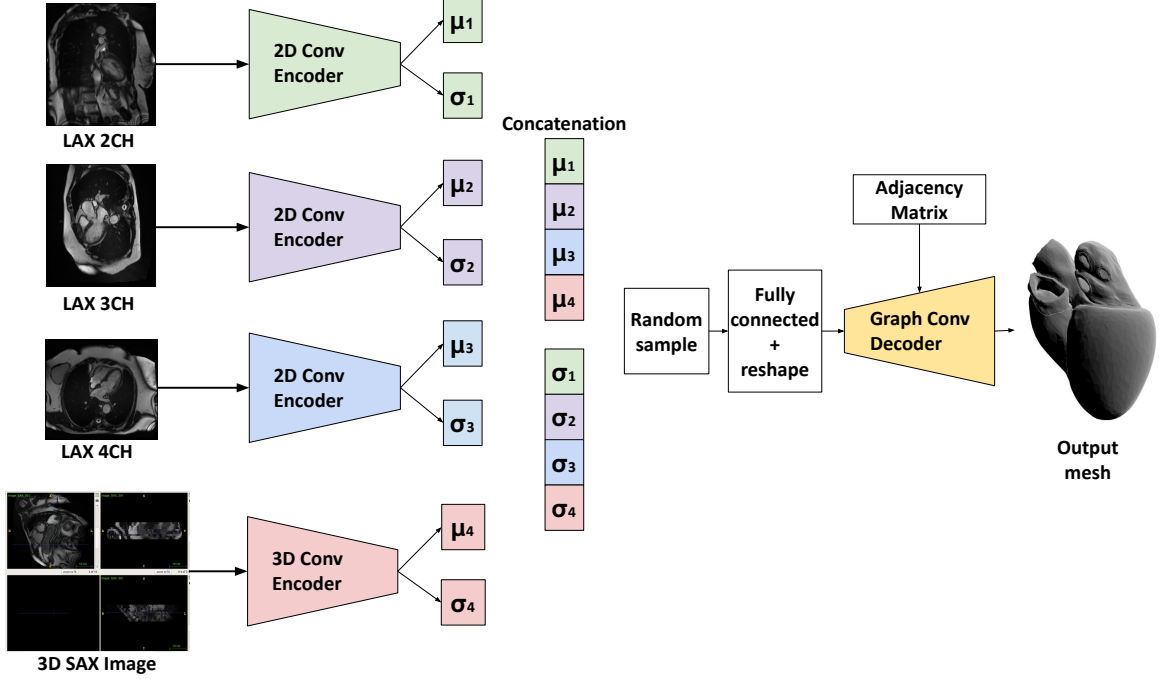


Figura 4.1: Arquitectura del modelo propuesto para la extracción de mallas cardíacas a partir de imágenes de resonancia magnética cardíaca.

La Figura 4.1 muestra el modelo propuesto, al que denominamos MV-HybridVNet, por *Multi-view Hybrid Volumetric Network*, de forma similar a la HybridGNet presentada en el capítulo anterior. MV-HybridVNet recibe varios tipos de imágenes de resonancia magnética cardíaca (CMR): la vista de eje corto (SAX, por *short axis*), una vista tridimensional en sección transversal adquirida perpendicular al eje largo, y tres vistas 2D de eje largo (LAX, por *long axis*) para dos, tres y cuatro cámaras del corazón (LAX 2CH, LAX 3CH y LAX 4CH, respectivamente). Estas vistas 2D se adquieren en paralelo al eje largo. Con estas cuatro imágenes como entrada, nuestro objetivo es generar una malla que represente las estructuras cardíacas de interés, ya sea una malla superficial o tetraédrica.

De manera similar al capítulo anterior, formulamos nuestro conjunto de datos $\mathcal{D} = \{(\mathbf{I}, \mathbf{G})_n\}_{0 \leq n \leq N}$, que consta de N conjuntos de imágenes CMR en múltiples vistas $\mathbf{I} = (\mathbf{I}^{\text{LAX 2CH}}, \mathbf{I}^{\text{LAX 3CH}}, \mathbf{I}^{\text{LAX 4CH}}, \mathbf{I}^{\text{SAX}})$ junto con sus correspondientes mallas cardíacas representadas como un grafo $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$. Aquí, V es el conjunto de vértices de M nodos, $\mathbf{A} \in \{0, 1\}^{M \times M}$ es la matriz de adyacencia que indica la conectividad entre pares de vértices, y $\mathbf{X} \in \mathbb{R}^{M \times s}$ es una función que asigna un vector de características a cada vértice ($s = 3$ para coordenadas espaciales tridimensionales). Por construcción del dataset, como veremos en la sección siguiente, asumiremos que V y $\mathbf{A}$ son iguales para todos los individuos del conjunto de datos, siendo la única diferencia entre ellos la posición espacial de los nodos definida en $\mathbf{X}$. De esta manera, nuestra formulación utiliza convoluciones espectrales en grafos, siguiendo nuevamente el trabajo de [22, 79]. Esta formulación nos permite utilizar la misma arquitectura para mallas superficiales o tetraédricas, simplemente cambiando el número M de nodos en V y su matriz de conectividad $\mathbf{A}$, sin necesidad de modificar la arquitectura.

El modelo consta de un autocodificador variacional, similar al presentado en el capítulo anterior, pero adaptado para múltiples datos de entrada. Los datos de entrada pasan a través de un codificador, f_e^I , que aprende una representación latente de la estructura cardíaca. Luego, el decodificador de convoluciones espectrales en grafo, f_d^G , genera una malla que representa el órgano. Dado que nuestras

entradas consisten en cuatro imágenes diferentes con tamaños y vistas distintas, empleamos lo que denominamos un codificador multi vista. Creamos cuatro codificadores independientes, cada uno especializado en una vista, y concatenamos sus salidas para aprender un espacio latente conjunto $\mathbf{z}$. Para las imágenes de eje largo, utilizamos el mismo codificador bidimensional que en el capítulo anterior, obteniendo $f_e^{\text{LAX } 2\text{CH}}$, $f_e^{\text{LAX } 3\text{CH}}$ y $f_e^{\text{LAX } 4\text{CH}}$. Para la imagen tridimensional de eje corto, utilizamos un codificador convolucional 3D, f_e^{SAX} , compuesto por bloques residuales tridimensionales intercalados con maxpoolings. Esto nos permite definir el codificador multi vista de la siguiente manera:

$$f_e^I = [f_e^{\text{LAX } 2\text{CH}}, f_e^{\text{LAX } 3\text{CH}}, f_e^{\text{LAX } 4\text{CH}}, f_e^{\text{SAX}}]. \quad (4.1)$$

Definimos el decodificador, f_d^G , utilizando bloques de convoluciones espectrales en grafo de la misma manera que en el capítulo anterior. Cada bloque convolucional consta de una convolución, una función de activación no lineal (ReLU), y una capa de normalización por capas (*Layer Normalization*, en ingles) [3]. Para capturar información a diferentes resoluciones, empleamos 5 de estos bloques, separados por capas de sobremuestreo.

En contraste con los contornos en dos dimensiones, cuya estructura se asemeja a un grafo circulante, la construcción de capas de muestreo y sobremuestreo para mallas triangulares o tetraédricas es un desafío. Adoptamos la técnica propuesta por Ranjan et al. [79], que nos permite abordar esta problemática. El proceso comienza con la estimación de una matriz de muestreo a través de un algoritmo iterativo de contracción de pares de vértices. El objetivo es reducir el número de vértices minimizando el error de aproximación entre superficies. Simultáneamente, este algoritmo genera una matriz de sobremuestreo para revertir este proceso y volver a la resolución original. Es importante destacar que este proceso es computacionalmente costoso, llevando entre minutos y horas para reducir el tamaño de una malla. Por lo tanto, optamos por realizarlo una sola vez en la malla atlas utilizada para construir la base de datos. Generamos 4 pares de capas de muestreo y sobremuestreo, cada uno reduciendo o aumentando a la mitad la cantidad de nodos, lo que nos proporciona una resolución mínima de 1/16 de la cantidad total de nodos de la salida. Las matrices de sobremuestreo permanecen fijas para el decodificador, independientemente del dato de entrada. Cada salida en resoluciones intermedias se somete a un proceso de supervisión profunda [54], donde estimamos versiones reducidas de las mallas *ground truth* durante el entrenamiento. Esto se logra utilizando las capas de muestreo fijas aprendidas previamente, con el objetivo de reducir el error de reconstrucción en cada nivel.

Además, se implementa una versión de una sola vista de nuestro enfoque propuesto, denominada simplemente HybridVNet, que utiliza exclusivamente las imágenes de eje corto como entrada, resultando en un modelo similar al enfoque presentado en el capítulo anterior. La Sección 3.2.3 del capítulo anterior proporciona más detalles sobre la formulación variacional y el entrenamiento del modelo, que se basa en una función de costo que utiliza el error cuadrático medio (MSE) y está regularizada mediante la divergencia de Kullback-Leibler. Para obtener más información sobre la arquitectura, consulte el Apéndice E, Sección II-B.

4.2.2. Evaluando diferentes términos de regularización

En nuestros experimentos, con el fin de obtener mallas suaves, utilizamos técnicas comunes del estado del arte como regularización de los vectores normales, del largo de las aristas, o mediante el suavizado laplaciano, detalladas en [93]. Sin embargo, este tipo de métricas, pensadas e implementadas en específico para mallas de superficie triangulares, no tienen en cuenta la estructura de elementos más complejos como los tetraedros en casos de mallas volúmetricas. Para superar esta limitación, proponemos una nueva función de regularización específicamente diseñada para mejorar la calidad de los elementos tetraédricos presentes en las mallas generadas por un modelo, durante el proceso de entrenamiento. La definimos de la siguiente manera:

$$\mathcal{L}_{ter} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{6} \sum_{j=1}^6 \left(\|\mathbf{e}_j^i\|_2 - \frac{1}{6} \sum_{k=1}^6 \|\mathbf{e}_k^i\|_2 \right)^2, \quad (4.2)$$

donde N_t es la cantidad de tetraedros de la malla, i representa el i -ésimo tetraedro, y $\mathbf{e}_j^i$ y $\mathbf{e}_k^i$ representan las aristas que forman el tetraedro. Este término fuerza la regularidad de los elementos

penalizando las aristas cuyo largo se aleje demasiado del largo promedio de las aristas del tetraedro al que pertenece.

4.3. Experimentos y resultados

4.3.1. Base de datos

Los datos para este estudio se recopilaron del UK Biobank (UKB). Los detalles del UKB se explican en [73], y el protocolo de adquisición de CMR se detalla en [74]. Realizamos nuestros experimentos en divisiones de entrenamiento/prueba de 4525 sujetos del UKB. Para asegurar una comparación justa con trabajos previos y facilitar la reproducibilidad de nuestros resultados, utilizamos las mismas divisiones de entrenamiento/prueba que [97], en las que se reservaron 600 sujetos como una división de prueba separada.

El conjunto de referencia de mallas de superficie utilizada en este estudio, al que referiremos como *ground truth*, se presentó en [97]. Se creó registrando un atlas de alta resolución del corazón humano [81]. Este atlas consta de un modelo de malla que incluye cinco estructuras separadas: el ventrículo izquierdo (LV), el ventrículo derecho (RV), la aurícula izquierda (LA), la aurícula derecha (RA) y la aorta. Por lo tanto, por construcción, cada sujeto contiene el mismo número de vértices y matriz de adyacencia que el atlas.

Dado que también queríamos probar el modelo en mallas volumétricas, generamos las anotaciones volumétricas a partir de las mallas de superficie, utilizando la correspondencia uno a uno entre los nodos de superficie para registrar un atlas volumétrico en el conjunto de datos de superficie. Utilizamos Simpleware's ScanIP [88] para construir la malla volumétrica del atlas, y luego registramos el atlas volumétrico en todo el conjunto de superficie utilizando deformaciones de malla a malla con el algoritmo thin-plate-splines, presente en la biblioteca Vedo [61].

Es importante notar que el procesamiento de las imágenes 3D de eje corto podía realizarse de dos maneras, estableciendo una región de interés centrada en el corazón, y descartando el resto de la información, como es común en la bibliografía [2, 97], o mediante la utilización de la imagen completa. En nuestros experimentos probamos ambas maneras, con un método que recibe la imagen completa, y otro que recibe un recorte.

Para más detalles sobre la construcción de las bases de datos, preprocesamiento de imágenes y de la construcción de la bases de datos volumétrica, ver el Apéndice E, Secciones II-A y III-A.

4.3.2. Modelo para comparación

Comparamos los resultados de los modelos propuestos con el modelo Multi-Cue Shape Inference Network (MCSI-Net) [97], que se considera el estado del arte en la generación de mallas cardíacas a partir de imágenes de CMR en este conjunto de datos en particular. MCSI-Net utiliza dos redes neuronales distintas: la primera infiere las coordenadas de la posición de la malla y un vector de rotación, mientras que la segunda es una CNN que infiere los parámetros de un modelo de distribución de puntos basado en PCA. Es importante destacar que MCSI-Net utiliza las mismas vistas de imágenes que nuestra MV-HybridVNet pero también incorpora información clínica de los pacientes durante el entrenamiento, lo cual constituye una diferencia importante en comparación con nuestro modelo.

4.3.3. Resultados en mallas de superficie

Para evaluar la calidad de las mallas cardíacas, utilizamos métricas basadas en mallas y máscaras. Generamos máscaras de segmentación densas a partir de las mallas superficiales y evaluamos métricas de segmentación clásicas como DSC, HD o la distancia promedio entre los contornos de referencia y los contornos predichos (MCD, por *mean contour distance*). Para evaluar las mallas completas utilizando métricas estándar, como el error cuadrático medio (MSE, por el inglés *mean squared error*), el error absoluto promedio (MAE *mean absolute error*) y la raíz del error cuadrático medio (RMSE, por el inglés *root mean squared error*) en milímetros, que utilizamos para comparar solo entre nuestros modelos.

Tabla 4.1: Resultados cuantitativos sobre de segmentación en mallas de superficie.

	Métricas	MCSI-Net SAX	HybridGNet3D		MCSI-Net SAX-LAX	MV-HybridGNet3D	
		Recorte	Imagen completa	Recorte	Recorte	Imagen Completa	Recorte
Malla completa	MAE ↓	-	2.56 (0.62)	2.43 (0.59)	-	2.26 (0.55)	2.18 (0.54)
	MSE ↓	-	12.20 (7.11)	11.27 (6.69)	-	9.29 (5.48)	8.80 (5.31)
	RMSE ↓	-	3.38 (0.89)	3.25 (0.86)	-	2.95 (0.76)	2.87 (0.76)
	DSC ↑	0.87 (0.05)	0.89 (0.05)	0.90 (0.04)	0.88 (0.05)	0.90 (0.04)	0.91 (0.04)
LV Endo	HD ↓	5.13 (1.97)	4.48 (1.32)	4.08 (1.22)	4.74 (1.75)	4.22 (1.22)	3.89 (1.18)
	MCD ↓	1.93 (0.83)	1.67 (0.55)	1.49 (0.49)	1.86 (0.79)	1.55 (0.51)	1.39 (0.46)
	DSC ↑	0.76 (0.09)	0.80 (0.06)	0.83 (0.05)	0.78 (0.08)	0.81 (0.05)	0.84 (0.04)
	HD ↓	5.31 (1.98)	4.71 (1.36)	4.23 (1.27)	4.75 (1.76)	4.40 (1.26)	3.96 (1.23)
LV Myo	MCD ↓	1.97 (0.95)	1.71 (0.56)	1.49 (0.51)	1.86 (0.82)	1.57 (0.52)	1.35 (0.46)
	DSC ↑	0.85 (0.06)	0.85 (0.05)	0.86 (0.05)	0.85 (0.06)	0.86 (0.05)	0.87 (0.05)
	HD ↓	7.11 (2.78)	6.97 (2.31)	6.44 (2.19)	7.06 (2.64)	6.79 (2.23)	6.13 (2.23)
	MCD ↓	2.34 (0.98)	2.10 (0.64)	1.90 (0.57)	2.27 (0.95)	1.99 (0.59)	1.76 (0.59)

Como se puede ver en la Tabla 4.1, en la comparación inicial, evaluamos nuestro modelo HybridVNet en dos configuraciones: una con imágenes completas y otra con versiones recortadas centradas en la estructura de interés, contra la MCSI-Net del estado del arte, en su versión que solo recibe imágenes SAX. Notablemente, HybridVNet supera a la MCSI-Net que solo utiliza imágenes SAX en todas las métricas y estructuras. Luego, comparamos nuestro modelo MV-HybridVNet con la MCSI-Net estándar, que también incorpora múltiples vistas. Los resultados confirman la superioridad de MV-HybridVNet en todas las métricas de segmentación, tanto para el ventrículo izquierdo como para el derecho. Es importante notar que nuestra variante de imagen completa logra resultados superiores respecto a la MCSI-Net en ambos casos, eliminando la necesidad de detectar la región de interés para realizar el recorte de la imagen. Además, MV-HybridVNet en recortes supera significativamente a las imágenes completas.

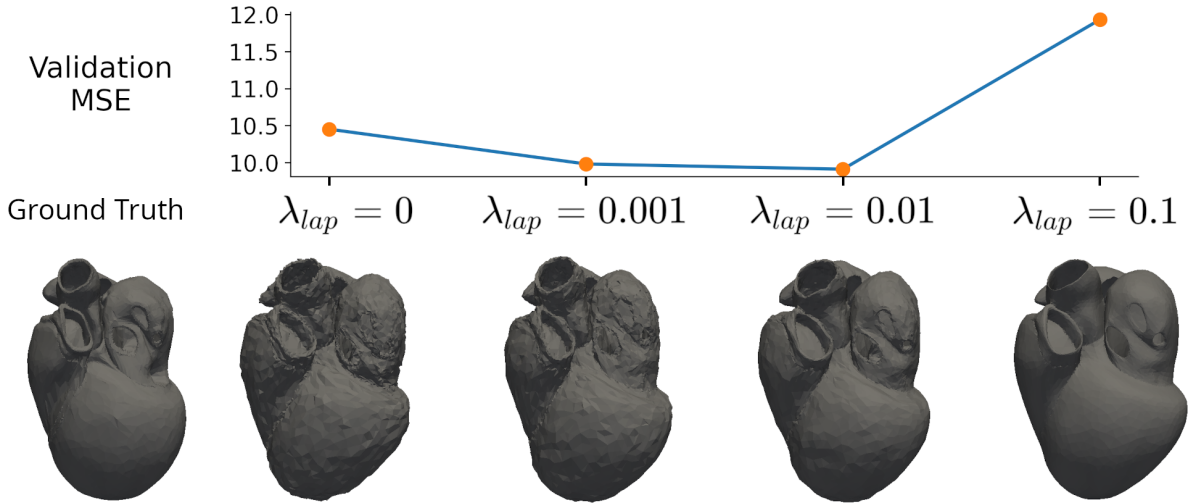


Figura 4.2: Análisis del suavizado laplaciano en mallas de superficie. Arriba: resultados cuantitativos en validación. Abajo: resultados cualitativos. Los mejores resultados se obtuvieron cuando $\lambda_{lap} = 0,01$.

En este contexto, se realizó además una evaluación exhaustiva de diversas funciones de pérdida de regularización de superficie para mejorar el rendimiento de nuestro modelo HybridVNet. La regularización de vectores normales y la regularización de la largo de aristas, aunque son comúnmente empleadas en tareas de regularización de mallas, no produjeron mejoras significativas en el rendimiento de nuestro modelo. En contraste, la incorporación del suavizado laplaciano produjo mallas de superficie notablemente más suaves y precisas. Esto se puede apreciar visualmente en la Figura 4.2, que presenta un análisis cualitativo de las mallas obtenidas a medida que se aumenta el peso en la función de pérdida para el suavizado laplaciano. Las figuras ilustran claramente la mayor suavidad y calidad de las mallas a medida que se ajusta el coeficiente de regularización.

Para un análisis más detallado de los resultados y de los términos de regularización en mallas de

superficie, ver el Apéndice E Sección IV-A.

4.3.4. Resultados en mallas volumétricas

Tabla 4.2: Resultados cuantitativos de segmentación en mallas volumétricas.

	Metrics	MV-HybridVNet			
		$\lambda_{ter} = 0$	$\lambda_{ter} = 1e - 4$	$\lambda_{ter} = 1e - 3$	$\lambda_{ter} = 1e - 2$
Mesh	MAE ↓	2.08 (0.63)	2.07 (0.64)	2.04 (0.61)	2.11 (0.61)
	MSE ↓	8.25 (6.14)	8.22 (6.12)	7.93 (5.63)	8.39 (6.00)
	RMSE ↓	2.74 (0.88)	2.73 (0.88)	2.69 (0.84)	2.77 (0.84)
LV Endo	DSC ↑	0.90 (0.04)	0.90 (0.04)	0.90 (0.05)	0.88 (0.05)
	HD ↓	4.36 (1.22)	4.32 (1.24)	4.41 (1.35)	5.21 (1.42)
	MCD ↓	1.52 (0.46)	1.51 (0.49)	1.58 (0.54)	1.89 (0.62)
LV Myo	DSC ↑	0.78 (0.04)	0.78 (0.04)	0.76 (0.05)	0.74 (0.06)
	HD ↓	5.27 (1.47)	4.98 (1.40)	5.17 (1.50)	5.30 (1.57)
	MCD ↓	1.86 (0.61)	1.81 (0.64)	1.95 (0.72)	1.96 (0.77)
RV Endo	DSC ↑	0.85 (0.06)	0.86 (0.05)	0.85 (0.05)	0.85 (0.06)
	HD ↓	7.22 (2.76)	6.97 (2.54)	7.38 (2.67)	7.55 (2.80)
	MCD ↓	2.05 (0.64)	2.02 (0.63)	2.09 (0.64)	2.13 (0.69)

Tabla 4.3: Resultados cuantitativos del análisis de calidad de los tetraedros en mallas volumétricas.

		Promedio	Desvío	Mínimo	Máximo	1 %	5 %	25 %	50 %	75 %
Mallas de referencia	Atlas	0.491	0.174	0.092	0.984	0.115	0.194	0.367	0.494	0.617
	Ground truth	0.355	0.156	-0.207	0.838	0.04	0.103	0.238	0.353	0.47
	Simpleware	0.524	0.185	0.064	0.992	0.128	0.202	0.387	0.535	0.667
MV-HybridGNet3D	$\lambda_{ter} = 0$	0.222	0.225	-0.759	0.876	-0.327	-0.144	0.065	0.219	0.384
	$\lambda_{ter} = 1e - 4$	0.229	0.23	-0.771	0.871	-0.337	-0.151	0.068	0.231	0.397
	$\lambda_{ter} = 1e - 3$	0.433	0.206	-0.719	0.904	-0.138	0.059	0.307	0.457	0.585
	$\lambda_{ter} = 1e - 2$	0.501	0.309	-0.931	0.943	-0.681	-0.298	0.434	0.577	0.688

En nuestro segundo experimento, nos centramos en la creación de mallas tetraédricas. Evaluamos varios factores de ponderación (λ_{ter}) para el término de regularización de elementos tetraédricos definido en la Eq. 4.2, con el objetivo de comprender su influencia tanto en la calidad de la malla como en el rendimiento de la segmentación del ventrículo. La Tabla 4.2 presenta los resultados, incluyendo métricas para la calidad de la malla y la segmentación del ventrículo, en diferentes valores de λ_{ter} .

Los resultados demostraron una relación entre λ_{ter} y el rendimiento del modelo. Para las métricas de segmentación, se obtienen los mejores resultados cuando $\lambda_{ter} = 1e - 4$, lo que genera resultados similares al modelo no regularizado, especialmente en términos de métricas del Endocardio del Ventrículo Izquierdo (LV Endo). Sin embargo, para las métricas relacionadas con el rendimiento de la calidad de la malla, $\lambda_{ter} = 1e - 3$ produce los mejores resultados, a pesar de una ligera disminución en el rendimiento en términos de métricas de segmentación.

Para evaluar la calidad de los elementos tetraédricos, utilizamos la métrica ampliamente adoptada llamada *Scaled Jacobian*. Esta métrica, que se mueve en el rango (-1,1), proporciona una medida cuantitativa de regularidad y simetría, con valores altos indicando alta calidad. Los resultados demuestran que, a partir del percentil 25 %, nuestros modelos regularizados superan a los elementos de referencia para $\lambda_{ter} = 1e - 3$ y mayores, respaldando nuestra hipótesis de que la pérdida de regularización mejora significativamente la calidad de la malla.

En comparación con la conversión directa de mallas superficiales en volumétricas, nuestro modelo muestra resultados competitivos y aborda desafíos asociados con esta conversión. Además, supera notablemente en velocidad al procedimiento típico, lo que resulta en un notable aumento de eficiencia

al generar el conjunto de vértices para una malla volumétrica. Para un análisis más detallado de los resultados y del término de regularización, ver el Apéndice E Sección IV-B.

Sección 5

Generación de anotaciones para grandes bases de datos a partir de datos multi-céntricos y métodos de validación automática de la calidad

Una de las principales limitaciones identificadas en esta tesis es la escasez de bases de datos adecuadas para el entrenamiento de modelos como los propuestos. Para abordar este desafío, se propone la obtención de anotaciones para generar una extensa base de datos anotada. Aunque existen numerosas bases de datos de imágenes de rayos X disponibles, es importante destacar que la mayoría no incluye anotaciones en forma de máscaras de segmentación anatómica, y cuando lo hacen, las anotaciones varían, no siendo uniformes en cuanto a los órganos que se abordan. En este capítulo, abordaremos la creación de una nueva base de datos de anotaciones de segmentación, aprovechando las principales bases de datos de rayos X disponibles, las cuales provienen de diversos centros clínicos ubicados en distintas partes del mundo. Como se discutió previamente en el Capítulo 3, Sección 3.3.4, esta diversidad de fuentes presenta desafíos significativos relacionados con el cambio de dominio. Sin embargo, hemos observado que los modelos basados en landmarks representan una sólida alternativa para abordar este problema, como se evidenciará en este capítulo. Finalmente, con el objetivo de brindar una métrica objetiva que indique la calidad de las segmentaciones automáticas generadas, trabajaremos en el desarrollo de métodos de control de calidad en ausencia de anotaciones de referencia o *ground truth*.

5.1. Antecedentes

El desarrollo de métodos de aprendizaje automático para la segmentación de imágenes médicas requiere el uso de extensas bases de datos, las cuales a menudo combinan imágenes provenientes de diversos centros médicos, también llamadas multi-céntricas. Normalmente, cada uno de estos centros proporciona diferentes tipos de etiquetas, conformando lo que se conoce como anotaciones heterogéneas (ver Figura 5.1). Esto plantea dos desafíos principales. En primer lugar, las bases de datos con anotaciones procedentes de distintos centros médicos pueden dar lugar a problemas de cambio de dominio [34], debido a las variaciones en la intensidad causadas por diferencias en los dispositivos de adquisición o en los protocolos de adquisición. En segundo lugar, la heterogeneidad de los datos dificulta el entrenamiento de modelos, ya que no todas las imágenes cuentan con anotaciones para el mismo conjunto de órganos, generando un problema de datos faltantes [47]. Como consecuencia, los modelos de segmentación a nivel pixel como UNet y nnUNet entrenados en estas bases de datos tienden a memorizar qué conjunto de etiquetas está presente en qué centro médico. Así, al momento de realizar una predicción en una nueva imagen de alguno de dichos centros, sólo generará las etiquetas que fueron vistas allí, pese a que otras estructuras se encuentren presentes.

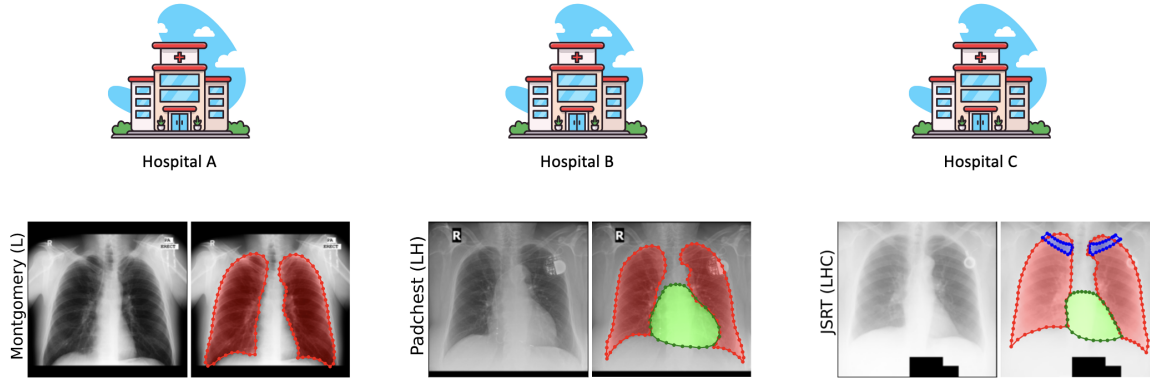


Figura 5.1: **Bases de datos multi-céntricas con anotaciones heterogéneas.** Ejemplo de imágenes provenientes de distintos centros médicos y con anotaciones heterogéneas, donde el Hospital A solo incluye pulmones, el B pulmones y corazón, y el C agrega además las clavículas.

La resolución conjunta de estos dos problemas se encuentra en la intersección de diversas áreas, como el aprendizaje con múltiples objetivos (*multi-task learning*), la adaptación de dominio y el aprendizaje débilmente supervisado (*weakly supervised learning*) [23]. La literatura ha propuesto varios métodos para abordar los problemas de cambio de dominio [34, 70, 83] y etiquetas heterogéneas [75, 47, 26] en la segmentación de imágenes médicas por separado. Aquí nos centraremos en el abordaje de ambos problemas al mismo tiempo.

En lo que respecta a la evaluación de la calidad de las segmentaciones en ausencia de *ground truth*, se han propuesto diversos enfoques. Algunos se basan en la incertidumbre predictiva (*predictive uncertainty*) [21, 20], asumiendo que las predicciones altamente inciertas se correlacionarán con píxeles erróneos. Sin embargo, estos métodos dependen en gran medida de la calidad de las estimaciones de incertidumbre proporcionadas por el modelo de segmentación. Por otro lado, en [27] se propuso un enfoque basado en el aprendizaje, donde se entrena una CNN para predecir el valor de DSC a partir de pares de imágenes con las correspondientes segmentaciones predichas. Este enfoque es independiente del modelo de segmentación, pero requiere entrenar una CNN adicional para predecir el valor de DSC.

En este capítulo, basaremos nuestra investigación en un enfoque introducido previamente por Valindria et al. [91], conocido como RCA. Este método propone la construcción de un clasificador inverso utilizando solo una imagen de prueba y su segmentación predicha como *pseudo ground truth*. Luego, evaluaremos este clasificador en un conjunto de datos de referencia que contiene segmentaciones disponibles, utilizando su rendimiento como una aproximación de la calidad de la segmentación predicha. Se han propuesto diferentes variantes de RCA, incluyendo *atlas-forest*, RCA basado en CNNs y RCA basado en propagación de etiquetas mediante atlas utilizando métodos de registración clásicos. En este capítulo, implementaremos nuestra propia variante de aprendizaje profundo de RCA basado en la propagación de etiquetas mediante el uso de varios atlas, que ha demostrado ser rápida y confiable en trabajos anteriores [56, 57].

5.2. Métodos propuestos

5.2.1. Descripción del problema

Nos enfrentamos a la tarea de entrenar modelos de segmentación utilizando imágenes provenientes de centros médicos distintos. Esto implica trabajar con un conjunto de datos $\mathcal{D}$ compuesto por n subconjuntos $\mathcal{D}_n$, cada uno conteniendo sus propias imágenes $\mathbf{I}_n$ y las correspondientes segmentaciones basadas en landmarks representadas como grafos $\mathbf{G}_n$. Cada conjunto de datos tiene un número fijo de nodos que representan las estructuras anatómicas de los pulmones, el corazón y las clavículas. Sin embargo, la disponibilidad de anotaciones varía según el centro clínico específico.

Además, disponemos de otros conjuntos de prueba, que llamaremos $\mathcal{D}_p$, para los cuales solo tenemos imágenes $\mathbf{I}$ y sus respectivas segmentaciones $S_{\mathbf{I}}$ generadas por un modelo de segmentación. En este caso, nuestro objetivo es estimar la calidad de estas segmentaciones.

5.2.2. Entrenando modelos basados en landmark con datos heterogéneos

Los modelos basados en landmarks, como la HybridGNet, proporcionan una manera natural de ser entrenados cuando los datos son heterogéneos, utilizando indexación. El modelo genera una matriz de $D \times 2$, donde D representa la cantidad fija de nodos y está secuencialmente ordenada: primero los nodos de los pulmones t_L (L por *lungs*), luego los del corazón t_H (H por *heart*) y finalmente los de las clavículas t_C (C por *clavicles*). Esto se corresponde con las etiquetas *ground truth target* = $[t_L, t_H, t_C]$ para el conjunto de datos que incluye todas las anotaciones. El tamaño de cada subconjunto es D_L , D_H y D_C , respectivamente.

Para entrenar el modelo, separamos los distintos conjuntos de datos y, en cada iteración, trabajamos con un lote de uno de los conjuntos de forma aleatoria. Esto garantiza que cada lote contenga solo un tipo de anotación. Por ejemplo, si un lote contiene imágenes con anotaciones solo para los pulmones, el modelo generará predicciones para todos los órganos, pero solo evaluaremos y retropropagaremos la función de costo en los primeros D_L nodos, sin retropropagar errores para el corazón y las clavículas. Este enfoque nos permite entrenar eficazmente una HybridGNet o cualquier modelo basado en landmarks.

Modelos basados en píxeles para comparación

Tal como se mencionó anteriormente, los modelos basados en píxeles y entrenados de forma estándar en escenarios multi-céntricos con etiquetas heterogéneas, tienden a memorizar qué set de etiquetas proviene de cada centro, tal como se puede observar en la Figura 5.3. Para realizar una evaluación justa, compararemos nuestra HybridGNet con modelos del estado del arte como la nnUNet [42] y la UNet estándar [82]. En ambos casos, utilizaremos las máscaras disponibles según las anotaciones de cada conjunto de datos. Con el objetivo de mantener la comparación equitativa, también entrenaremos una UNet siguiendo el mismo enfoque de entrenamiento heterogéneo (HT) que aplicamos a la HybridGNet, pero ignorando mapas de segmentación a nivel pixel en lugar de nodos en particular. En este modelo, generaremos salidas independientes para las máscaras de pulmón, corazón y clavícula, utilizando funciones de pérdida binarias e independientes para cada máscara. Así, sólo retropropagaremos la función de costo para las anotaciones disponibles en cada lote, en lugar de utilizar funciones multi-clase, evitando confundir a la red. Para obtener más detalles sobre la implementación de los modelos, consultar el Apéndice F, Sección 2.

5.2.3. Evaluación automática de la calidad mediante *reverse classification accuracy*

Una vez que hemos entrenado un modelo y segmentado una base de datos $\mathcal{D}_{prueba}$, el proceso de evaluación automática de la calidad comienza con una imagen $\mathbf{I}$ de dicha base de datos y su correspondiente segmentación $S_{\mathbf{I}}$, para la cual deseamos estimar la calidad. A esta imagen la denominaremos *atlas*. Por otro lado, contamos con la base de datos de referencia (utilizada para entrenar el modelo de segmentación) que contiene imágenes J_i junto con sus respectivas máscaras de segmentación ground-truth $S_{J_i}^{GT}$. La intuición detrás del método de estimación automática de calidad RCA es la siguiente: si una segmentación estimada es precisa, entonces se la podrá usar como atlas para segmentar otras imágenes, para las cuales sí se posee anotaciones de referencia. La calidad de las segmentaciones obtenidas en ese proceso (que puede ser medida sobre las imágenes de referencia ya que poseen anotaciones *ground truth*) actuará como *proxy* para determinar la calidad de la segmentación predicha.

El proceso de evaluación comienza aplicando una registración deformable [28] para alinear el atlas con las imágenes anotadas en la base de datos de referencia. Este proceso consta de un paso de registración afín inicial, seguido por la aplicación de un campo de deformación D_i que alinea las estructuras anatómicas. Dichas transformaciones son también aplicadas a la segmentación predicha

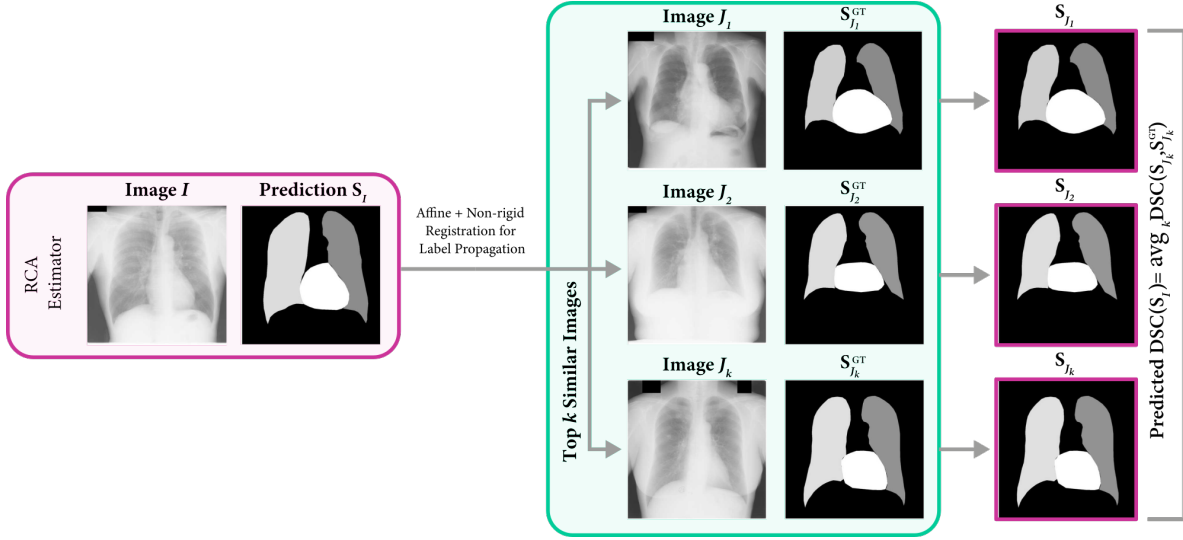


Figura 5.2: Metodología propuesta para la evaluación automática de calidad mediante RCA.

S_I , con el objetivo de propagar estas etiquetas y poder evaluar su calidad al compararla con las segmentaciones de referencia.

Posteriormente, evaluamos la calidad de estas máscaras propagadas en comparación con las máscaras ground-truth utilizando como métrica el DSC. Para obtener una métrica global, agregamos estos resultados utilizando una operación que puede ser el máximo o el promedio. En nuestra implementación, utilizamos el promedio, como se muestra en la sección de resultados donde validamos este método.

Para acelerar la evaluación automática y como una modificación al marco de trabajo clásico de propagación de etiquetas basado en atlas, propusimos utilizar métodos basados en aprendizaje profundo. Además, restringimos las imágenes de las bases de referencia utilizadas en cada iteración sólo a las top- k imágenes que son más similares al atlas que deseamos evaluar para acelerar dicho proceso, tal como se propuso en [56]. Este enfoque, que se ilustra en la Figura 5.2, se fundamenta en la suposición de que si la segmentación es de alta calidad, las máscaras estimadas por la propagación de etiquetas deberían tener buenos resultados en las imágenes de referencia. De esta manera, podemos evaluar la calidad de una segmentación sin necesidad de contar con su máscara de referencia. Para obtener más detalles sobre estos procesos, consultar el Apéndice G, Sección *Reverse Classification Accuracy to Measure Segmentation Quality*.

5.3. Experimentos y Resultados

5.3.1. Bases de Datos

Para entrenar nuestros modelos en la segmentación de datos heterogéneos y llevar a cabo la validación automática de la calidad, utilizamos la base de datos que generamos en el Capítulo 3 y describimos en detalle en la Sección 3.3.5. Esta base de datos está compuesta por 911 imágenes obtenidas de conjuntos de datos como JSRT [85] (con anotaciones de pulmones, corazón y clavícula), un subconjunto de Padchest [11] (con anotaciones de pulmones y corazón), Montgomery [13] (solo pulmón) y Shenzhen [43] (solo pulmón).

Por otro lado, para crear nuestra base de datos a gran escala, que denominaremos CheXmask, combinamos varios conjuntos de datos públicos de rayos X que no incluyen segmentaciones anatómicas. Estos son CANDID-PTX [25], Chest x-ray8 [94], CheXpert [41], MIMIC-CXR-JPG [44], Padchest [11] y VinDr-CXR [65]. Cada uno de estos conjuntos de datos contiene entre 18.000 y 300.000 imágenes, lo que suma un total superior a los 650 mil imágenes. Esto representa más de 650 veces la cantidad

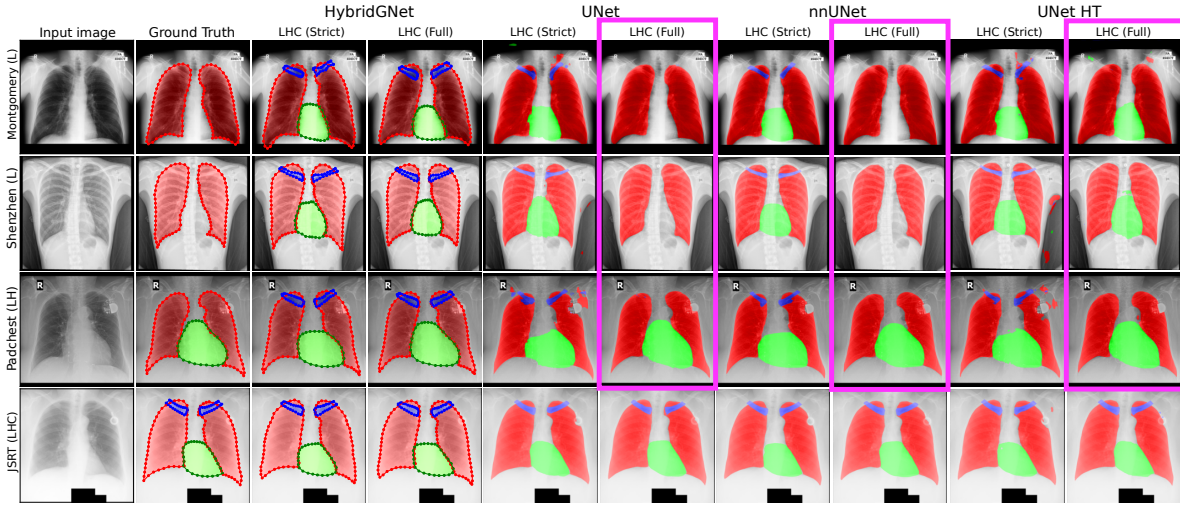


Figura 5.3: **Evaluación cualitativa.** Ejemplos de los 4 datasets (filas), segmentados por los distintos métodos en las distintas configuraciones de anotación (columnas).

de datos de nuestra base de datos original, incluyendo una mayor diversidad geográfica y diferentes condiciones clínicas, demográficas y patológicas. Dado que presentan heterogeneidad en formatos, vistas y tamaños de imágenes, realizamos un proceso de selección y preprocesamiento de datos exhaustivo, que se describe en detalle en la sección *Data Preparation* del Apéndice G.

5.3.2. Entrenando modelos con datos heterogéneos

Para evaluar el rendimiento de los modelos, creamos dos contextos de evaluación dado que no todas las etiquetas están disponibles para cada conjunto de datos: uno estricto (*Strict*) y otro completo (*Full*). En el estricto, excluimos las imágenes que no contienen todas las etiquetas necesarias, lo que lleva a descartar muchos datos con etiquetas incompletas. En el completo, utilizamos todos los datos disponibles.

La Figura 5.3 presenta uno de nuestros hallazgos clave: cuando entrenamos con datos heterogéneos de diferentes conjuntos de datos, los modelos basados en píxeles tienden a aprender a segmentar según el conjunto de datos de origen de las imágenes (marcado en rosa). En cambio, la UNet heterogénea (HT) solo segmenta las clases que no entran en conflicto y aprende a distinguir entre conjuntos de datos si hay clases en conflicto. En este contexto, decimos que las clases están en conflicto cuando un píxel se clasifica de una manera en un conjunto de datos y de otra manera en otro, como en el caso de las clavículas y el pulmón. Por otro lado, la HybridGNet siempre predice todas las clases.

Dado que este fenómeno se observa en todas las muestras de cada conjunto de datos, los modelos basados en píxeles parecen memorizar qué estructura se ha anotado en cada conjunto de datos, utilizando las diferentes distribuciones de intensidad de cada centro clínico para decidir qué anotaciones generar durante las pruebas. Para comprender mejor este fenómeno, realizamos una reducción de dimensionalidad de las características obtenidas en el espacio latente de la HybridGNet y en las capas intermedias de las UNet, utilizando la técnica de reducción de dimensionalidad UMAP [58] en los modelos entrenados en configuraciones heterogéneas.

La Figura 5.4 (columna izquierda) muestra la proyección 2D de las imágenes de todos los conjuntos de datos para cada modelo. Es evidente una agrupación por parte de los modelos UNet. Dado que el conjunto de datos JSRT contiene imágenes con pulmones significativamente más grandes que el resto (debido a la distancia de adquisición), también llevamos a cabo experimentos en los que escalamos todas las imágenes para que tengan el mismo área pulmonar antes de realizar las predicciones, con el fin de eliminar los efectos asociados a esta información. Esto se muestra en la columna derecha de la Figura 5.4, donde la agrupación se elimina por completo en el espacio latente de la HybridGNet, pero no afecta en absoluto a los modelos UNet. El hecho de que este experimento elimine la agrupación

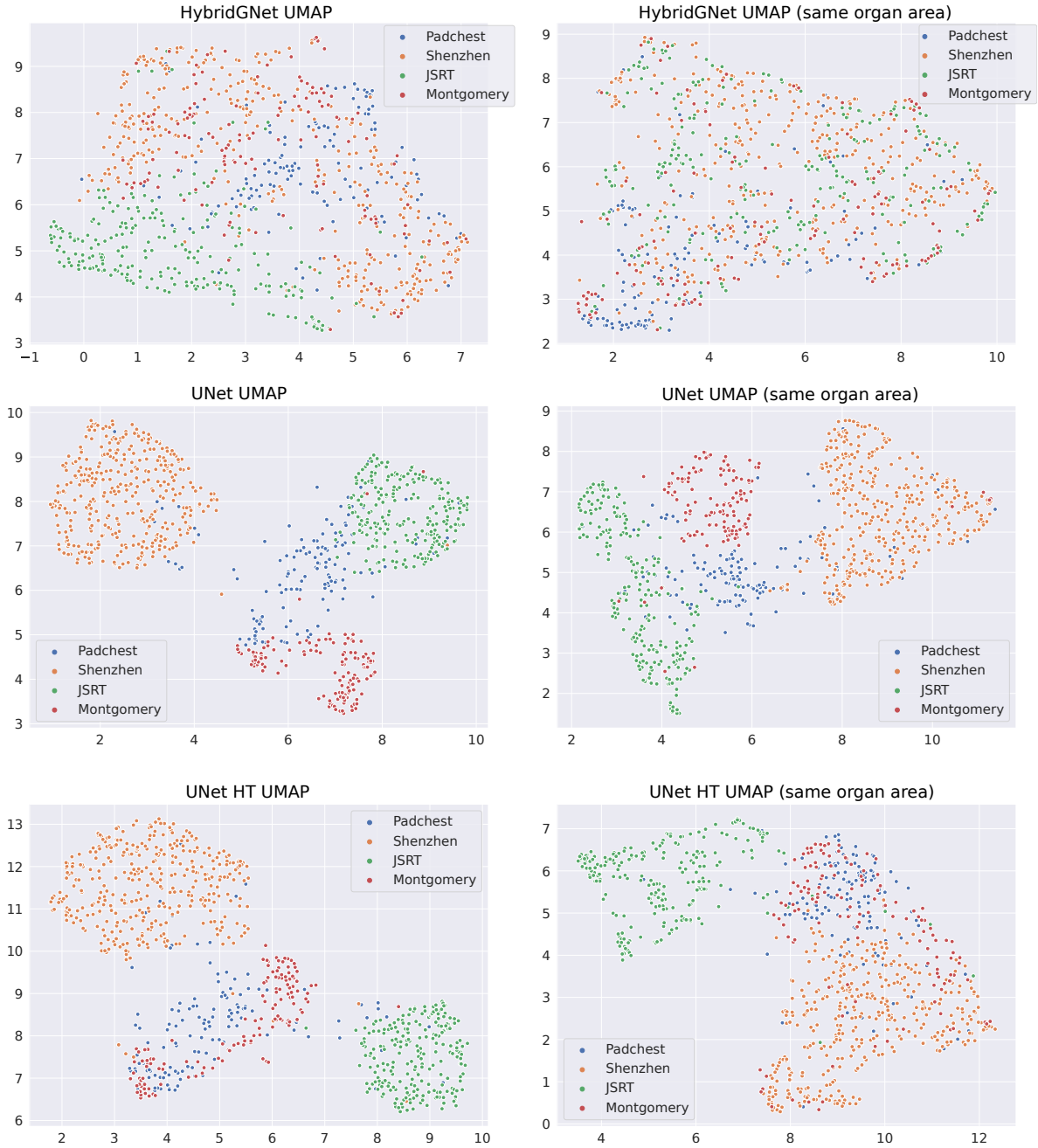


Figura 5.4: **Inspección del espacio latente mediante UMAP:** Diferentes conjuntos de datos se muestran en colores, lo que permite ver cómo los modelos UNet y UNet HT tienden a agrupar las imágenes por conjunto de datos, mientras que HybridGNet no lo hace, lo que explica la mayor robustez ante la memorización de etiquetas de dominio.

presente en JSRT en la HybridGNet proporciona información sobre el tipo de conocimiento que el modelo adquiere, generando agrupaciones que se relacionan más con la forma o el tamaño y evitando agrupar imágenes en función de las diferencias de intensidad.

En el Apéndice F se detallan en profundidad estos resultados, en conjunto con otros experimentos cuantitativos adicionales que muestran la superioridad de la HybridGNet frente a las arquitecturas basadas en píxeles en estos escenarios.

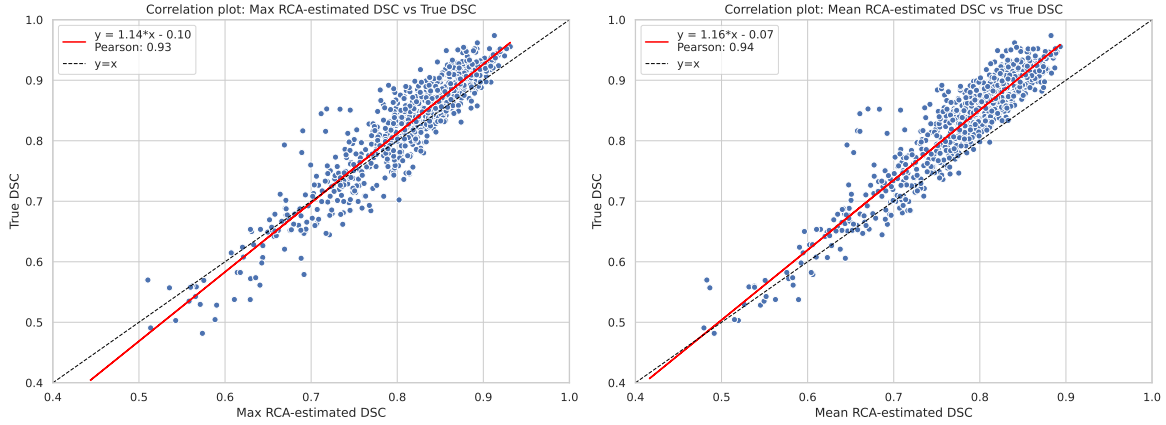


Figura 5.5: Correlación de los valores de RCA contra los valores reales de DSC en la partición de prueba

5.3.3. Validación del método de evaluación de calidad mediante *reverse classification accuracy*

Para validar el método de evaluación de calidad propuesto a través de RCA, se llevaron a cabo experimentos separando la base de datos en un 80 % de entrenamiento y un 20 % de prueba, utilizando solo las anotaciones de los pulmones para registrar. Dado que el proceso requiere una registración deformable para alinear el atlas con las imágenes de referencia, se entrenó primero un paso de registración afín seguido del entrenamiento de un modelo deformable que aprendiera un campo de deformación para lograr una alineación precisa en términos anatómicos.

Posteriormente, tomando cada imagen del conjunto de pruebas del 20 % como un atlas, se propagaron sus máscaras de segmentación a las 5 imágenes de referencia más parecidas mediante la registración deformable. Luego, se calculó el DSC entre las máscaras propagadas y las máscaras de referencia, utilizando el método de agregación promedio o máximo.

Para validar la precisión de esta metodología, se entrenaron modelos UNet en el mismo conjunto de datos con diferentes duraciones de entrenamiento, lo que resultó en segmentaciones de calidad variada. Luego, se evaluó la calidad de estas segmentaciones variadas obtenidas por RCA en comparación con la calidad real. La correlación de Pearson entre el DSC estimado mediante RCA y el DSC real en estas segmentaciones simuladas fue de 0.93 para el valor máximo y 0.94 para el valor promedio. Esto demostró una fuerte correlación entre la estimación de calidad mediante RCA y la calidad real de las segmentaciones. Para más detalles técnicos sobre esta validación, ver Apéndice G.

5.3.4. Segmentación masiva de rayos X utilizando HybridGNet

Luego de investigar los efectos de la memorización y la robustez de HybridGNet frente a variaciones como el cambio de dominio u oclusiones, se decidió emplear este modelo para llevar a cabo la segmentación masiva de los seis conjuntos de datos previamente mencionados. Se utilizó el conjunto de entrenamiento completo para entrenar un modelo heterogéneo con el fin de generar segmentaciones de pulmones y corazón. Se empleó la mejor configuración de hiperparámetros obtenida en validaciones previas y se fijó el número de épocas de entrenamiento para evitar dividir los datos en conjuntos de entrenamiento, validación y prueba, y, de este modo, no perder información.

Las anotaciones generadas se evaluaron mediante tres estrategias diferentes: validación clínica, validación automática con RCA y el entrenamiento de modelos sobre estas imágenes de prueba, explorando su utilidad como conjunto de datos de referencia para el entrenamiento de modelos de segmentación.

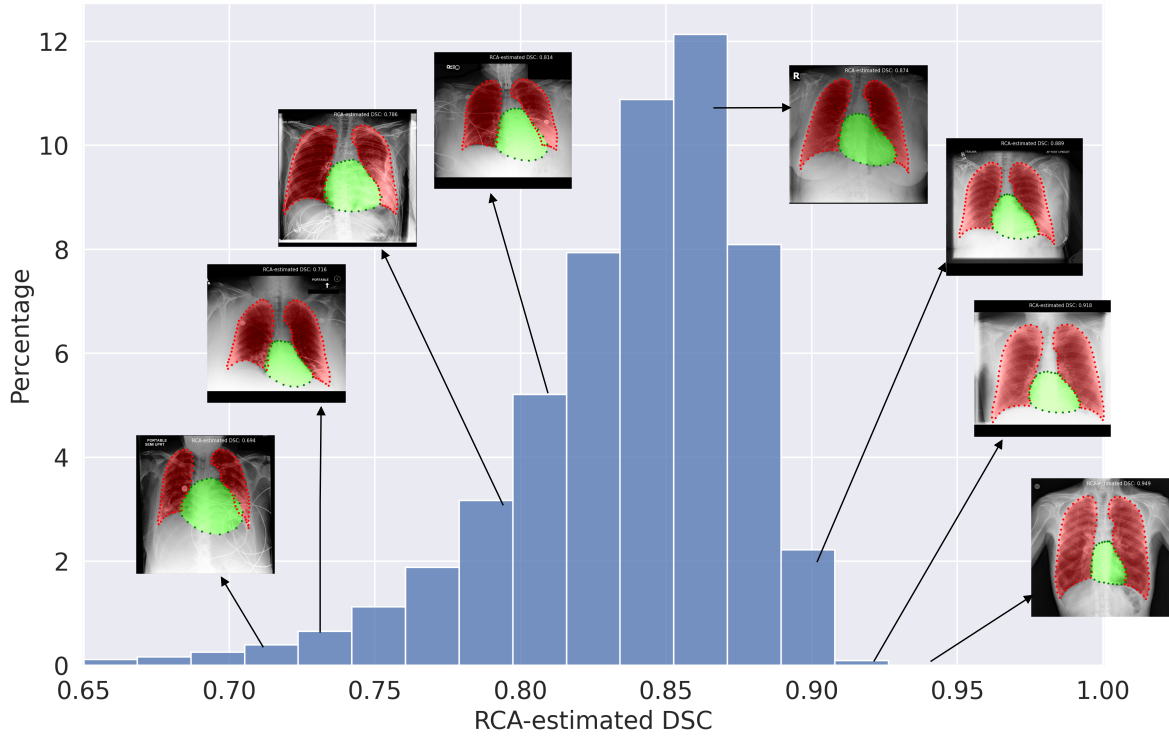


Figura 5.6: Histograma de DSC estimado por RCA para la base de datos CheXmask completa

Validación clínica

Se creó un conjunto de máscaras de referencia etiquetadas por una radióloga experimentada para evaluar la calidad de las segmentaciones de pulmones y corazón en CheXmask. Para ello, se realizó una revisión manual de las segmentaciones basadas en landmarks del pulmón derecho, pulmón izquierdo y corazón. Se tomaron 30 imágenes de cada conjunto de datos, lo que resultó en un conjunto de referencia final de 180 imágenes con revisión manual.

La Tabla 5.1 resume los resultados de la validación manual realizada para la segmentación de pulmones y corazón en los seis conjuntos de datos en subconjuntos de 30 muestras. Se informan métricas de evaluación como el DSC, la HD y la HD al 95 % (HD95). Estos resultados muestran, en general, una alta concordancia entre las anotaciones proporcionadas y los resultados generados por el modelo para este subconjunto de imágenes. Para obtener más detalles sobre estos resultados, consulte el Apéndice G, Sección *Technical Validation*.

Tabla 5.1: Resultados de la validación manual clínica, con desvíos estandar ($n = 30$ por dataset)

Dataset	Lungs			Heart		
	DSC	HD	HD95	DSC	HD	HD95
CANDID-PTX	0.981 ± 0.038	23.7 ± 43.0	11.9 ± 23.6	0.950 ± 0.111	23.6 ± 51.9	21.0 ± 46.6
CheXpert	0.947 ± 0.087	39.3 ± 57.7	27.0 ± 46.3	0.961 ± 0.083	18.0 ± 37.2	14.8 ± 30.3
ChestX-Ray8	0.961 ± 0.097	32.3 ± 63.4	18.8 ± 44.1	0.958 ± 0.133	18.6 ± 58.6	17.2 ± 55.1
MIMIC-CXR-JPG	0.965 ± 0.077	29.7 ± 51.1	15.1 ± 32.2	0.954 ± 0.101	22.9 ± 51.9	21.2 ± 47.8
Padchest	0.966 ± 0.089	23.0 ± 42.7	12.7 ± 27.5	0.980 ± 0.035	12.8 ± 24.3	10.4 ± 18.8
VinDr-CXR	0.978 ± 0.048	18.8 ± 40.1	8.9 ± 18.5	0.953 ± 0.119	21.9 ± 55.4	20.6 ± 52.6

Evaluación de las segmentaciones generadas usando RCA

Como parte de la validación, presentamos un análisis estadístico de la distribución de las estimaciones de DSC mediante RCA para cada conjunto de datos en la Tabla 5.2. Se observa que, para cada conjunto de datos, la media es superior a 0.88. Además, la Figura 5.6 muestra una representación gráfica del RCA estimado global para las seis bases de datos en conjunto. En el Apéndice G, Sección *Technical Validation*, se incluye una desagregación más detallada de estos resultados en términos de atributos demográficos y cómo el RCA puede utilizarse indirectamente para detectar imágenes fuera de la distribución mediante el análisis de los valores más bajos.

Tabla 5.2: Análisis estadístico del DSC estimado por RCA en las distintas bases de datos segmentadas

Dataset name	Sample size (n)	Mean	Std	Min	1 %	5 %	25 %	50 %	75 %	Max
CANDID-PTX	19235	0.853	0.031	0.573	0.756	0.795	0.837	0.858	0.875	0.920
Chestx-ray8	112120	0.841	0.044	0.148	0.691	0.764	0.823	0.850	0.869	0.925
CheXpert	187825	0.830	0.038	0.400	0.718	0.761	0.808	0.835	0.857	0.918
MIMIC-CXR-JPG	243334	0.831	0.051	0.149	0.650	0.738	0.811	0.842	0.864	0.921
Padchest	96184	0.851	0.042	0.103	0.672	0.790	0.838	0.858	0.874	0.927
VinDr-CXR	18000	0.850	0.033	0.469	0.738	0.794	0.835	0.855	0.872	0.919

Evaluación usando los datos generados para entrenar nuevos modelos

Realizamos un control de calidad final a nivel de conjunto de datos para evaluar la calidad de las segmentaciones de HybridGNet por conjunto de datos (no de forma individual, como se hizo con el DSC estimado por RCA). Nuestra hipótesis es que si las segmentaciones son de alta calidad, pueden utilizarse como conjunto de entrenamiento para un modelo de segmentación que obtendrá un alto rendimiento al ser evaluado en imágenes para las cuales sí se posee anotaciones de referencia. En este sentido, entrenamos seis nuevos modelos de HybridGNet utilizando el conjunto de máscaras predichas de cada uno de los seis conjuntos de datos públicos como conjunto de entrenamiento. Luego, evaluamos el rendimiento de los modelos en el conjunto de datos con anotaciones de referencia. También incluimos como referencia los valores obtenidos al entrenar en el 80 % de los datos de entrenamiento originales y probar en el 20 % restante.

La Tabla 5.3 muestra los resultados para la segmentación de pulmones y corazón. Se observa una ligera disminución en términos de DSC, que pasa de 0.967 a aproximadamente 0.955 cuando se utilizan los conjuntos de datos segmentados como conjunto de entrenamiento (en comparación con los valores de referencia). Esto indica que la calidad general de las segmentaciones es lo suficientemente buena como para entrenar modelos que se desempeñan bien en comparación con las anotaciones de expertos existentes.

Tabla 5.3: Estimación a nivel dataset obtenida por el reentrenamiento de la HybridGNet en los nuevos datos generados

Training Dataset	Lungs ($n = 911$)			Heart ($n = 383$)		
	DSC	HD	HD 95	DSC	HD	HD 95
CANDID-PTX	0.959 ± 0.020	36.4 ± 23.3	18.2 ± 10.5	0.919 ± 0.045	41.1 ± 22.1	36.8 ± 21.7
Chestx-ray8	0.960 ± 0.019	36.2 ± 22.7	17.7 ± 10.3	0.927 ± 0.031	39.9 ± 16.5	35.2 ± 16.3
CheXpert	0.946 ± 0.028	42.0 ± 23.5	22.9 ± 12.1	0.888 ± 0.053	58.2 ± 27.0	54.0 ± 26.4
Padchest	0.956 ± 0.028	37.7 ± 23.7	19.0 ± 12.3	0.924 ± 0.037	40.9 ± 19.6	36.6 ± 19.2
MIMIC-CXR	0.956 ± 0.020	37.5 ± 22.8	19.2 ± 10.5	0.921 ± 0.031	44.2 ± 19.9	39.6 ± 19.6
VinDr-CXR	0.959 ± 0.023	35.4 ± 23.3	17.7 ± 11.7	0.925 ± 0.035	38.4 ± 16.8	34.0 ± 16.2
20 % test split	Lungs ($n = 181$)			Heart ($n = 76$)		
HybridGNet (entrenado 80 %-20 %)	0.967 ± 0.017	32.5 ± 20.6	15.6 ± 10.9	0.937 ± 0.029	34.6 ± 14.8	29.4 ± 13.4

Sección 6

Conclusiones

En esta tesis se presentaron avances significativos en la extracción de grafos a partir de imágenes en diversas aplicaciones de procesamiento de imágenes y visión por computadora. A lo largo de los capítulos, se han introducido enfoques desacoplados y métodos unificados que abordan desafíos específicos, adaptándose al dominio en el que se aplicaron.

Se desarrolló un método para extraer grafos que representan la arquitectura de las raíces de plantas, centrándose en la planta modelo *Arabidopsis thaliana*. Este enfoque, basado en redes neuronales convolucionales, esqueletización y seguimiento de grafos, automatiza la segmentación y el seguimiento de las raíces a lo largo del tiempo. Los resultados demuestran un rendimiento prometedor al permitir la observación detallada del crecimiento de las raíces a través del tiempo y la extracción de parámetros fenotípicos previamente inaccesibles.

Se presentó la arquitectura HybridGNet, un modelo unificado en 2D que sobresale en la segmentación anatómica basada en landmarks en imágenes de rayos X. Este enfoque combina redes neuronales convolucionales con redes neuronales convolucionales en grafos, posicionándose como estado del arte para métodos basados en grafos en este contexto y logrando resultados altamente competitivos con el estado del arte en segmentación basada en píxeles, mejorando su robustez y la plausibilidad anatómica de sus resultados, al incorporar restricciones anatómicas por construcción. Además, se demostró su capacidad de adaptación a diferentes dominios, lo que lo hace adecuado para aplicaciones clínicas en las que se pueden encontrar datos de diferentes orígenes y características demográficas.

Posteriormente, se presentó un método unificado en 3D para la extracción de mallas a partir de imágenes de resonancia magnética cardíaca (CMR), abordando el desafío de generar mallas tridimensionales precisas y anatómicamente plausibles del corazón. Nuestra arquitectura MV-HybridVNet demostró un rendimiento sobresaliente en la generación de mallas de superficie y volumétricas, superando al estado del arte representado por MCSI-Net en términos de calidad de segmentación. Evaluamos diferentes términos de regularización y se propuso un nuevo término de regularización de elementos tetraédricos, que mejoró significativamente la calidad de los elementos de las mallas generadas, permitiendo obtener mallas volumétricas listas para simulación directamente a partir de imágenes.

Por último, se abordó la creación de anotaciones para grandes bases de datos a partir de datos multicéntricos y se desarrollaron métodos de validación automática de la calidad de las segmentaciones obtenidas. En particular, se demostró que los modelos basados en landmarks, como HybridGNet, son una solución sólida para entrenar con datos heterogéneos, ya que pueden manejar de manera eficiente conjuntos de datos con anotaciones variables. Este modelo se utilizó para la creación de CheXmask, una extensa base de datos multicéntrica que combina más de 650,000 imágenes de rayos X de diversos conjuntos de datos públicos. Se utilizó el método RCA para evaluar automáticamente la calidad de las segmentaciones en ausencia de anotaciones de referencia, mostrando una fuerte correlación con métricas reales de calidad.

En conclusión, esta tesis abordó el problema de extracción de grafos a partir de imágenes 2D y 3D, proponiendo tanto modelos desacoplados como arquitecturas unificadas de redes neuronales para llevar esta tarea.

Sección 7

Publicaciones

A continuación se listan todos los trabajos relacionados con la tesis en los que se participó durante el desarrollo del doctorado.

Publicaciones en revistas

- **Gaggion, N.**, Ariel, F., Daric, V., Lambert, É., Legendre, S., Roulé, T., Camoirano, A., Milone, D. H., Crespi, M., Blein, T., & Ferrante, E. (2021). ChronoRoot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture. *GigaScience*, 10(7), giab052.
- **Gaggion, N.**, Mansilla, L., Mosquera, C., Milone, D. H., & Ferrante, E. (2022). Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Medical Imaging*, 42(2), 546-556.
- Roulé, T., Legascue, M. F., Barrios, A., **Gaggion, N.**, Crespi, M., Ariel, F., & Blein, T. (2023). The long intergenic noncoding RNA ARES modulates root architecture in Arabidopsis. *IUBMB life*.

Trabajos en eventos científicos de alto impacto

- **Gaggion, N.**, Mansilla, L., Milone, D. H., & Ferrante, E. (2021). Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. *MICCAI 2021: International Conference on Medical Image Computing and Computer Assisted Intervention*, Strasbourg, France. Proceedings, Part I 24 (pp. 600-610). Springer International Publishing.
- **Gaggion, N.**, Vakalopoulou, M., Milone, D. H., & Ferrante, E. (2023, April). Multi-center anatomical segmentation with heterogeneous labels via landmark-based models. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE.
- **Gaggion, N.**, Echeveste, R., Mansilla, L., Milone, D. H., & Ferrante, E. (2023, October). Un-supervised bias discovery in medical image segmentation. In *Workshop on Clinical Image-Based Procedures* (pp. 266-275). Cham: Springer Nature Switzerland.

Publicaciones en revisión al momento del envío de la tesis

- **Gaggion, N.**, Mosquera, C., Mansilla, L., Aineseder, M., Milone, D. H., & Ferrante, E. (2023). CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. *arXiv preprint arXiv:2307.03293*.

- Barrios, A., **Gaggion, N.**, Mansilla, N., Lucero, L., Blein, T., Sorin, C., Ferrante, E., Crespi, M. & Ariel, F. D. (2023). The transcription factor NF-YA10 determines the area explored by *Arabidopsis thaliana* roots through direct regulation of LAZY and TAC genes. *bioRxiv*, 2023-04.
- **Gaggion, N.**, Matheson B., Xia, Y. Bonazzola, R., Ravikumar, N., Milone, D. H., Frangi, A., Ferrante, E (2023). A multi-view hybrid graph convolutional architecture for Volume-to-Mesh extraction in cardiovascular MR images.

Bibliografía

- [1] Armengaud, P., Zambaux, K., Hills, A., Sulpice, R., Pattison, R.J., Blatt, M.R., Amtmann, A.: EZ-Rhizo: integrated software for the fast and accurate measurement of root system architecture. *Plant J.* **57**(5), 945–956 (Mar 2009)
- [2] Attar, R., et al.: Quantitative CMR population imaging on 20,000 subjects of the UK biobank imaging study: Lv/rv quantification pipeline and its evaluation. *Medical image analysis* **56**, 26–42 (2019)
- [3] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
- [4] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional Encoder-Decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (Dec 2017)
- [5] Bai, W., Shi, W., Ledig, C., Rueckert, D.: Multi-atlas segmentation with augmented features for cardiac mr images. *Medical image analysis* **19**(1), 98–109 (2015)
- [6] Bhalodia, R., Elhabian, S., Adams, J., Tao, W., Kavan, L., Whitaker, R.: Deepssm: A blueprint for image-to-shape deep learning models. *arXiv preprint arXiv:2110.07152* (2021)
- [7] Bhalodia, R., Elhabian, S.Y., Kavan, L., Whitaker, R.T.: Deepssm: a deep learning framework for statistical shape modeling from raw images. In: *International Workshop on Shape in Medical Imaging*. pp. 244–257. Springer (2018)
- [8] Bohlender, S., Oksuz, I., Mukhopadhyay, A.: A survey on shape-constraint deep learning for medical image segmentation. *arXiv preprint arXiv:2101.07721* (2021)
- [9] Bonazzola, R., Ravikumar, N., Attar, R., Ferrante, E., Syeda-Mahmood, T., Frangi, A.F.: Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in CMR images. In: *MICCAI*. pp. 699–708. Springer (2021)
- [10] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* **34**(4), 18–42 (2017)
- [11] Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2020)
- [12] Cai, J., Zeng, Z., Connor, J.N., Huang, C.Y., Melino, V., Kumar, P., Miklavcic, S.J.: Rootgraph: a graphic optimization tool for automated image analysis of plant roots. *Journal of Experimental Botany* **66**(21), 6551–6562 (2015)
- [13] Candemir, S., Jaeger, S., Palaniappan, K., Musco, J.P., Singh, R.K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., McDonald, C.J.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging* **33**(2), 577–590 (2014). <https://doi.org/10.1109/TMI.2013.2290491>
- [14] Chen, H., Giuffrida, M.V., Doerner, P., Tsiftaris, S.A.: Blind inpainting of large-scale masks of thin structures with adversarial and reinforcement learning. *CoRR* **abs/1912.02470** (2019), <http://arxiv.org/abs/1912.02470>

- [15] Chen, H., Valerio Giuffrida, M., Doerner, P., Tsaftaris, S.A.: Adversarial large-scale root gap inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
- [16] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**(4), 834–848 (2017)
- [17] Choudhary, A., Tong, L., Zhu, Y., Wang, M.D.: Advancing medical imaging informatics by deep learning-based domain adaptation. *Yearbook of medical informatics* **29**(01), 129–138 (2020)
- [18] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: European conference on computer vision. pp. 484–498. Springer (1998)
- [19] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Training models of shape from sets of examples. In: BMVC92, pp. 9–18. Springer (1992)
- [20] Cubero, L., Serrano, J., Castelli, J., de Crevoisier, R., Acosta, O., J., P.: Exploring uncertainty for clinical acceptability in head and neck deep learning-based oar segmentation. In: IEEE ISBI 2023. IEEE (2023)
- [21] Czolbe, S., Arnavaz, K., Krause, O., Feragen, A.: Is segmentation uncertainty useful? In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27. pp. 715–726. Springer (2021)
- [22] Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems* **29** (2016)
- [23] Dorent, R., Booth, T., Li, W., Sudre, C.H., Kafiabadi, S., Cardoso, J., Ourselin, S., Vercauteren, T.: Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. *Medical image analysis* **67**, 101862 (2021)
- [24] Fedele, M., Quarteroni, A.: Polygonal surface processing and mesh generation tools for the numerical simulation of the cardiac function. *International Journal for Numerical Methods in Biomedical Engineering* **37**(4), e3435 (2021)
- [25] Feng, S., Azzollini, D., Kim, J.S., Jin, C.K., Gordon, S.P., Yeoh, J., Kim, E., Han, M., Lee, A., Patel, A., et al.: Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence* **3**(6), e210136 (2021)
- [26] Filbrandt, G., et al.: Learning from partially overlapping labels: Image segmentation under annotation shift. In: Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health, pp. 123–132. Springer (2021)
- [27] Fournel, J., Bartoli, A., Bendahan, D., Guye, M., Bernard, M., Rauso, E., Khanji, M.Y., Petersen, S.E., Jacquier, A., Ghattas, B.: Medical image segmentation automatic quality control: A multi-dimensional approach. *Medical Image Analysis* **74**, 102213 (2021)
- [28] Frangi, A., Prince, J., Sonka, M. (eds.): *Medical Image Analysis*. Elsevier, 1st edn. (September 20 2023)
- [29] Frangi, A.F., Niessen, W.J., Rueckert, D., Schnabel, J.A.: Automatic 3d asm construction via atlas-based landmarking and volumetric elastic registration. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 78–91. Springer (2001)
- [30] French, A., Ubeda-Tomás, S., Holman, T.J., Bennett, M.J., Pridmore, T.: High-throughput quantification of root growth using a novel image-analysis tool. *Plant physiology* **150**(4), 1784–1795 (2009)

- [31] Galkovskyi, T., Mileyko, Y., Bucksch, A., Moore, B., Symonova, O., Price, C.A., Topp, C.N., Iyer-Pascuzzi, A.S., Zurek, P.R., Fang, S., Harer, J., Benfey, P.N., Weitz, J.S.: GiA roots: software for the high throughput analysis of plant root system architecture. *BMC Plant Biol.* **12**, 116 (Jul 2012)
- [32] van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis* **10**(1), 19–40 (2006)
- [33] Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*. Pearson Higher Ed (Nov 2011)
- [34] Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering* **69**(3) (2021)
- [35] He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R-CNN. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Oct 2017)
- [36] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
- [37] Heitz, G., Rohlfing, T., Maurer Jr, C.R.: Automatic generation of shape models using nonrigid registration with a single segmented template mesh. In: *VMV*. pp. 73–80 (2004)
- [38] Hoopes, A., Iglesias, J.E., Fischl, B., Greve, D., Dalca, A.V.: Topofit: Rapid reconstruction of topologically-correct cortical surfaces. In: *Medical Imaging with Deep Learning (2022)*, <https://openreview.net/forum?id=-JiHeZNDY3a>
- [39] Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis* **24**(1), 205–219 (2015)
- [40] Ingram, P.A., Malamy, J.E.: *Root System Architecture, Advances in Botanical Research*, vol. 55, pp. 75–117. Elsevier (2010)
- [41] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 590–597 (2019)
- [42] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
- [43] Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R.K., Antani, S., Thoma, G., Wang, Y.X., Lu, P.X., McDonald, C.J.: Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging* **33**(2), 233–245 (2014). <https://doi.org/10.1109/TMI.2013.2284099>
- [44] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019)
- [45] Jurdia, R.E., Petitjean, C., Honeine, P., Cheplygina, V., Abdallah, F.: High-level prior-based loss functions for medical image segmentation: A survey. *arXiv preprint arXiv:2011.08018* (2020)
- [46] Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: *International MICCAI brainlesion workshop*. pp. 450–462. Springer (2017)

- [47] Kemnitz, J., et al.: Combining heterogeneously labeled datasets for training segmentation networks. In: International Workshop on Machine Learning in Medical Imaging. pp. 276–284. Springer (2018)
- [48] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
- [49] Kong, F., Shadden, S.C.: Whole heart mesh generation for image-based computational simulations by learning free-from deformations. In: MICCAI. pp. 550–559. Springer (2021)
- [50] Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011)
- [51] Larrazabal, A.J., Martinez, C., Ferrante, E.: Anatomical priors for image segmentation via post-processing with denoising autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–593. Springer (2019)
- [52] Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE TMI (2020)
- [53] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
- [54] Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. pp. 562–570 (2015)
- [55] Lobet, G.: Image analysis in plant sciences: Publish then perish. Trends Plant Sci. **22**(7), 559–566 (Jul 2017)
- [56] Mansilla, L., Ferrante, E.: Segmentación multi-atlas de imágenes médicas con selección de atlas inteligente y control de calidad automático. In: XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018). (2018)
- [57] Mansilla, L., Milone, D.H., Ferrante, E.: Learning deformable registration of medical images with anatomical constraints. Neural Networks **124**, 269–279 (2020)
- [58] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
- [59] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 4th international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
- [60] Milletari, F., Rothberg, A., Jia, J., Sofka, M.: Integrating statistical prior knowledge into convolutional neural networks. In: Medical Image Computing and Computer-Assisted Intervention. Springer (2017)
- [61] Musy, M., others.: vedo: 2022.4.1 (Oct 2022). <https://doi.org/10.5281/zenodo.7222019>
- [62] Naeem, A., French, A.P., Wells, D.M., Pridmore, T.P.: High-throughput feature counting and measurement of roots. Bioinformatics **27**(9), 1337–1338 (May 2011)
- [63] Narisetti, N., Henke, M., Seiler, C., Shi, R., Junker, A., Altmann, T., Gladilin, E.: Semi-automated root image analysis (saRIA). Sci. Rep. **9**(1), 19674 (Dec 2019)
- [64] Neic, A., Gsell, M.A., Karabelas, E., Prassl, A.J., Plank, G.: Automating image-based mesh generation and manipulation tasks in cardiac modeling workflows using meshtool. SoftwareX **11**, 100454 (2020)
- [65] Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. Scientific Data **9**(1), 429 (2022)

- [66] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., et al.: Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging* **37**(2), 384–395 (2017)
- [67] Ordas, S., Oubel, E., Leta, R., Carreras, F., Frangi, A.F.: A statistical shape model of the heart and its application to model-based segmentation. In: *Medical Imaging 2007: Physiology, Function, and Structure from Medical Images*. vol. 6511, pp. 490–500. SPIE (2007)
- [68] Orlando, J.I., Prokofyeva, E., Blaschko, M.B.: A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. *IEEE transactions on Biomedical Engineering* **64**(1), 16–27 (2016)
- [69] Pak, D.H., Liu, M., Kim, T., Liang, L., McKay, R., Sun, W., Duncan, J.S.: Distortion energy for deep learning-based volumetric finite element mesh generation for aortic valves. In: *MICCAI*. pp. 485–494. Springer International Publishing (2021)
- [70] Palladino, J.A., Slezak, D.F., Ferrante, E.: Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter mr images. In: *16th International Symposium on Medical Information Processing and Analysis*. vol. 11583, p. 1158302. SPIE (2020)
- [71] Palmer, C.M., Bush, S.M., Maloof, J.N.: Phenotypic and developmental plasticity in plants. In: *John Wiley & Sons, Ltd (ed.) eLS*, vol. 59, p. 1127. John Wiley & Sons, Ltd, Chichester, UK (May 2001)
- [72] Paulsen, R., Larsen, R., Nielsen, C., Laugesen, S., Ersbøll, B.: Building and testing a statistical shape model of the human ear canal. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 373–380. Springer (2002)
- [73] Petersen, S.E., et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. *Journal of Cardiovascular Magnetic Resonance* **15**(1), 46 (Dec 2013)
- [74] Petersen, S.E., et al.: UK biobank's cardiovascular magnetic resonance protocol. *Journal of cardiovascular magnetic resonance* **18**(1), 1–7 (2015)
- [75] Petit, O., et al.: Handling missing annotations for semantic segmentation with deep convnets. In: *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 20–28. Springer (2018)
- [76] Pound, M.P., French, A.P., Atkinson, J.A., Wells, D.M., Bennett, M.J., Pridmore, T.: Rootnav: navigating images of complex root architectures. *Plant physiology* **162**(4), 1802–1814 (2013)
- [77] Puyol-Anton, E., et al.: A multimodal spatiotemporal cardiac motion atlas from mr and ultrasound data. *Medical image analysis* **40**, 96–110 (2017)
- [78] Quantitative plant, <https://www.quantitative-plant.org>, Accessed: 2020-12-18
- [79] Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 704–720 (2018)
- [80] Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. *nature communications* **13**(1), 4581 (2022)
- [81] Rodero, C., et al.: Linking statistical shape models and simulated function in the healthy adult human heart. *PLOS Computational Biology* **17**(4), 1–28 (04 2021). <https://doi.org/10.1371/journal.pcbi.1008851>

- [82] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
- [83] Scalbert, M., Vakalopoulou, M., Couzinié-Devy, F.: Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology. In: MICCAI – MICCAI 2022. pp. 120–129. Springer Nature Switzerland (2022)
- [84] Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., Kokkinos, I.: Subcortical brain structure segmentation using f-cnn’s. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). pp. 269–272. IEEE (2016)
- [85] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. *American Journal of Roentgenology* **174**(1), 71–74 (2000)
- [86] Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine* **30**(3), 83–98 (2013)
- [87] Sozou, P.D., Cootes, T.F., Taylor, C.J., Di Mauro, E., Lanitis, A.: Non-linear point distribution modelling using a multi-layer perceptron. *Image and Vision Computing* **15**(6), 457–463 (1997)
- [88] Synopsis: Simpleware (2021), <https://www.synopsys.com/simpleware.html>, Last opened on 2022-07-11
- [89] Tetteh, G., Efremov, V., Forkert, N.D., Schneider, M., Kirschke, J., Weber, B., Zimmer, C., Piraud, M., Menze, B.H.: Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes. *Frontiers in Neuroscience* p. 1285 (2020)
- [90] Tóthová, K., Parisot, S., Lee, M., Puyol-Antón, E., King, A., Pollefeys, M., Konukoglu, E.: Probabilistic 3D surface reconstruction from sparse MRI information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 813–823. Springer (2020)
- [91] Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging* **36**(8), 1597–1606 (2017)
- [92] Ventura, C., Pont-Tuset, J., Caelles, S., Maninis, K.K., Van Gool, L.: Iterative deep learning for road topology extraction. *arXiv preprint arXiv:1808.09814* (2018)
- [93] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3D mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–67 (2018)
- [94] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
- [95] Win rhizo website, https://regentinstruments.com/assets/winrhizo_mostrecent.html, Accessed: 2020-12-18
- [96] Wolterink, J.M., van Hamersvelt, R.W., Viergever, M.A., Leiner, T., Išgum, I.: Coronary artery centerline extraction in cardiac ct angiography using a cnn-based orientation classifier. *Medical image analysis* **51**, 46–60 (2019)

- [97] Xia, Y., Chen, X., Ravikumar, N., Kelly, C., Attar, R., Aung, N., Neubauer, S., Petersen, S.E., Frangi, A.F.: Automatic 3d+t four-chamber cmr quantification of the uk biobank: integrating imaging and non-imaging data priors at scale. *Medical Image Analysis* **80**, 102498 (2022). <https://doi.org/https://doi.org/10.1016/j.media.2022.102498>, <https://www.sciencedirect.com/science/article/pii/S1361841522001451>
- [98] Yasrab, R., Atkinson, J.A., Wells, D.M., French, A.P., Pridmore, T.P., Pound, M.P.: RootNav 2.0: Deep learning for automatic navigation of complex plant root architectures. *Gigascience* **8**(11) (Nov 2019)
- [99] Zhang, T., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM* **27**(3), 236–239 (1984)
- [100] Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* **15**(5), 749–753 (2018)
- [101] Zitova, B., Flusser, J.: Image registration methods: a survey. *Image and vision computing* **21**(11), 977–1000 (2003)

Apéndice

Contribuciones

ChronoRoot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture

En este trabajo se presentó el sistema ChronoRoot, que combina hardware y software para generar fenotipado temporal de raíces de *Arabidopsis thaliana*. Mi contribución incluyó el desarrollo de la herramienta de software y el sistema de seguimiento temporal que permite el fenotipado de raíces a través del tiempo. Además, participé en el procesamiento digital de los experimentos mediante el sistema propuesto, en el análisis de los resultados, en la redacción del manuscrito y en su revisión.

The transcription factor NF-YA10 determines the area explored by *Arabidopsis thaliana* roots through direct regulation of LAZY and TAC genes

En este trabajo, analizamos el papel del factor de transcripción NF-YA10 de *Arabidopsis thaliana* en el gravitropismo de las raíces laterales. Mi contribución consistió en desarrollar nuevas métricas incorporadas al sistema ChronoRoot para fenotipar temporalmente el comportamiento de las raíces laterales, incluyendo métricas como el ángulo de emergencia lateral y el ángulo total de la raíz lateral. También participé en el procesamiento digital de los experimentos mediante el uso de ChronoRoot y en la redacción de la sección que detalla las nuevas métricas implementadas en el manuscrito.

The long intergenic noncoding RNA ARES modulates root architecture in *Arabidopsis*

En este trabajo, mi contribución se centró en el procesamiento digital de los experimentos mediante el uso de ChronoRoot.

Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

En este trabajo, se presentó el modelo unificado de extracción de grafos a partir de imágenes, específicamente en el contexto de segmentación basada en landmarks en imágenes de rayos-x. Mi contribución incluyó el desarrollo del código, la ejecución de los experimentos, el análisis de los resultados y la revisión del manuscrito.

Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis

En este trabajo, se presentó una extensión del modelo unificado en el contexto de imágenes de rayos-x, incluyendo supervisión profunda y conexiones residuales de imagen a grafo, junto con experimentos en diversos escenarios. Mi contribución abarcó el desarrollo del código, la ejecución de los experimentos, el análisis de los resultados, la escritura y revisión del manuscrito.

A multi-view hybrid graph convolutional architecture for Volume-to-Mesh extraction in cardiovascular MR images

En este trabajo, se presentó el modelo unificado para extracción de mallas a partir de imágenes volumétricas de resonancia magnética cardíaca. Mi contribución incluyó el desarrollo del código, la ejecución de los experimentos, el análisis de los resultados, la escritura y revisión del manuscrito.

Multi-center anatomical segmentation with heterogeneous labels via landmark-based models

En este trabajo, se exploró cómo las anotaciones provenientes de distintos centros clínicos pueden influir en el entrenamiento de modelos de aprendizaje profundo. Se desarrolló una nueva metodología para entrenar en contexto de datos heterogéneos. Mi contribución incluyó el desarrollo del código, la ejecución de los experimentos, el análisis de los resultados, la escritura y revisión del manuscrito.

CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images

En este trabajo, se utilizó el modelo unificado aplicado a imágenes de rayos-x para segmentar masivamente más de 650 mil imágenes de rayos-x y se realizó una validación automática de la calidad de los resultados. Mi contribución incluyó el proceso de segmentación, la implementación del modelo de registración afin, el análisis de los resultados, la escritura y revisión del manuscrito.

Unsupervised bias discovery in medical image segmentation

En este trabajo, se evaluó la posibilidad de utilizar métodos de evaluación automática de la calidad basados en RCA para la detección de sesgos en segmentación de imágenes médicas. Mi contribución incluyó el desarrollo del código, la ejecución de los experimentos y la revisión del manuscrito.

ChronoRoot: High-throughput
phenotyping by deep segmentation
networks reveals novel temporal
parameters of plant root system
architecture

ChronoRoot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture

Nicolás Gaggion¹, Federico Ariel², Vladimir Daric³, Éric Lambert³, Simon Legendre³, Thomas Roulé³, Alejandra Camoirano², Diego H. Milone¹, Martin Crespi³, Thomas Blein³, and Enzo Ferrante¹

¹Research Institute for Signals, Systems and Computational Intelligence (sinc(i)), CONICET, FICH, Universidad Nacional del Litoral, Ciudad Universitaria UNL, Santa Fe, Argentina.

²Instituto de Agrobiotecnología del Litoral (IAL), CONICET, FBCB, Universidad Nacional del Litoral, Colectora Ruta Nacional 168 km 0, Santa Fe, Argentina.

³Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, University Paris-Saclay and University of Paris Bâtiment 630, 91192 Gif sur Yvette, France.

Abstract

Background:

Deep learning methods have outperformed previous techniques in most computer vision tasks, including image-based plant phenotyping. However, massive data collection of root traits and the development of associated artificial intelligence approaches have been hampered by the inaccessibility of the rhizosphere. Here we present ChronoRoot, a system which combines 3D printed open-hardware with deep segmentation networks for high temporal resolution phenotyping of plant roots in agarized medium.

Results:

We developed a novel deep learning based root extraction method which leverages the latest advances in convolutional neural networks for image segmentation, and incorporates temporal consistency into the root system architecture reconstruction process. Automatic extraction of phenotypic parameters from sequences of images allowed a comprehensive characterization of the root system growth dynamics. Furthermore, novel time-associated parameters emerged from the analysis of spectral features derived from temporal signals.

Conclusions:

Altogether, our work shows that the combination of machine intelligence methods and a 3D-printed device expands the possibilities of root high-throughput phenotyping for genetics and natural variation studies as well as the screening of clock-related mutants, revealing novel root traits.

1 Background

Plants are sessile organisms unable to seek out optimal environmental conditions for development and survival. Strikingly, a remarkable developmental plasticity allows plants to complete their life cycle under changing growth conditions [46]. Understanding plant root plastic growth is crucial to assess how different populations may respond to the same soil properties or environmental conditions and to link this developmental adaptation to their genetic background [52]. Under controlled conditions,

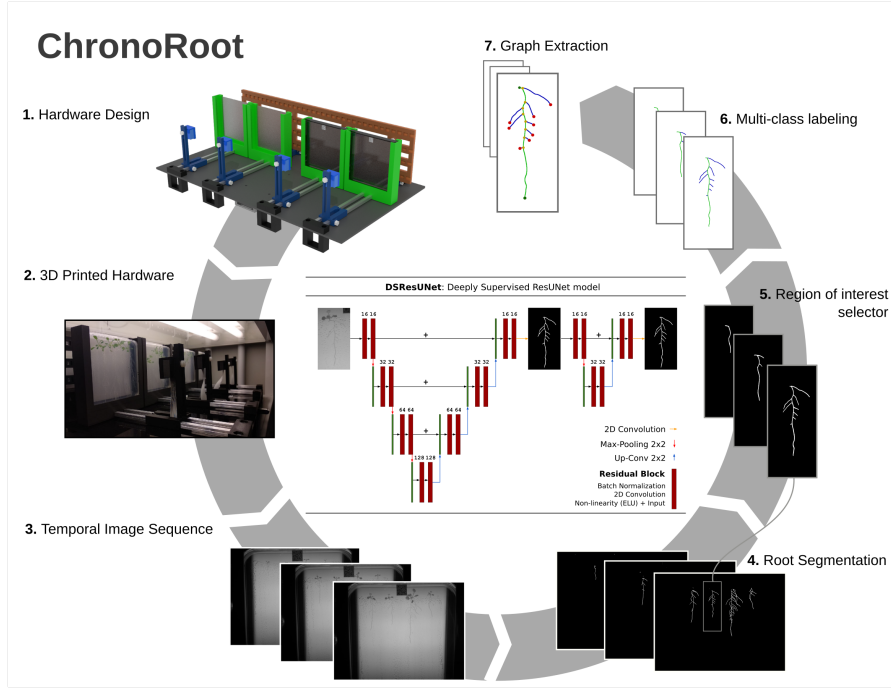


Figure 1: Main components of ChronoRoot. 1) Open hardware specification (see the Supplementary File 1 for a detailed description). 2) 3D printed ChronoRoot system mounted in a plant growth chamber. 3) Temporal sequence of images acquired by the system are provided as input to the CNN based segmentation module. The diagram corresponding to the proposed CNN architecture is included in the center of the Figure. 4) The deep learning model produces dense segmentation maps for all the plants, which are enhanced taking into account the temporal consistency of the results. 5) Independent plants can be selected to be processed individually. 6,7) The roots are skeletonized and a graph is constructed by traversing the skeleton. Pixels in the skeleton are identified as belonging to the main root (green) or lateral root (blue). The graph nodes are labeled as being the main root seed and tip (green), lateral root tip (red) or bifurcation (yellow).

root development is generally observed based on images of plants growing vertically on the surface of a semisolid agarized medium. Root system architecture (RSA) is then characterized by parametrization of a grown plant, which relies on the combination of a subset of variables like main root (MR) length or density and length of the lateral roots (LRs) [31]. Several semi-automatic tools have been developed to assist root phenotyping at specific time points [44]. However, temporal phenotyping is generally hindered by technological limitations, ignoring potentially useful phenotypical parameters that may be linked to the temporal dynamics of root growth. Here we present ChronoRoot, a low-cost system based on off-the-shelf electronics, 3D printed hardware components and deep learning models, allowing high-throughput temporal phenotyping of *Arabidopsis thaliana* RSA. Figure 1 illustrates the different components of ChronoRoot. Temporal sequences of pictures, automatically snapped, are processed for root segmentation through a convolutional neural network (CNN) model. We leverage the latest advances in CNNs for image segmentation and propose an architecture for RSA delineation which incorporates deep supervision, producing fast and accurate segmentations. The root extraction workflow is completed by a temporal consistency refinement step and a final graph generation process, which generates a labeled root graph for every image. An exploratory approach assessing root growth under alternative photoperiods served to demonstrate that temporal phenotyping performed by ChronoRoot allows deciphering the evolution of the traditional RSA parameters throughout time. Moreover, novel

parameters emerged, including architectural features, oscillating growth speed and other characteristics derived from spectral analysis of the growth signals in the Fourier domain. The combination between a low-cost automatic device for image acquisition and machine intelligence methods for image segmentation gave rise to a powerful tool for root phenomics potentially applicable to natural variation studies, the characterization of root-related subtle disorders and the screening for clock-associated mutants.

2 Data Description

2.1 Plant material and growth conditions

Arabidopsis thaliana ecotype Col-0 seeds were surface sterilized and stratified at 4°C for 2d before being grown under long day conditions (16h light, $140\mu Em^{-2}s^{-1}$ / 8h dark), or continuous light (24h light, $140\mu Em^{-2}s^{-1}$) at 22°C, on half-strength Murashige and Skoog media (1/2 MS) (Duchefa, Netherlands) with 0.8% plant agar (Duchefa, Netherlands). Four seeds were used per plate. All the experiments were performed under laboratory conditions according to the local institutional guidelines.

2.2 Datasets

We generated two different datasets in this work: the first one was used to train and evaluate the segmentation performance of the CNN models, while the second one served as an exploratory use case, to assess root growth under alternative photoperiods and provide an example of the novel temporal phenotypical parameters that can be extracted with ChronoRoot. Note that all these images were obtained with the proposed 3D printed hardware, and both are available to encourage reproducible research. Importantly, when splitting the training, validation and test partitions, we were careful not to include images corresponding to the same video on different partitions, to avoid overoptimistic biased evaluations.

- **Dataset used to train and validate the deep learning models for root segmentation:**
The dataset used for training consisted of 331 images from 55 videos (on average 6 images from the same plate at different states of growth), 11 of those videos were annotated by an expert biologist. The dataset used for testing consisted of 55 images from 11 different videos, all

Model	Dice		Recall		Hausdorff Distance (mm)		Time	# Params
	Thresh	CRF	Thresh	CRF	Thresh	CRF		
UNet	0.769±0.048	0.774±0.044	0.871±0.044	0.830±0.056	10.25±7.45	9.39±7.94	0.29s	488.212
ResUNet	0.768±0.050	0.770±0.047	0.862±0.046	0.823±0.057	8.83±6.71	7.53±5.91	0.33s	505.046
Deeply Supervised	0.769±0.048	0.772±0.045	0.861±0.044	0.815±0.057	8.14±7.34	6.95±5.42	0.49s	532.336
ResUNet (our)								
SegNet	0.768±0.043	0.773±0.040	0.862±0.044	0.824±0.053	7.42±6.40	6.81±5.65	1.49s	29.460.450
DeepLab	0.666±0.055	0.609±0.079	0.763±0.077	0.600±0.113	7.58±7.79	7.56±7.52	1.86s	58.009.410
Ensamble (our)	0.772±0.048	0.774±0.044	0.864±0.044	0.804±0.061	6.68±5.08	6.45±4.98	4.5s	

Table 1: **Quantitative evaluation for the different CNN architectures compared in this work.** We measured the Dice coefficient, recall and Hausdorff distance for dense root segmentation task. We compared state-of-the-art models (including UNet, ResUNet, SegNet and DeepLab) and compared with the proposed Deeply Supervised ResUNet and the ensemble of multiple models and architectures. On the one side, we found that our Ensemble of Multiple Models and Architectures produced equal or more accurate results than the rest of the models in terms of Dice and Hausdorff, at the expense of increasing the processing time. On the other side, the proposed Deeply Supervised ResUNet is fast (less than half a second), shows a significantly lower value for Hausdorff distance than the other fast models, while keeping equivalently good Dice and Recall.

annotated by the same expert. The tool used for the manual annotation was ITK-SNAP[58]. In total, 240 plants distributed over the 66 videos were used for training/testing the methods.

- **Use case dataset for plant phenotyping under alternative photoperiods:** We used 12 videos for each photoperiod, with pictures taken every 15 minutes. We took the first 17 days (1632 frames), and after processing the videos we proceeded to discard the results from the first 3 days prior to seed germination. We selected 25 plants from each photoperiod to perform the temporal growth analyses.

3 Analyses

We designed an automatic method to perform RSA delineation in temporal image sequences of plant roots. Our framework takes a sequence of images as input and outputs a labeled graph for each frame, representing the current root growing state. Graphs are powerful data structures particularly useful to represent curvilinear shapes like plant roots (details on the graph generation process are provided in the Methods section). The main module of the RSA delineation method is a deep CNN which produces a dense segmentation mask, where every pixel is classified as belonging to the root or the background. We proposed different CNN architectures for this task (described in the Methods section), and compared their performance with state-of-the-art models using manual annotations produced by expert biologists. We measured three different metrics: (1) Dice coefficient quantifies the overlapping between the prediction and the ground-truth, (2) Hausdorff distance indicates the maximum distance between them and (3) the recall (or sensitivity) refers to the fraction of root pixels retrieved over the total amount of root pixels. Quantitative results are included in Table 1. Based on these results, we chose two models, depending on whether we aim at having a faster or more accurate method:

- **Fast method:** The fastest models are the proposed UNet [48] variants, requiring up to half a second to process a high resolution image using a standard GPU. These models have lower parameter complexity compared to state-of-the-art architectures like SegNet [10] and DeepLab [15], which explains the lower running time. Among the fast UNet models, we observed that the proposed Deeply Supervised ResUNet (DSResUNet) shows a significantly lower value for Hausdorff distance, while keeping equivalently good Dice and Recall. The proposed DSResUNet architecture (depicted in Figure 1) combines residual blocks [28] with deep supervision [38], improving the results of a standard UNet with a minimum increase in model complexity.
- **Accurate method:** We proposed to combine all the implemented architectures into a single ensemble method, increasing model diversity by creating an ensemble of multiple models and architectures [33]. This ensemble of deep models increased the running time by a factor of 9, but achieved the best performance across all metrics, outperforming state-of-the-art models like SegNet and DeepLab.

ChronoRoot implements both fast and accurate variants, giving the users the opportunity to decide according to their requirements. In this study, we used the fast method based on the proposed Deeply Supervised ResUNet model, which offered a good trade off between running time and accuracy. We apply several post-processing steps after segmentation, which are independent of the CNN model. We first apply a Conditional Random Field (CRF) [45, 34] model to improve the homogeneity of the labels assigned to neighboring pixels. Then, we enhance the temporal consistency of the segmentations by considering its weighted average. These steps serve to remove spurious segmentations by analyzing a temporal sequence of images, which ultimately translates into generating more stable phenotypic measurements. A graph structure is then constructed where every node is assigned a class label indicating whether it is associated with the plant seed, main root, lateral root, bifurcation or root tip. Temporal consistency on the graph structures is finally improved by tracking the labeled nodes and solving conflicting cases. A more detailed description of these steps can be found in the Methods section. After the graph generation process, we proceed to extract phenotypic features for RSA characterization.

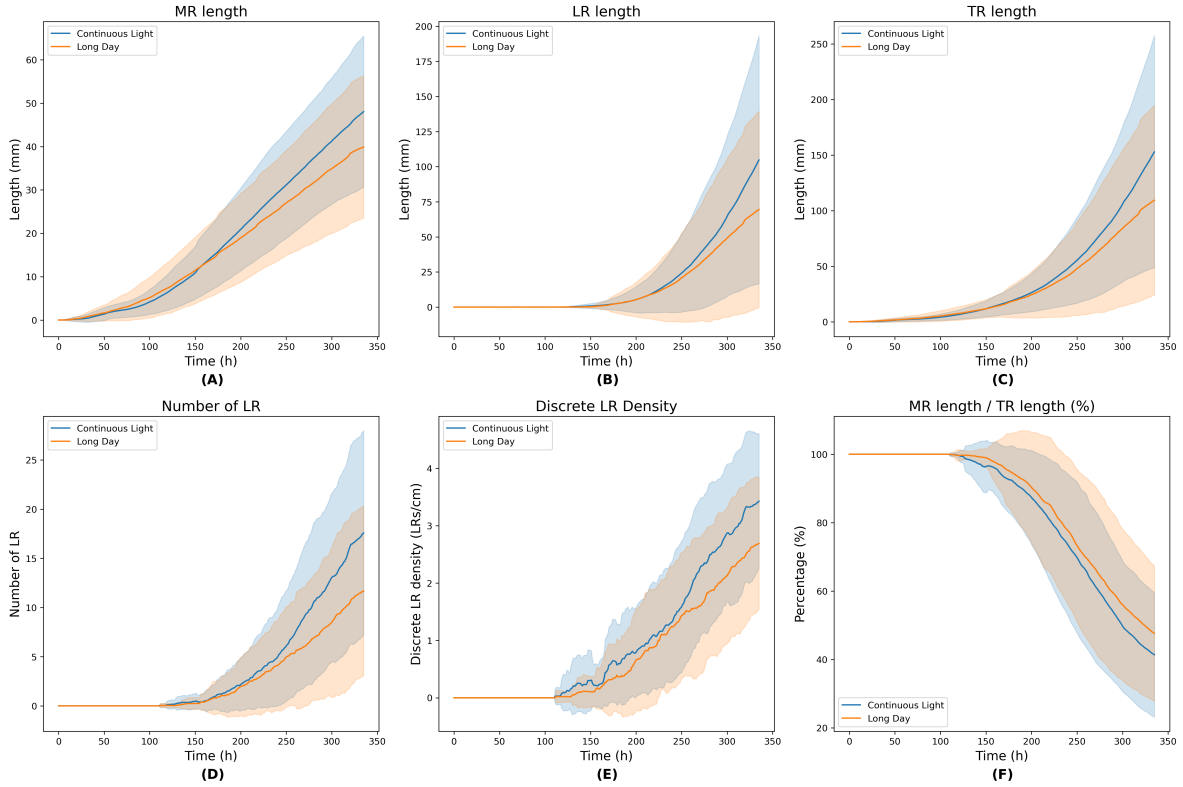


Figure 2: Traditional RSA parameters expanded to the temporal dimension. A. Main root (MR) length; B. The sum of all LRs length; C. Total root (TR) length, representing the sum of LR total length and MR length; D. Number of lateral roots (LRs); E. LR density, expressed as LR number / MR length; and F. MR component of the RSA, expressed as MR length / total root (TR) length, which is the sum of MR and LRs. Data is shown for plants grown under continuous light and long day. The lines indicate the mean, and the shadows represent the standard deviation (SD) throughout the experiment.

3.1 Temporal dimension of traditional and novel RSA parameters

We analyzed temporal sequences of plant roots growing under different conditions. In order to assess the potential of ChronoRoot, we decided to compare RSA of *Arabidopsis thaliana* ecotype Col-0 grown under two distinctive photoperiods, i.e. long day (LD; 16 h of light, 8 h of dark) or continuous light (CL; 24 h of light). Light availability and photosynthesis in the shoot determine the amount of sugar transported to the roots, thus modulating underground plant growth. Moreover, ample evidence suggests that root developmental plasticity depends on the light environment, involving a more sophisticated impact on endogenous signaling pathways [26].

Traditional parameterization of RSA expanded to temporal dynamics revealed the progression of root growth under continuous light (CL) and long day (LD) conditions. A representation of root automatic segmentation is shown in Figure 1 (see Methods and Figure 6 for more details). Our experiments show that main root (MR) length, the sum of lateral root (LR) length and the resulting total root system (TR) begin to differ between conditions at approximately 200-250 h (8-10 days) after germination (Figure 2A, B and C), together with LR number (Figure 2D). Notably, root growth was not only faster under CL, but also resulted in a different RSA, exhibiting a higher density of LRs and a lower component of the MR over the total root system (Figure 2E and F). Notably, between 250 h (10 days) and the end of the experiment (336 h, 14 days), the contrast between both photoperiods increased

gradually in every measured parameter, hinting at a temporal reorganization of root development under different light conditions.

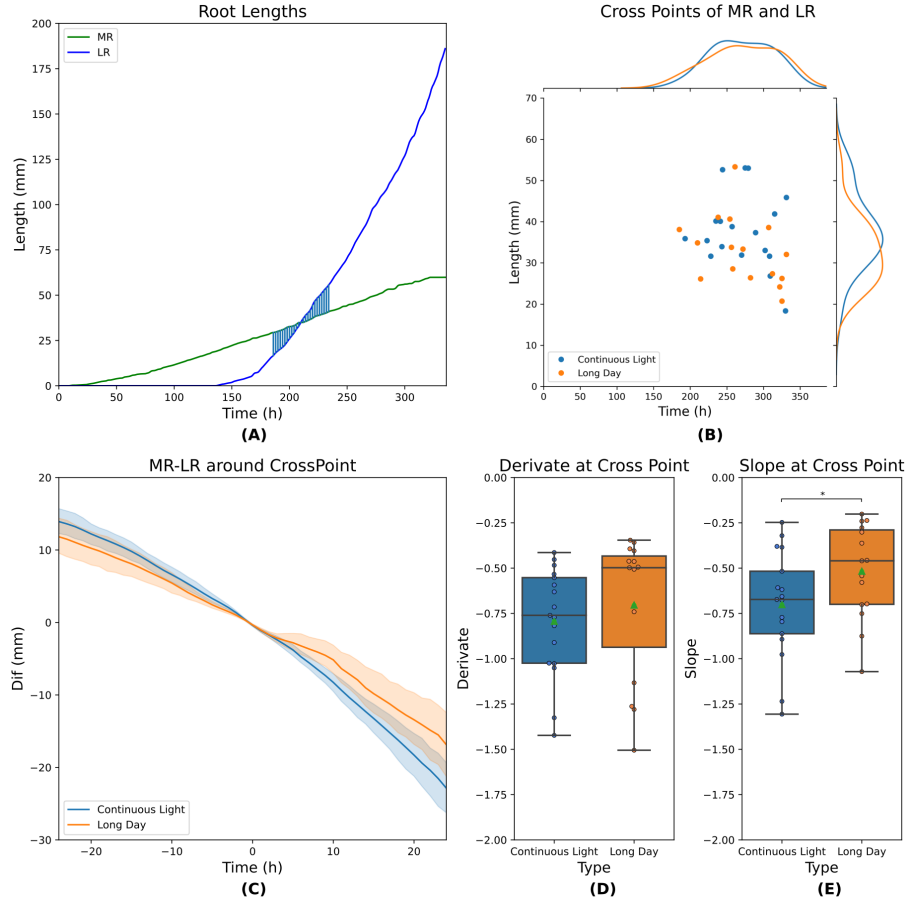


Figure 3: Novel RSA parameters analyzed along time. A. Example of one individual main root (MR) and the sum of lateral root (LR) length along time, revealing the time point of the intersection between the two curves. B. Distribution of the intersection points for all individuals from both conditions (continuous light, CL, and long day, LD). At the top and on the right, the distribution of both populations are represented. For this experiment, no significant difference was observed with respect to the intersection time point. C. The difference between the MR length and total LRs length at each time point was calculated and aligned around the point of MR-LRs = 0 mm for each individual considering ± 24 h. D. Distribution of the derivative value for the individual curves from panel C at time of equal length of MR and total LRs. The difference between the means is not statistically significant according to a non-parametric Mann Whitney U test (p -value > 0.05). E. For every individual, the tendency of the MR-LR curves shown in panel C was determined by fitting a linear function to every curve, considering ± 24 h. The boxplot shows the distribution of the slope of the fitted curves, revealing a clear difference between LC and CL. Difference between the means is statistically significant according to a non-parametric Mann Whitney U test (p -value < 0.05).

Based on the information derived from temporal phenotyping, we explored in more detail the reconfiguration of RSA under alternative photoperiods. We identified the time point at which the sum of LRs length equals the MR length as a novel parameter of RSA dynamics (Figure 3A). However, no significant difference was observed in the distribution of individual time-length points of plants grown under CL or under LD (Figure 3B). The analysis of the relationship between the MR and LRs along

time, determined by the difference between both measurements (MR-LRs) aligned to the time point at which MR and LRs are of the same length (time 0) is shown in Figure 3C. It reflects that the difference between MR and LRs tends to have a significantly larger absolute value for plants grown under CL than under LD. Moreover, we extracted different indicators to analyze the dynamics of these curves. Figure 3D shows the approximate derivative (computed by means of finite differences) at the time point at which MR and LRs are of the same lengths (time 0). These differences are not statistically significant according to a Mann Withney U test. However, when extending the analysis to the full ± 24 h range by fitting a linear function to every curve from Figure 3C and plotting the corresponding slopes (Figure 3E), we found strong differences in the distribution (statistically significant according to a Mann Withney U test) for both photoperiods. This novel time-related parameter reflects the dynamics of root growth by determining how long it takes for the system to be composed of more LRs than the MR.

In order to assess the impact of RSA reconfiguration on the area explored by roots under distinct photoperiods, we calculated the dynamic convex hull for each subset of plants. Interestingly, the observation of the convex hull resulting from the overlap of all individuals grown in the corresponding conditions reveals an extended high density of LRs along the MR axe at the end of the experiment under CL (14 days after germination, Figure 4A, B and C). Notably, the area of the average convex hulls between 8 and 14 days does not differ between CL and LD conditions (Figure 4D). Nonetheless, the quantification of the sum of LRs length over the convex hull area indicates that the density of LRs is higher under CL between 10 and 14 days (Figure 4E). Collectively, our analyses indicate that global LR length increases under CL as a result of more numerous LRs growing simultaneously, although the area explored by the RSA does not differ between the two photoperiods. Thus, the global density of the resulting RSA is higher under CL.

3.2 Novel speed-based parameters derived from temporal phenotyping

The information derived from the temporal dimension of traditional and novel RSA parameters indicated that the difference in root growth rate became broader throughout time under CL vs. LD. It has been shown that the *Arabidopsis thaliana* MR exhibits an oscillating growth which likely depends on the lunisolar tide [26, 21] and light-associated carbon partitioning [57]. Therefore, based on the segmentations obtained with our deep learning models, we calculated the growth speed throughout the experiment in both conditions, showcasing how novel speed-based parameters can be derived via ChronoRoot. MR speed grew steadily until approximately 150 h under LD and 200 h under CL post germination, and the average maximum speed reached in CL was higher than in LD (Figure 5A). Strikingly, the difference in the growth speed of the global root system (TR) between the two conditions became increasingly larger since the moment when the speed of the MR was stabilized (Figure 5B), hinting at a different acceleration rate between conditions. The observed root growth dynamics further supports the rising relevance of LRs as a main component of RSA throughout time.

Notably, the analyses of growth speed uncovered an oscillating behavior in both conditions (Figure 5A and B). To better understand the different growth patterns exhibited by LD and CL conditions, we performed a Fourier decomposition of the growth speed signals. Fourier transform decomposes functions depending on time into functions depending on frequency. In other words, the Fourier transform of a given function describes how much of any given frequency is present in the original signal. When comparing growth speed signals, analyzing their Fourier spectrum helps us to see how much this signal correlates with particular oscillation frequencies. For example, if high Fourier coefficients are associated with the frequency $1/24$ h, it means that the plant growth speed follows a daily oscillation (corresponding to what is known as circadian rhythm). Differences in the Fourier coefficients at a given frequency between growth conditions would indicate an alteration in the oscillatory pattern of plant growth. A Fast Fourier Transformation (FFT) of the signal of MR growth speed in CL vs. LD revealed a major energy difference in the components corresponding to the frequencies of $1/24$ h and $1/12$ h, respectively (Figure 5C). Remarkably, these two components distinguish circadian and ultradian rhythms displayed by plants grown in LD, with a pronounced local minimum of the growth speed

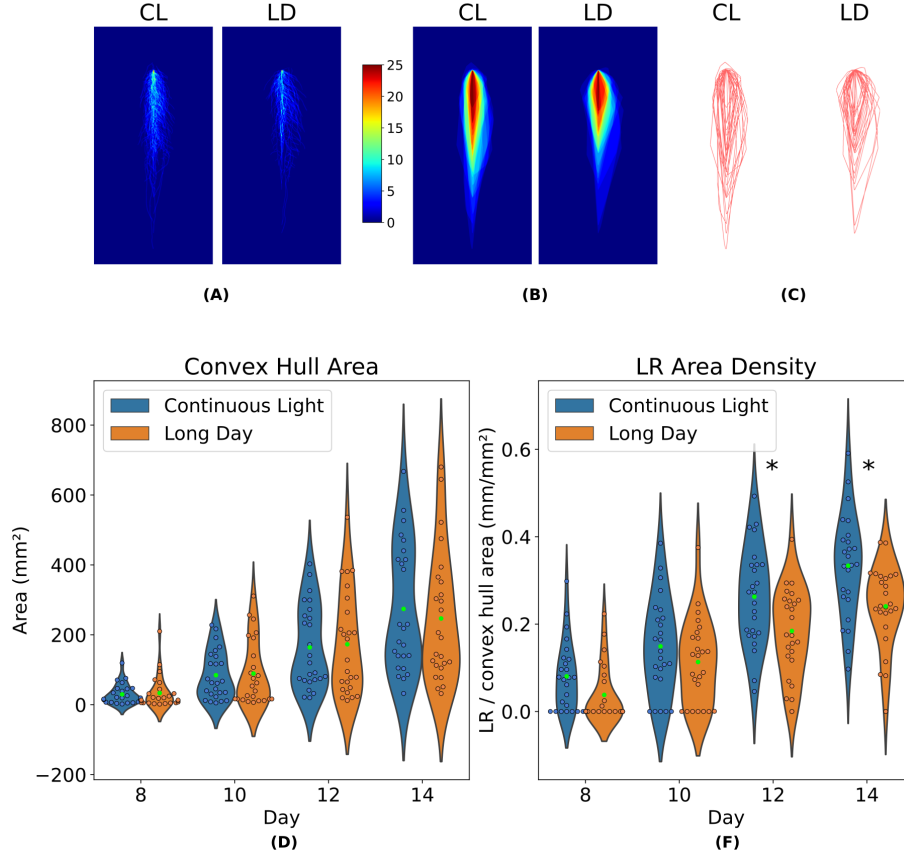


Figure 4: Area and density of the RSA analyzed along time. A. Overlapped segmentations of the whole root system at 14 days after germination. Blue background indicates no roots. The brightness of the signal increases as more roots occupy the same position. B. Same as A, represented as a heat map of the convex hull extracted for each individual. C. Overlapped convex hull contours for each condition. D. Average convex hull area for different time points under CL or LD, represented as violin plots. The mean is indicated as a green point. E. LR density calculated as the sum of LR length over the area included in the respective convex hull. The distribution of each population for the corresponding time points is shown as violin plots. The green points indicate the mean. Asterisks indicate that the difference is statistically significant. We used Shapiro-Wilk test to assess gaussianity, Levene test to confirm equal variances and t-test to confirm that differences between the means of both populations are statistically significant (p-value < 0.01).

at 1/24h and a minor local minimum at 1/12h (Figure 5D). Strikingly, the most pronounced differences revealed by FFT (Figure 5C) served to uncover a root growth clock-related disorder suffered under CL, coinciding with a blurred daily oscillation of growth speed, in comparison with the corresponding sine curves (Figure 5D; a detailed comparison of MR, LR and TR growth speed analyses is shown Supplementary Figure 1). Although an oscillating behavior can be observed under CL towards the end of the experiment (Figure 5A and B), the energy at 1/24 and 1/12 frequencies was higher under a LD photoperiod throughout the complete time lapse analyzed (Figure 5E and F). Notably, the difference between conditions of the TR oscillating speed of growth is mainly due to the MR contribution (Supplementary Figure 2). Altogether, our study of wild-type *Arabidopsis thaliana* Col-0 plants growing under alternative photoperiods using ChronoRoot served to reveal novel temporal parameters of root

development, notably including clock-related features depending on the light environment.

3.3 3D-printed device for temporal image acquisition

The ChronoRoot device is an affordable and modular imaging system based on 3D-printed and laser cut pieces and off-the-shelf electronics (Figure 1.1 and 1.2). Each module consists of a Raspberry Pi (v3)-embedded computer controlling four fixed-zoom and fixed-focus cameras (RaspiCam v2), and an array of infrared (IR) LED back-light. In between each camera and the corresponding IR array, there is a vertical 12 x 12 cm plate for seedling growth, allowing automatic image acquisition repeatedly along the experiment without any modification or movement of the imaging setup. The four-plate module is small (62 x 36 x 20 cm) and can be placed in any growth chamber. The different parts of the imaging setup (back-light, plate support and camera) can be positioned along a horizontal double-rail to control the field of view of the camera and accurate lightning. In addition, the camera can be moved vertically. ChronoRoot allows image acquisition at a high temporal resolution (a set of pictures every minute). The use of an IR back-light (850 nm) and optional long pass IR filters (> 830 nm) allow acquiring images of the same quality independently from the light conditions required for the experiment, during day and night.

Each module is connected to the network either by Wi-Fi or Ethernet cable. A web interface allows the control of the device offering live feed of the cameras for field of view and focus setup. The user can program the activation of cameras and IR back-light, starting and ending dates, the time basis for picture acquisition, and finally follow the progression of the experiment. The pictures are saved directly on an external drive plugged on the Raspberry Pi. Once the experimental setup is ready, each module is completely independent from the external environment and the access to the network (for more details see Methods, Supplementary Figures 3-7 and the Supplementary 3D printing and laser cutting files).

4 Discussion

The plant phenotype can be defined as the integration of structural, physiological, and performance-related traits of a genotype in a given environment. Plant phenotyping is therefore the act of determining the quantitative or qualitative values of these traits [19]. The advent of novel imaging technologies and image processing have revolutionized plant phenotyping, expanding the frontiers of phenotypic trait measurement. Plant roots have a major role in plant anchorage and resource acquisition while offering environmental benefits such as carbon sequestration and soil erosion mitigation [41]. The growing knowledge linking genetics with functional properties of plant roots is of crucial interest to plant breeding, notably for the design of novel strategies for sustainable agriculture and environmental stewardship in the face of the impending climate change. Whereas high throughput genotyping, sequencing-based genotyping and genomic breeding are behind current agricultural practices in the era of omics technologies, the collection of phenotypic data for a thorough characterization of the RSA is increasingly becoming a limiting factor [35]. Although significant advancements in the application of imaging sensors for high-throughput data collection have allowed comprehensive plant phenotyping [17], the characterization of root traits has been hampered by the inaccessibility of the rhizosphere.

Large and sophisticated phenotyping platforms are deployed worldwide and allow the simultaneous phenotyping of several hundreds of plants (see the International Plant Phenotyping Network[2]). However, their big dimensions and high automatization reserve their implementation on specialized locations and mainly for large phenotyping experiments. Root imaging systems of intermediate complexity, like the one proposed in [54], serve to address the temporal phenotyping of plant roots using vertical plates for plant growth in agarized medium. However, such system still requires a single-axis mobile robot, which implies more expensive electronics and plant chamber space devoted to the equipment setup. In contrast, low-cost ChronoRoot modules can be located easily in already existing facilities without major modifications or permanent movement. The number of modules to be built

and used will only depend on the available space and the experimental design (e.g. a few modules for the characterization of given genotypes vs. multiple units for GWAS approaches using tens to hundreds of plant accessions). It also allows to scale the hardware setup according to available funding and experiment requirements progressively. In parallel, the advancement of the Do It Yourself (DIY) movement has promoted the development of a growing number of low-cost phenotyping devices combining 3D-printed, laser cut, captor and microcontroller coming from open-source and open-hardware communities such as Arduino[1] or Raspberry Pi[5]. Successful inexpensive devices have allowed monitoring plant leaves [53], including Phenotiki, an affordable open software and hardware platform for image-based phenotyping of plant aerial organs [42]. More recently, the Phenotiki sensor interface was used to characterize cotton RSA on soil-containing big Rhizoboxes, allowing the determination of basic architectural parameters [11] (like total root area, convex hull area, total root length, etc). In contrast, ChronoRoot allows for a more fine-grained high-throughput temporal phenotyping, e.g. making it possible to distinguish between MR and LR. Finally, modular rhizotrons of more sophisticated designs (including more expensive cameras and light flashes, precluding observations during the night) also served for RSA characterization of crops [55].

In the last few years, images of the root system from different plant species have been acquired manually using a flat-bed scanner or a camera positioned in front of vertical petri dishes. Root phenotyping is generally performed upon single-time-point images or using several images acquired during growth in time-lapse sequences. More recently, semi-automatic devices and softwares have also helped to increase the efficiency of image acquisition and associated analyses [32, 29, 9, 16, 40, 24]. The great need of throughput in screening experiments to uncover the genetic basis of root development, justifies the use of simplified artificial culture conditions and standardized environments to make the RSA accessible to image acquisition [19]. Software tools use input images of root systems grown under a variety of conditions, including hydroponic and aeroponic systems, agarized medium, paper pouches or soil [40, 16]. Here we propose to use vertical square petri dishes for plant growth on the surface of transparent agarized medium, for automatic acquisition of photographs allowing a high resolution temporal phenotyping of the RSA.

According to Quantitative Plant [39, 4], over 40 image processing softwares are available for root system analysis [47, 40]. RSA parameters are extracted from various types of 2D images captured from agar plates or washed roots extracted from soil. Moreover, 3D RSA reconstruction is possible using X-ray computed tomography [51] or magnetic resonance imaging [20]. Nearly all the reported tools need human input to be operated and retrieve precise numbers. RootTrace [22, 43] for example, which also focuses on high-throughput analyses of root growth, employs traditional image processing and tracking techniques, resulting in a program that can only extract MR length and count the number of emerged LRs. On the contrary, our model relies on deep networks producing a detailed segmentation of the RSA which is then classified into MR and LR, allowing for fine grained measurements like the total length of the LRs, which is not provided by RootTrace. Other tools such as GiaRoots [24] and EZ Rhizo [9] employ simple threshold strategies for root segmentation. In contrast to ChronoRoot, these tools fall short at handling segmentation problems emerging from drops due to water condensation, they require manual human calibration and do not take advantage of the redundancy provided by the temporal resolution of the high-throughput videos to filter out spurious segmentations. Another alternative tool is Win Rhizo [8], a commercial and non-open source tool designed to work with images captured with high resolution desktop optical scanners. Such a requirement makes it virtually impossible to capture high-throughput temporal sequences of growing plants. On the contrary, ChronoRoot is open-source and designed to work with low-cost cameras. Another option is BRAT [50], designed for high-throughput phenotyping of root growth and development. The main disadvantage of BRAT is that it can only handle early root growth, and does not provide measurements for LRs.

The previously discussed methods are mostly based on conventional image processing approaches and extract a limited number of RSA features. Advances in machine learning applied to image analysis allowed partially overcoming these limitations. For example, deep learning techniques have been used to improve the consistency of classic approaches enhancing the quality of root segmentations [14, 13]. Closest to our work is the recent RootNav 2 [56], which is also based on deep learning models

and provides fine-grained metrics distinguishing between MR and LRs. However, RootNav 2 does not exploit the redundancy provided by the temporal resolution and follows a different architectural design, which makes ground truth annotations more difficult to obtain and prevents us from training the model with our dataset. Compared to ChronoRoot, RootNav 2 employs a more complex neural network architecture with two output paths: the first one is used to predict root segmentation masks (differentiating between MR and LRs) while the second one produces heat maps associated with root tips. This design choice requires the ground truth annotations to be composed of 3 parts: (1) MR pixel level annotations, (2) LR pixel level annotations and (3) root tip annotations. Conversely, ChronoRoot just requires binary segmentation maps (background vs foreground root) for training, since the MR and LR labeling is performed after segmentation following a depth first search approach on the skeletonized binary segmentation. Thus, our dataset is just composed of images with foreground/background pixel level annotations, which is not enough for training the RootNav 2 model.

5 Potential implications

ChronoRoot expands the possibilities for high throughput root phenotyping, which is of major importance for natural variation and GWAS, as well as mutant characterization and screening. Notably, it has been shown that clock-related mutants exhibit a differential oscillating MR growth under alternative conditions [21, 57, 30]. ChronoRoot offers an ideal platform for the identification of genotypes associated with altered clock traits, based on the analysis of spectral features extracted from temporal signals.

Note that the specification of all ChronoRoot hardware components (e.g. camera, etc) is released in this paper. Thus, for anybody installing the system and using the same imaging setup, our software for *Arabidopsis thaliana* analysis should work without further retraining. In case the system fails due to different lighting conditions or hardware components, a minimal fine-tuning of the model by using a few annotated images from the new setup may be required.

6 Methods

6.1 Hardware description

An automated imaging setup was designed and built in the shape of an independent module of 62 x 36 x 20 cm (Figure 1). It is aimed at imaging up to four vertical plates either in color or in near-infrared (NIR) lightning. Each module consists of a single board computer (Raspberry Pi) controlling four cameras through a multiplexer module and an array of NIR LED illumination through a relay. The main support of each module is a 620 x 36 x 5 mm acrylic sheet cut using a laser cutter to allow to screw the different parts or let pass strips connecting the camera to the camera multiplexer. Several 3D pieces were designed and printed to place the different components of the module. Each module is separated in four subparts, each of them along a double aluminum axis. This axis allows to adjust the distance of the different parts of the imaging setup: NIR illumination, plate support, camera. The underpart of the module was used to fix the LED AC/DC adaptor, the relay and the computer. The supports under the platform raise and stabilize the module. Supplementary File 1 includes a full description of the components and the steps for the assembly of the device. The 3D-printing and laser cut plans are available online at <https://github.com/ThomasBlein/ChronoRootModuleHardware> under the CERN Open Hardware Licence Version 2.

6.2 Computational methods

We evaluated different state-of-the-art architectures for image segmentation, and proposed new variants which achieved a good compromise between processing time, model complexity and accuracy, as discussed in the Results section. This segmentation module is followed by several post-processing stages

including a CRF post-processing to enhance label homogeneity, a temporal consistency refinement step, skeletonization, graph construction and node tracking. ChronoRoot outputs a labeled graph per image indicating which nodes correspond to the seed, main root, lateral roots, bifurcations and the root tips. For each time step, the complete RSA is saved following the RSML format [6].

6.3 Deep learning models for root segmentation

CNNs are representation learning methods with multiple abstraction levels, which compose simple but nonlinear modules transforming representations at one level into a representation at a higher, slightly more abstract level [37]. These models are specially suited for computer vision tasks, in particular for image segmentation [25]. We explored six different convolutional neural network architectures to perform plant root segmentation. Four of them are state-of-the-art existing architectures, while the other two were proposed in this work. In what follows, we first present a brief description of the state-of-the-art architectures (namely the UNet [48], ResUNet [60], SegNet [10] and DeepLab [15]) and then discuss the two models proposed in this work.

UNet: The first model is a modified lightweight version of the standard UNet [48], which employs a fully convolutional encoder-decoder architecture and produces a dense segmentation map at the pixel level. Based on the original UNet model, we implemented a lightweight version reducing by 4 the number of feature maps per convolutional layer. Skip connections were implemented via summations of the signals in the up-sampling part of the network, instead of the concatenation used in the original version. We also replaced the max-pooling layers with avg-pooling, and used an Exponential Linear Unit (ELU) as non-linearity instead of a Rectified Linear Unit (ReLU). See Supplementary Table 1 for a detailed description of the implemented architecture.

ResUNet: For the second model (ResUNet), we replaced the convolutional layers in the aforementioned UNet architecture by residual blocks [28]. Residual blocks help to prevent the degradation problem which occurs in very deep neural networks by learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. Recent works suggest that residual blocks are effective at segmenting tubular structures like plant roots or roads in a map [60]. See Supplementary Table 2 for a detailed description of the implemented architecture.

SegNet: The SegNet architecture [10] is a fully convolutional encoder-decoder neural network, widely adopted by the computer vision community to perform dense image segmentation. The architecture of the encoder is identical to the first 13 layers of VGG-16 [49] and the role of the decoder network is to map the low resolution encoder feature maps to full input resolution feature maps for pixel-wise classification. Differently from the UNet where skip connections are used to propagate the complete feature maps from the encoder to the decoder, the upsampling in the decoder part of the SegNet model uses the memorized max-pooling indices from the corresponding encoder level. Our implementation was based on a publicly available model[7].

DeepLab v3: The DeepLab V3 model [15] follows a different approach to generate dense segmentation maps. Differently from the previous models which use skip connections (UNet) or memorized max-pooling indices (SegNet), this model employs atrous convolutions with upsampled filters to extract dense feature maps and capture long range context.

6.3.1 Proposed Models

On top of these state-of-the-art architectures, we propose two different CNN models. In the first model, named DSResUNet, we aimed at improving the segmentation accuracy while keeping at the same time a fast lightweight model. In the second case, we focused on increasing the robustness and boosting the accuracy of the segmentation method, at the expense of a more complex model which follows the principle of ensemble learning.

DSResUNet: Taking the ResUNet as a baseline model, we propose here a new architecture which combines residual connections and deep supervision [38] to improve the accuracy of the results. Deep supervision integrates additional loss terms which are computed using feature maps from the

intermediate CNN layers, instead of the last one only. We concatenated the ResUNet output with the original input image, and processed these feature maps with two additional convolutional layers. This resulted in a cascade of two networks which are trained jointly, where the first one produces an initial segmentation map that is then refined by the second part of the network. We computed two loss terms, one after the output of the standard ResUNet and another one after the additional convolutions. The sum of both terms constitute the loss function used to train the DSResUNet model. See Figure 1 for a graphical illustration of the architecture, and Supplementary Table 2 for a detailed description.

Ensemble: Our final segmentation method is an ensemble model. The idea of ensembling is that we can create higher performing models by combining multiple predictors using an aggregation function. One of the most common strategies to implement ensemble models is bagging [12], where the same classifier is trained multiple times using different samples of the training set, and the final output is obtained as the average of the independent predictions. In this work, we followed a different principle which had been successfully applied in the context of medical image segmentation, where instead of combining several instances of the same model trained with different training samples, we combined different models and architectures trained with the same datasets [33]. The idea is to average out the bias infused by individual model configurations, to approximate more reliably the true posterior distribution. In the context of image segmentation, given a dataset $L = (x, y)_i$ where x is an intensity image and y the corresponding ground truth segmentation, we aim at learning the underlying conditional distribution $P(y|x)$ which maps input images x into segmentation maps y . This is commonly approximated by a model $P(y|x; \theta_m)$ which has trainable parameters, determined in our case by the neural network architecture. These parameters were learnt so that they minimize a particular loss function (see next section for more details in the loss functions used in our work) using the dataset. Given different architectures, we obtained independent estimates of and combined them following [33] approximating the posterior $P(y|x)$ as:

$$P(y|x) \approx \frac{1}{M} \sum_m^M P(y|x, \theta_m) \quad (1)$$

We implemented this ensemble of multiple models and architectures by averaging the predictions of the 5 previous models (UNet, ResUNet, DSResUNet, SegNet and DeepLab v3), obtaining a more robust and accurate segmentation method that significantly outperforms the independent instances.

Training details: All the CNN models were trained using binary cross entropy as the loss function, Adam optimizer with default parameters, learning rate of 0.0001 and weight decay = 1e-8 for UNet-like models, 1e-9 for DeepLab and 1e-10 for SegNet. The hyper-parameters were chosen by grid search using the validation data. All models were implemented in TensorFlow 1, and the source code is publicly available. The training was done on a standard workstation with Intel(R) Core(TM) i7-8700 CPU, 64 GB RAM and a NVIDIA Titan X graphics processing unit.

Since we are dealing with a relatively small dataset, data augmentation was crucial to achieve good segmentation performance. We implemented online data augmentation through a variety of patch-based augmentation procedures including addition of Gaussian noise, random Gamma corrections to simulate different lighting conditions, artificial blur and horizontal flipping. These transformations were applied to both the images and their corresponding ground-truth segmentation masks. The proposed architectures are all fully convolutional, enabling a patch-based training procedure. As this is a highly unbalanced problem (we have fewer pixels corresponding to root class than background) we implemented the following patch sampling strategy: we sampled patches from random positions centered in root pixels with the same probability as patches centered in background pixels. After performing a grid search of hyperparameters, the size of the training patches was set to 256x256 and we used batches of 8 patches. At test time, we worked with the full resolution images which can be fed to the network and processed by the fully convolutional architectures.

CRF Post-processing: The CNN segmentations are post-processed using a standard fully connected CRF[36]. The CRF operates under the hypothesis that pixels which are contiguous and have similar intensity values should be assigned the same label. We used an efficient publicly available

implementation[3] of a dense CRF[34] with Potts compatibility function and hand-tuned parameters $\theta_\alpha = 5$ and $\theta_\beta = 3$.

6.4 Graph generation and temporal consistency improvement

The CNN output can be interpreted as a soft segmentation. Since we processed temporal sequences of growing plant roots, we applied a post-processing step to improve temporal consistency using a variation of the weighted trailing average. The current segmentation and an accumulation of the previous ones are averaged to avoid losing parts of the root due to droplets or other type of occlusion. Given the current segmentation s^t at time t , and the accumulated mask up to the previous time step a^{t-1} , we compute the current map $a^t = s^t + \alpha a^{t-1}$. The weight α is chosen depending on the size of the time-step (we used $\alpha = 0.9$ in our experiments). The aim is to use the root segmentation masks obtained in previous time steps to correct for potentially missing root segments. In our experiments, we processed images every 15 minutes to ensure that the plant has not grown much between two time steps. The average helped to alleviate certain problems caused by root occlusion or water droplets, as the probability maps associated to previous frames act as memory mechanisms, resulting in more stable segmentations (see Figure 6 for a visual example).

At this point, as the user selects a Region of Interest (ROI) for each plant, the algorithm starts working one by one. We proceed to threshold the accumulated probability map for the selected plant, perform closing and opening morphological operations[27] to eliminate spurious pixels and then we select the biggest connected component as the root segmentation. Finally, we proceed to skeletonize[59] the segmentation and construct a graph that represents the root system architecture.

We run a depth first search (DFS) algorithm[18] in order to label the bifurcation and end nodes of the unlabeled root graph given by the skeletonized binary segmentation. We use the DFS algorithm, starting from a seed that can be automatically chosen as the top pixel in the plant ROI or manually specified. For assigning labels to the MR, we work based on the assumption that in early growing stages, there will only be a MR with seed (top pixel) and tip (bottom pixel). We then use nearest neighbours for matching the node graphs in the succeeding iterations. As more nodes appear deviating from the MR, they will be added as bifurcation (more than one neighbour) or lateral root tip (one neighbour, different from the MR tip). In case that one LR collides with the main root or another LR, the tip will still be a tip because of the matching process. Following this procedure, labels are assigned for the seed, main root tip, bifurcation and lateral root tip nodes. Node graph matching based on a nearest neighbor criterion was performed between the labeled nodes of successive graphs in the temporal sequence to track the evolution of the root. These graph structures allowed us to extract phenotyping features such as main root length, total lateral roots length or number of lateral roots at every temporal step. By processing the complete temporal sequence for a given root, we can obtain temporal features such as growing speed or information about the root behavior on day-night cycles, enabling the emergence of novel temporal plant phenotypes, as those shown in the Results section. Figure 7 includes several examples of RSAs extracted from images with different levels of complexity. Note that we visualize the graphs using a simplified version where only nodes corresponding to seed, bifurcation and tips are shown and connected. However, it is important to highlight that MR and LR length are computed considering the real length along the labeled skeleton, which are stored as an edge attribute in the simplified graph for optimization reasons.

7 Availability of source code and requirements

The source code corresponding to ChronoRoot, namely the deep learning model and the graph generation procedures:

- Project name: ChronoRoot: High-throughput phenotyping by deep learning reveals novel temporal parameters of plant root system architecture

- Project home page: <https://github.com/ngaggion/ChronoRoot>
- Research Resource Identifier (RRID): SCR_021259
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Python > 3.3, Anaconda, TensorFlow 1.15, PyDenseCRF
- License: GNU GPL

The source code corresponding to ChronoRoot imaging controller, namely the web interface to check and set up the image acquisition parameters:

- Project name: ChronoRoot: Module Controller
- Project home page: <https://github.com/ThomasBlein/ChronoRootControl>
- Operating system(s): GNU/Linux
- Programming language: Python
- Other requirements: NGINX, uWSGI, Python >= 3.5, Flask >= 1.1.0, APScheduler, RPi.GPIO, picamera, WTForms, smbus2
- License: OSI-approved CeCILL-2.1 license

8 Availability of supporting data and materials

All data gathered and reported in this study are available as supplementary material. The two datasets of images and annotations described in the "Datasets" section, as well as the 3D printing and laser cutting files are publicly available at <https://github.com/ThomasBlein/ChronoRootModuleHardware> under the CERN Open Hardware Licence Version 2 - Strongly Reciprocal licence. Supplementary figures and tables referenced in this work, as well as a detailed description of the hardware system are available in the annex Supplementary File 1. Snapshots of our code and other data further supporting this work are openly available in the GigaScience Repository, GigaDB[23].

8.1 Additional Files

Supplementary File 1 includes the following figures and tables:

- Supplementary Table 1: Detailed description of the UNet architecture implemented in this work.
- Supplementary Table 2: Detailed description of the Residual UNet implemented in this work and the proposed Deeply Supervised Residual UNet.
- Supplementary Figure 1: Novel time-derived parameters of RSA. Supplementary Figure 2: Novel time-derived parameters of RSA.
- Supplementary Figure 3: Low-cost device for automatic image acquisition of plant plates.
- Supplementary Figure 4: LED near-infrared panel front view and back view.
- Supplementary Figure 5: Plate support.
- Supplementary Figure 6: The camera setup.
- Supplementary Figure 7: Electronic connection of a module.

Supplementary File 2 includes a Video Abstract of this work.

9 Declarations

9.1 List of abbreviations

CL: continuous light;
CNN: convolutional neural network;
CRF: conditional random field;
DFS: depth first search;
DSResUNet: Deeply Supervised ResUNet;
ELU: exponential linear unit;
FCN: fully convolutional network;
FFT: fast Fourier Transform;
GPU: graphical processing unit;
GWAS: genome-wide association studies;
GT: ground truth;
IR: infra-red;
LD: long day;
LR: lateral root;
MR: main root;
NIR: near infra-red;
RELU: rectified linear unit;
ROI: region of interest;
RSA: root system architecture;
RSML: Root System Markup Language;
SD: standard deviation;
TR: total root;

9.2 Consent for publication

Not applicable

9.3 Competing Interests

The authors declare no competing interests.

9.4 Funding

This work was supported by grants from French State (Saclay Plant Sciences, reference n° ANR-17-EUR-0007, EUR SPS-GSR) managed by the French National Research Agency under an Investments for the Future program (reference n° ANR-11-IDEX-0003-02) to VD, TR, MC and TB; CNRS through the MITI interdisciplinary programs to TB; AXA Research Fund, ANPCyT (PICT2018-3907) and UNL (CAI+D 50220140100084LI and 50620190100145LI.) to EF; ANPCyT (PICT2019-04137) to FA; ANPCyT (PICT 2018-3384) to DM; and CNRS (Laboratoire International Associé NOCOSYM) to MC and FA.

9.5 Author's Contributions

TB, MC, EF and FA conceived the project. NG and EF designed the deep learning models. NG implemented the deep learning models, ran the numerical experiments and generated the figures. TB, VD, EL and SL designed and built the hardware system. TB and VD implemented the web control interface. TR prepared the plates with *Arabidopsis thaliana* seeds and launched the experiments for image acquisition. AC and NG annotated the images used to train the deep learning models. NG,

EF, FA, DM and TB analyzed and interpreted the results. EF, FA, TR, DM, MC, TB, NG wrote the paper.

10 Acknowledgements

We would like to thank Fablab Digiscope | LRI | UPSACLAY, and in particular Romain Di Vozzo, for fruitful discussions, his advice in the design and for the access to their digital fabrication equipment. We thank Jean-Paul Bares and Maël Jeuffrard from IPS2 for support and assembling of the ChronoRoot modules. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp used for this research.

11 Authors' information

FA, DM and EF are researchers of CONICET; NG and AC are fellows of the same institution. TB and MC are researchers and VD is an engineer of CNRS. EL and SL are technicians and TR is a fellow of University Paris-Saclay.

References

- [1] Arduino, <https://www.arduino.cc>, Accessed: 2020-12-18
- [2] International plant phenotyping network, <https://www.plant-phenotyping.org/>, Accessed: 2020-12-18
- [3] Pydensecrf implementation, <https://github.com/lucasb-eyer/pydensecrf>, Accessed: 2020-12-18
- [4] Quantitative plant, <https://www.quantitative-plant.org>, Accessed: 2020-12-18
- [5] Raspberry pi, <https://www.raspberrypi.org>, Accessed: 2020-12-18
- [6] RootSystemML home page, <http://rootssystemml.github.io/>, Accessed: 2020-12-18
- [7] Segnet implementation in tensorflow, <https://github.com/aizawan/segnet>, Accessed: 2020-12-18
- [8] Win Rhizo Website. https://regentinstruments.com/assets/winrhizo_mostrecent.html
- [9] Armengaud, P., Zambaux, K., Hills, A., Sulpice, R., Pattison, R.J., Blatt, M.R., Amtmann, A.: EZ-Rhizo: integrated software for the fast and accurate measurement of root system architecture. *Plant J.* **57**(5), 945–956 (Mar 2009)
- [10] Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional Encoder-Decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (Dec 2017)
- [11] Bontpart, T., Concha, C., Giuffrida, M.V., Robertson, I., Admkie, K., Degefu, T., Girma, N., Tesfaye, K., Haileselassie, T., Fikre, A., Fetene, M., Tsaftaris, S.A., Doerner, P.: Affordable and robust phenotyping framework to analyse root system architecture of soil-grown plants. *Plant J.* (Jun 2020)
- [12] Breiman, L.: Bagging predictors. *Machine learning* **24**(2), 123–140 (1996)

- [13] Chen, H., Giuffrida, M.V., Doerner, P., Tsiftaris, S.A.: Blind inpainting of large-scale masks of thin structures with adversarial and reinforcement learning. CoRR **abs/1912.02470** (2019), <http://arxiv.org/abs/1912.02470>
- [14] Chen, H., Valerio Giuffrida, M., Doerner, P., Tsiftaris, S.A.: Adversarial large-scale root gap inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
- [15] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
- [16] Clark, R.T., Famoso, A.N., Zhao, K., Shaff, J.E., Craft, E.J., Bustamante, C.D., McCouch, S.R., Aneshansley, D.J., Kochian, L.V.: High-throughput two-dimensional root system phenotyping platform facilitates genetic analysis of root growth and development. Plant Cell Environ. **36**(2), 454–466 (Feb 2013)
- [17] Coppens, F., Wuyts, N., Inzé, D., Dhondt, S.: Unlocking the potential of plant phenotyping data through integration and data-driven approaches. Current Opinion in Systems Biology **4**, 58–63 (Aug 2017)
- [18] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction To Algorithms. MIT Press (2001)
- [19] Dhondt, S., Wuyts, N., Inzé, D.: Cell to whole-plant phenotyping: the best is yet to come. Trends Plant Sci. **18**(8), 428–439 (Aug 2013)
- [20] van Dusschoten, D., Metzner, R., Kochs, J., Postma, J.A., Pflugfelder, D., Bühler, J., Schurr, U., Jahnke, S.: Quantitative 3D analysis of plant roots growing in soil using magnetic resonance imaging. Plant Physiol. **170**(3), 1176–1188 (Mar 2016)
- [21] Fisahn, J., Yazdanbakhsh, N., Klingele, E., Barlow, P.: Arabidopsis thaliana root growth kinetics and lunisolar tidal acceleration. New Phytol. **195**(2), 346–355 (Jul 2012)
- [22] French, A., Ubeda-Tomás, S., Holman, T.J., Bennett, M.J., Pridmore, T.: High-throughput quantification of root growth using a novel image-analysis tool. Plant physiology **150**(4), 1784–1795 (2009)
- [23] Gaggion, N., Ariel, F., Daric, V., Éric Lambert, Legendre, S., Roulé, T., Camoirano, A., Milone, D.H., Crespi, M., Blein, T., Ferrante, E.: Supporting data for "chronoroot: High-throughput phenotyping by deep segmentation networks reveals novel temporal parameters of plant root system architecture". GigaScience Database (2021), <http://dx.doi.org/10.5524/100911>
- [24] Galkovskyi, T., Mileyko, Y., Bucksch, A., Moore, B., Symonova, O., Price, C.A., Topp, C.N., Iyer-Pascuzzi, A.S., Zurek, P.R., Fang, S., Harer, J., Benfey, P.N., Weitz, J.S.: GiA roots: software for the high throughput analysis of plant root system architecture. BMC Plant Biol. **12**, 116 (Jul 2012)
- [25] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J.: A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing **70**, 41–65 (2018)
- [26] van Gelderen, K., Kang, C., Pierik, R.: Light signaling, root development, and plasticity. Plant Physiol. **176**(2), 1049–1060 (Feb 2018)
- [27] Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Higher Ed (Nov 2011)

- [28] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
- [29] Hund, A., Trachsel, S., Stamp, P.: Growth of axile and lateral roots of maize: I development of a phenotyping platform. *Plant and Soil* **325**(1-2), 335–349 (2009)
- [30] Iijima, M., Matsushita, N.: A circadian and an ultradian rhythm are both evident in root growth of rice. *J. Plant Physiol.* **168**(17), 2072–2080 (Nov 2011)
- [31] Ingram, P.A., Malamy, J.E.: Root System Architecture, *Advances in Botanical Research*, vol. 55, pp. 75–117. Elsevier (2010)
- [32] Iyer-Pascuzzi, A.S., Symonova, O., Mileyko, Y., Hao, Y., Belcher, H., Harer, J., Weitz, J.S., Benfey, P.N.: Imaging and analysis platform for automatic phenotyping and trait ranking of plant root systems. *Plant Physiol.* **152**(3), 1148–1157 (Mar 2010)
- [33] Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. In: *International MICCAI Brainlesion Workshop*. pp. 450–462. Springer (2017)
- [34] Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: *Advances in neural information processing systems*. pp. 109–117 (2011)
- [35] Kuijken, R.C.P., van Eeuwijk, F.A., Marcelis, L.F.M., Bouwmeester, H.J.: Root phenotyping: from component trait in the lab to breeding. *J. Exp. Bot.* **66**(18), 5389–5401 (Sep 2015)
- [36] Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *ICML: Proceedings of the Eighteenth International Conference on Machine Learning* (2001)
- [37] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (May 2015)
- [38] Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: *Artificial intelligence and statistics*. pp. 562–570 (2015)
- [39] Lobet, G.: Image analysis in plant sciences: Publish then perish. *Trends Plant Sci.* **22**(7), 559–566 (Jul 2017)
- [40] Lobet, G., Draye, X., Périlleux, C.: An online database for plant image analysis software tools. *Plant Methods* **9**(1), 38 (Oct 2013)
- [41] Lobet, G., Paez-Garcia, A., Schneider, H., Junker, A., Atkinson, J.A., Tracy, S.: Demystifying roots: A need for clarification and extended concepts in root phenotyping. *Plant Sci.* **282**, 11–13 (May 2019)
- [42] Minervini, M., Giuffrida, M.V., Perata, P., Tsafaris, S.A.: Phenotiki: an open software and hardware platform for affordable and easy image-based phenotyping of rosette-shaped plants. *Plant J.* **90**(1), 204–216 (Apr 2017)
- [43] Naeem, A., French, A.P., Wells, D.M., Pridmore, T.P.: High-throughput feature counting and measurement of roots. *Bioinformatics* **27**(9), 1337–1338 (May 2011)
- [44] Narisetti, N., Henke, M., Seiler, C., Shi, R., Junker, A., Altmann, T., Gladilin, E.: Semi-automated root image analysis (saRIA). *Sci. Rep.* **9**(1), 19674 (Dec 2019)
- [45] Orlando, J.I., Manterola, H.L., Ferrante, E., Ariel, F.: Arabidopsis roots segmentation based on morphological operations and crfs. *arXiv preprint arXiv:1704.07793* (2017)

- [46] Palmer, C.M., Bush, S.M., Maloof, J.N.: Phenotypic and developmental plasticity in plants. In: John Wiley & Sons, Ltd (ed.) *eLS*, vol. 59, p. 1127. John Wiley & Sons, Ltd, Chichester, UK (May 2001)
- [47] Pound, M.P., Atkinson, J.A., Townsend, A.J., Wilson, M.H., Griffiths, M., Jackson, A.S., Bulat, A., Tzimiropoulos, G., Wells, D.M., Murchie, E.H., Pridmore, T.P., French, A.P.: Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Giga-science* **6**(10), 1–10 (Oct 2017)
- [48] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)
- [49] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations* (2015)
- [50] Slovak, R., Göschl, C., Su, X., Shimotani, K., Shiina, T., Busch, W.: A scalable Open-Source pipeline for Large-Scale root phenotyping of arabidopsis. *Plant Cell* **26**(6), 2390–2403 (Jun 2014)
- [51] Teramoto, S., Takayasu, S., Kitomi, Y., Arai-Sanoh, Y., Tanabata, T., Uga, Y.: High-throughput three-dimensional visualization of root system architecture of rice using x-ray computed tomography. *Plant Methods* **16**, 66 (May 2020)
- [52] Tracy, S.R., Nagel, K.A., Postma, J.A., Fassbender, H., Wasson, A., Watt, M.: Crop improvement from phenotyping roots: Highlights reveal expanding opportunities. *Trends Plant Sci.* **25**(1), 105–118 (Jan 2020)
- [53] Valle, B., Simonneau, T., Boulord, R., Sourd, F., Frisson, T., Ryckewaert, M., Hamard, P., Brichet, N., Dauzat, M., Christophe, A.: PYM: a new, affordable, image-based method using a raspberry pi to phenotype plant leaf area in a wide diversity of environments. *Plant Methods* **13**, 98 (Nov 2017)
- [54] Wells, D.M., French, A.P., Naeem, A., Ishaq, O., Traini, R., Hijazi, H., Bennett, M.J., Pridmore, T.P.: Recovering the dynamics of root growth and development using novel image acquisition and analysis methods. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1595), 1517–1524 (2012)
- [55] Wu, J., Wu, Q., Pagès, L., Yuan, Y., Zhang, X., Du, M., Tian, X., Li, Z.: RhizoChamber-Monitor: a robotic platform and software enabling characterization of root growth. *Plant Methods* **14**, 44 (Jun 2018)
- [56] Yasrab, R., Atkinson, J.A., Wells, D.M., French, A.P., Pridmore, T.P., Pound, M.P.: RootNav 2.0: Deep learning for automatic navigation of complex plant root architectures. *Gigascience* **8**(11) (Nov 2019)
- [57] Yazdanbakhsh, N., Sulpice, R., Graf, A., Stitt, M., Fisahn, J.: Circadian control of root elongation and C partitioning in arabidopsis thaliana. *Plant Cell Environ.* **34**(6), 877–894 (Jun 2011)
- [58] Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G.: User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**(3), 1116–1128 (Jul 2006)
- [59] Zhang, T., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM* **27**(3), 236–239 (1984)
- [60] Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters* **15**(5), 749–753 (2018)

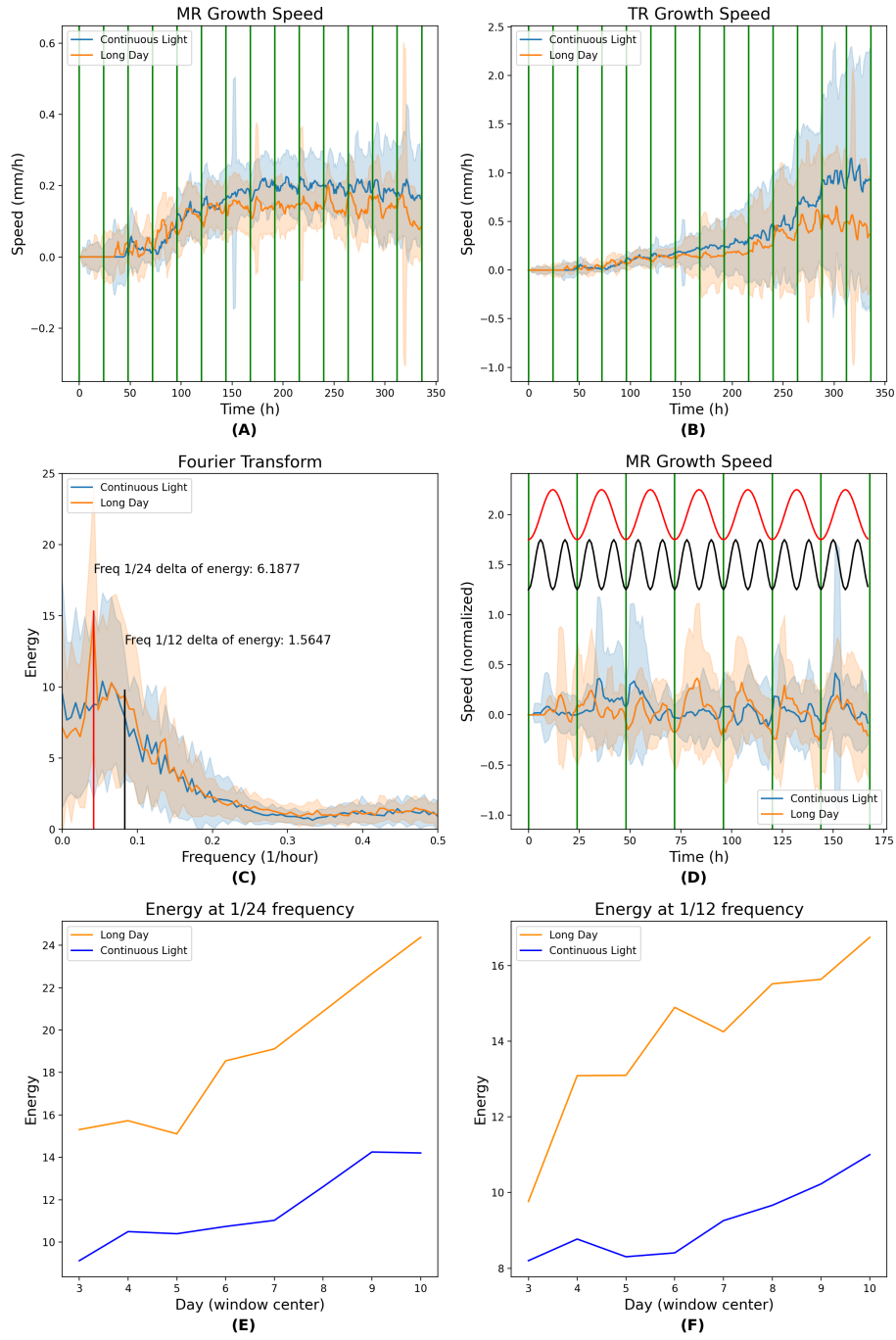


Figure 5: Novel time-derived parameters of RSA. A. Main root (MR) and B. Total root (TR) growth speed along time. C. Fourier Fast Transform of the growth speed signal of MR. The largest energy differences are indicated in the graph. D. The post-processed (high-pass followed by normalization) MR growth speed showing a 7-day-window centered on day 3. The sine curves corresponding to the frequency 1/12 (in black) and 1/24 (in red) found in C are indicated at the top. Note the correlation between the LD growth speed oscillation and the two components 1/12 and 1/24. E. The energy at 1/24 frequency calculated in a 7-day-window centered at consecutive time points for MR. F. The energy at 1/12 frequency calculated in a 7-day-window centered at consecutive time points for MR.

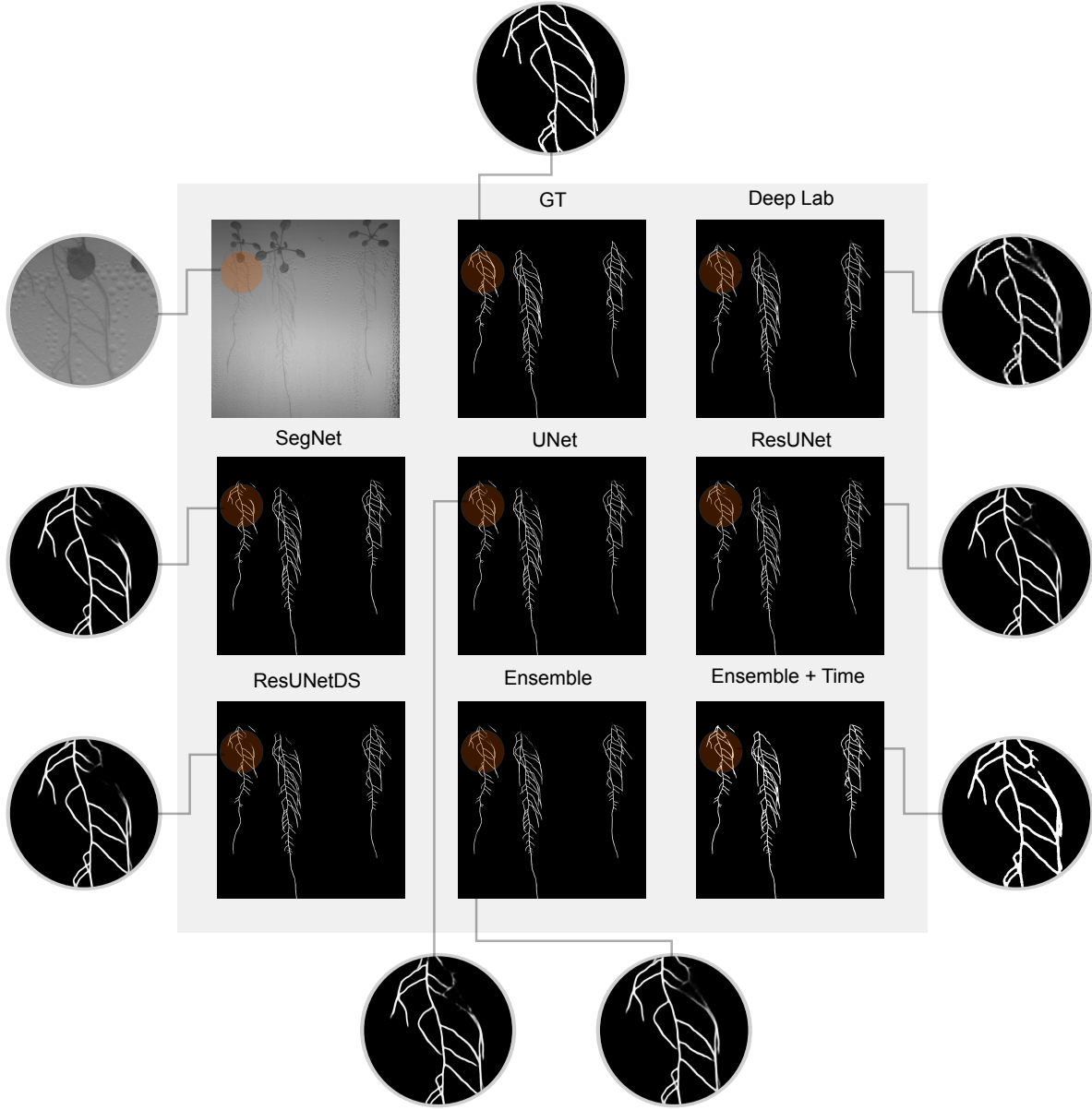


Figure 6: Qualitative segmentation results obtained with the benchmarked methods. We observe how the ensemble and the ensemble with temporal consistency improve the quality of the results, specially in areas with root occlusion.

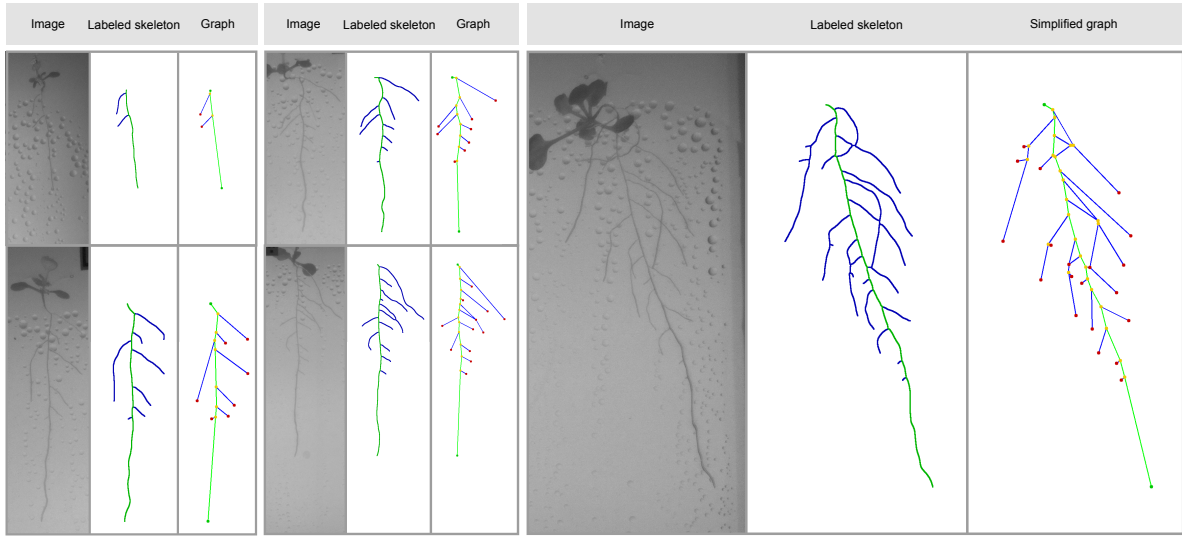


Figure 7: Examples of images, labeled skeleton and simplified graphs corresponding to RSAs exhibiting different levels of complexity. Note that we visualize the graphs using a simplified version where only nodes corresponding to seed, bifurcation and tips are shown and connected. However, since the full skeleton is labeled, the MR and LR length are computed considering the real length along the skeleton.

Supplementary Files for ChronoRoot:
High-throughput phenotyping by deep segmentation networks
reveals novel temporal parameters of plant root system
architecture

CNN Architectures

U-Net			
Block	Input Dimensions	Feature Maps	Output Dimensions
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	256x256	16	256x256x16
Average Pooling 2x2	256x256x16		128x128x16
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	128x128x16	32	128x128x32
Average Pooling 2x2	128x128x32		64x64x32
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	64x64x32	64	64x64x64
Average Pooling 2x2	64x64x32		32x32x64
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	32x32x64	128	32x32x128
Dropout Layer			
Upconv	32x32x128	64	64x64x64
Elu(Skip Connection (ADD))	64x64x64		64x64x64
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	64x64x64	64	64x64x64
Dropout Layer			
Upconv	64x64x64	32	128x128x32
Elu(Skip Connection (ADD))	128x128x32		128x128x32
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	128x128x32	32	128x128x32
Dropout Layer			
Upconv	128x128x32	16	256x256x16
Elu(Skip Connection (ADD))	256x256x16		256x256x16
Conv / Batch Norm /Elu Conv / Batch Norm /Elu	256x256x16	16	256x256x16
Conv	256x256x16		256x256x2

Table 1: Detailed description of the UNet architecture implemented in this work.

DS Residual U-Net				
Block	Input Dimensions	Feature Maps	Output Dimensions	Model
Residual Block	256x256	16	256x256x16	Residual U-Net
Max Pooling 2x2	256x256x16		128x128x16	
Residual Block	128x128x16	32	128x128x32	
Max Pooling 2x2	128x128x32		64x64x32	
Residual Block	64x64x32	64	64x64x64	
Max Pooling 2x2	64x64x32		32x32x64	
Residual Block	32x32x64	128	32x32x128	
Dropout Layer				
Upconv	32x32x128	64	64x64x64	
Elu(Skip Connection (ADD))	64x64x64		64x64x64	
Residual Block	64x64x64	64	64x64x64	
Dropout Layer				
Upconv	64x64x64	32	128x128x32	
Elu(Skip Connection (ADD))	128x128x32		128x128x32	
Residual Block	128x128x32	32	128x128x32	
Dropout Layer				
Upconv	128x128x32	16	256x256x16	
Elu(Skip Connection (ADD))	256x256x16		256x256x16	
Residual Block	256x256x16	16	256x256x16	
Conv (ResUNet OUT)	256x256x16		256x256x2	
Concat (ResUNet OUT) with original input	256x256x2		256x256x3	Deeply Supervised Residual U-Net
Residual Block	256x256x3	16	256x256x16	
Max Pooling 2x2	256x256x16		128x128x16	
Residual Block	128x128x16	32	128x128x32	
Dropout Layer				
Upconv	128x128x32	16	256x256x16	
Elu(Skip Connection (ADD))	256x256x16		256x256x16	
Residual Block	256x256x16	16	256x256x16	
Conv (DSResUNet OUT)	256x256x16		256x256x2	

Table 2: Detailed description of the Residual U-Net implemented in this work and the proposed Deeply Supervised Residual U-Net.

Supplementary Figures

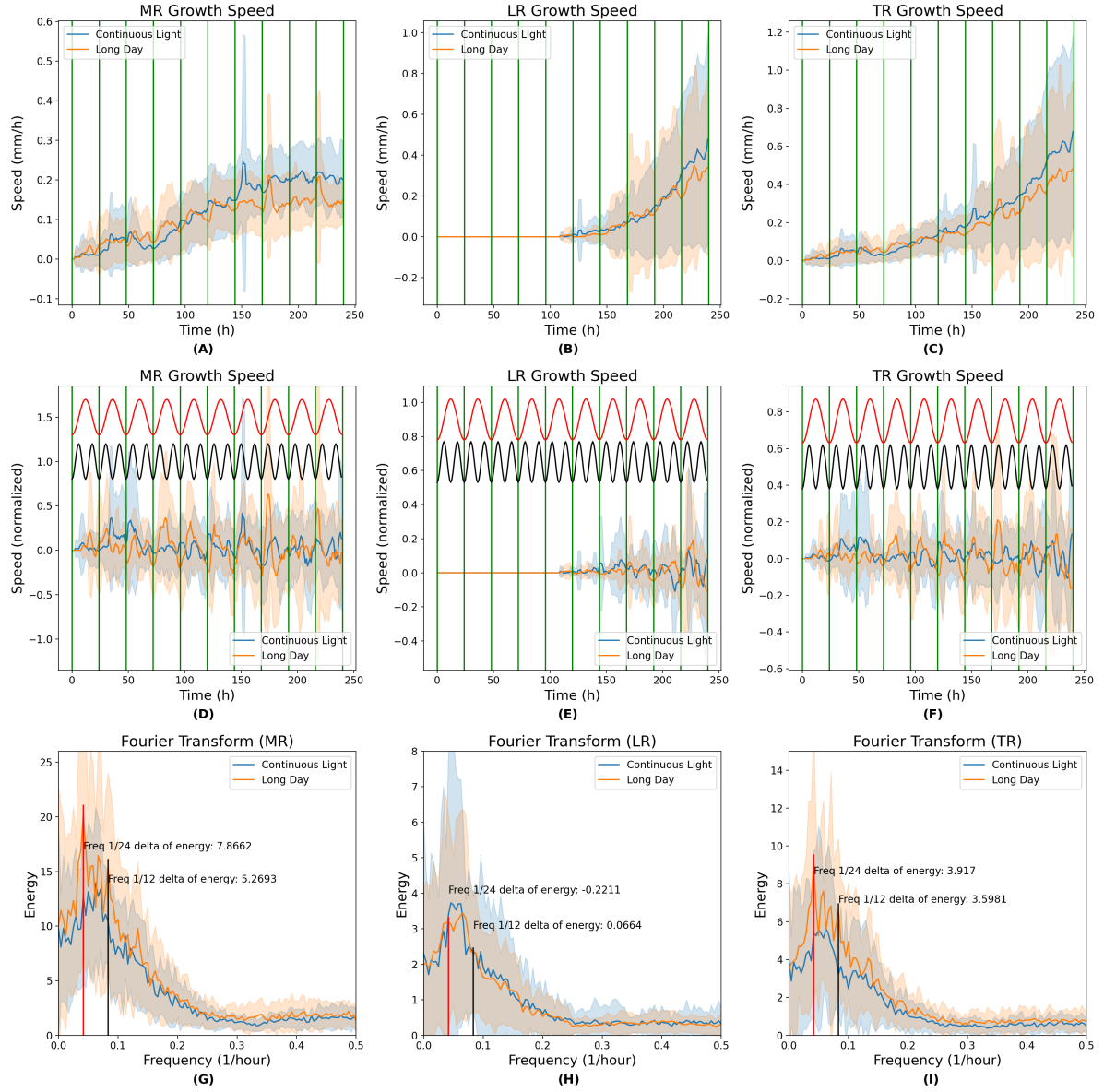


Figure 1: **Novel time-derived parameters of RSA.** A. Main root (MR), B. Lateral Roots (LR) and C. Total root (TR) growth speed along time. D-F. Post-processed speed along time (application of a high-pass filter followed by normalization) of (D) MR, (E) LR and (F) TR, respectively. The sine curves corresponding to the frequency 1/12 (in black) and 1/24 (in red) found in D-F are indicated at the top of each panel. G-I. Fourier Fast Transform of the growth speed signal of (G) MR, (H) LR and (I) TR, respectively.

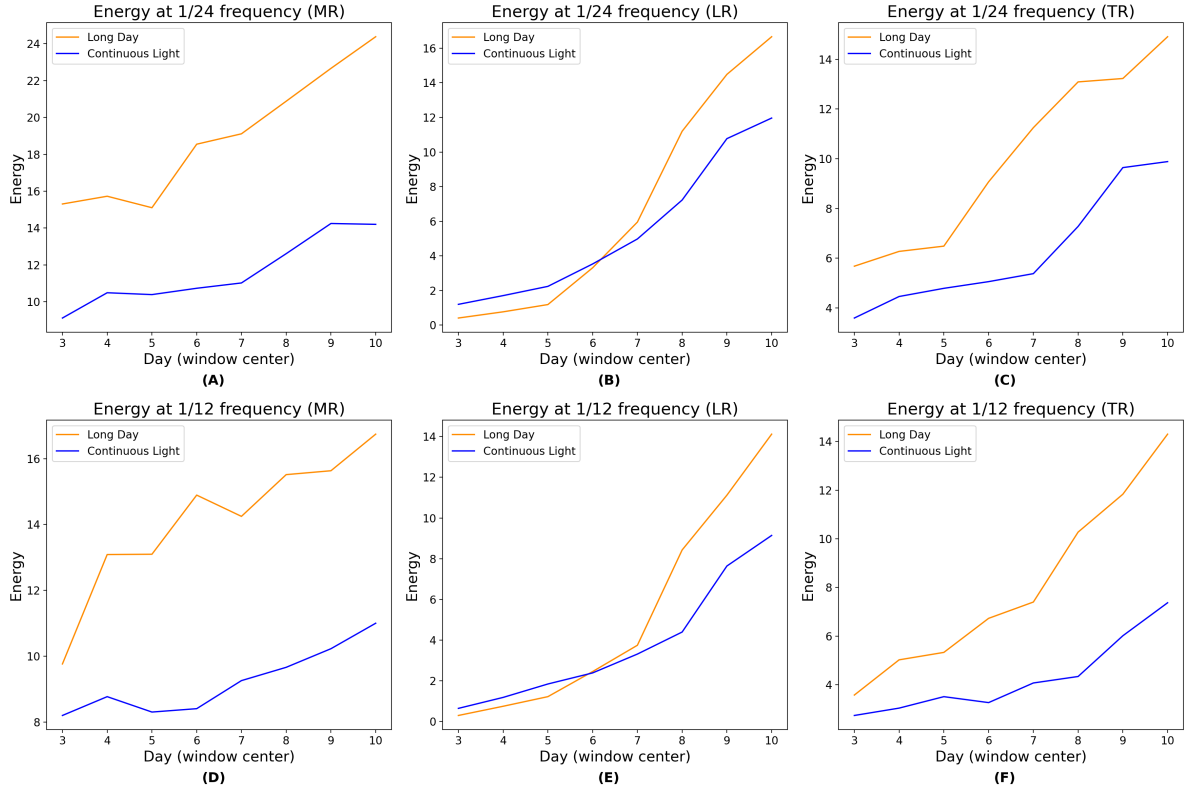


Figure 2: **Novel time-derived parameters of RSA.** A-C. The energy at 1/24 frequency calculated in a 7-day-window centered at consecutive time points for (A) MR, (B) LR and (C) TR. D-F. The energy at 1/12 frequency calculated in a 7-day-window centered at consecutive time points for (D) MR, (E) LR and (F) TR.

Hardware Description

In this section we include supplementary notes with additional details about the hardware specification. These notes are provided as open-hardware specifications, to encourage other scientists to 3D print and mount the device in their own laboratories.

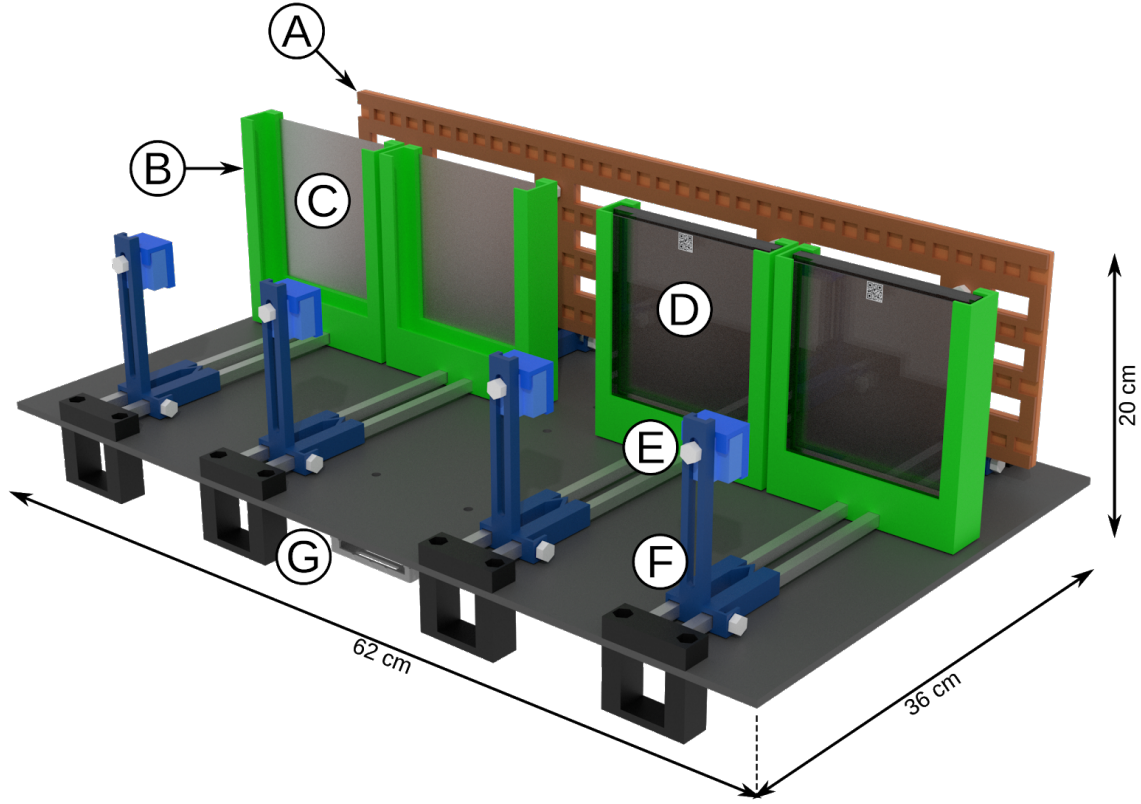


Figure 3: **Low-cost device for automatic image acquisition of plant plates.** 3D rendering of a ChronoRoot module. (A) mobile and controlled infrared (IR) backlight (850 nm), (B) mobile plate support, (C) diffusing filter, (D) 12 cm x 12 cm square plate with QR code on the top, (E) Camera case equipped with an IR long pass filter (> 830 nm), (F) mobile camera support, (G) Raspberry Pi computer controlling the module (IR backlight and camera).

The NIR illumination was built with four rows of LED flexible tape (tri-chip SMD5050-150-IR 850 nm, Huake LTD, China) fixed in a sandwich between two acrylic plates (580 mm x 140 mm x 3 mm) laser cuts to allow air flow and prevent the LED strip from ungluing. The four strips of LED were connected in parallel to a 12V AC/DC adaptor fixed under the module. The power supply cable of the adaptor is under the control of relay (Single Relay Board #27115, Parallax Inc), in a box fixed near the adaptor. It allows the control of the NIR illumination by the computer. The LED array is maintained vertically by L-shaped supports able to move along the aluminum horizontal axes as four individual panels in the module, corresponding to the respective plant plates.

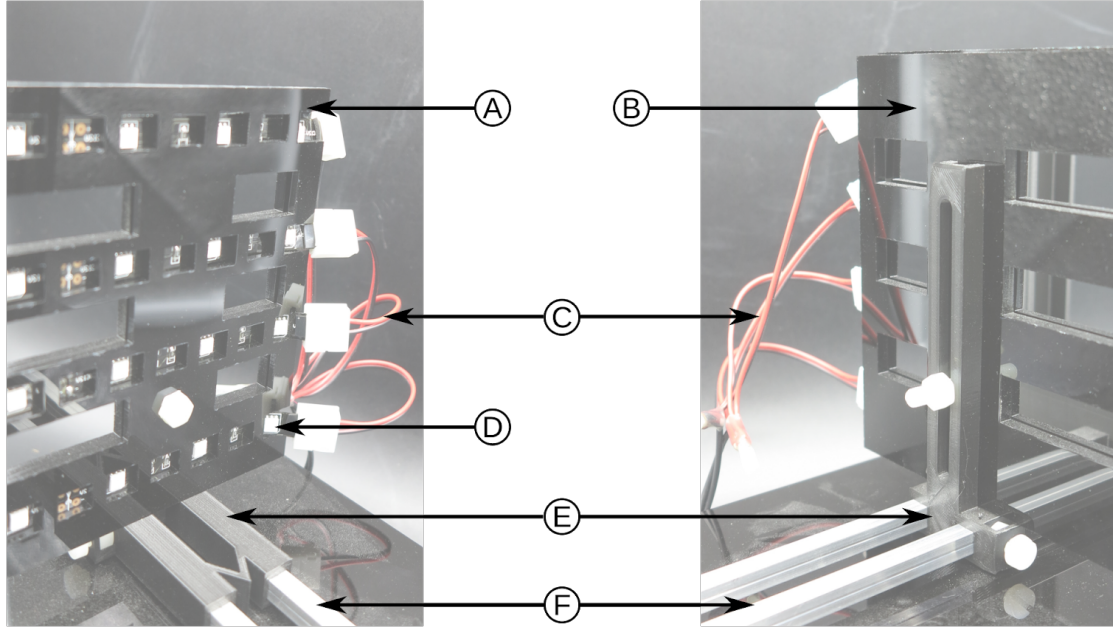


Figure 4: **LED near-infrared** panel front view (left) and back view (right). The LED strips (D) are squeezed in between two acrylic panels (A and B) to prevent them from unsticking from the back panel. The strips are linked in parallel to the AC/DC adaptor through strip connector (C). The full LED panel is kept in vertical position via an L-shaped support (E) able to move along the double aluminum axis (F). Acrylic screws and nuts are used to maintain the LED panel on the L-shape support and fixed the horizontal position on the horizontal axis. U-shaped plate holders were designed to hold 12.5 cm x 12.5 cm square petri dishes used to grow the plants. The plate holder carries a diffusion filter (Cinegel R3000 Tough Rolux, Rosco Laboratories Inc.) at the back of the plate to allow homogeneous backlight illumination of the plates by the NIR LED array. The four plate holders can be moved independently along their respective horizontal axis.



Figure 5: **Plate support.** Annotated 3D rendering (left) and picture (right) of the plate support. The U-shaped plate support (A) carries a diffusion filter at the back (B) allowing an homogeneous near-infrared illumination. Each support moves horizontally on the double axis (C). A classical 12.5 cm x 12.5 cm square plate fit into the plate support (E).

Each camera is a regular Raspberry Pi NoIR V2 module (Sony IMX219 8-megapixel sensor) which is able to capture NIR wavelengths in addition to the classical visible spectrum. It is positioned in a box that can be moved vertically on an L-shaped support which can be shifted along the horizontal axis. The focus of the camera is adjusted manually before the start of the experiment by turning the objective lens using the provided crown. To be able to record pictures exclusively from the NIR spectrum, an IR long pass filter (12.5 mm diameter, RG830 Schott AG) can be positioned in front of the camera to exclude light below 830 nm and provide consistent images independently of the lightning of the growth chamber.

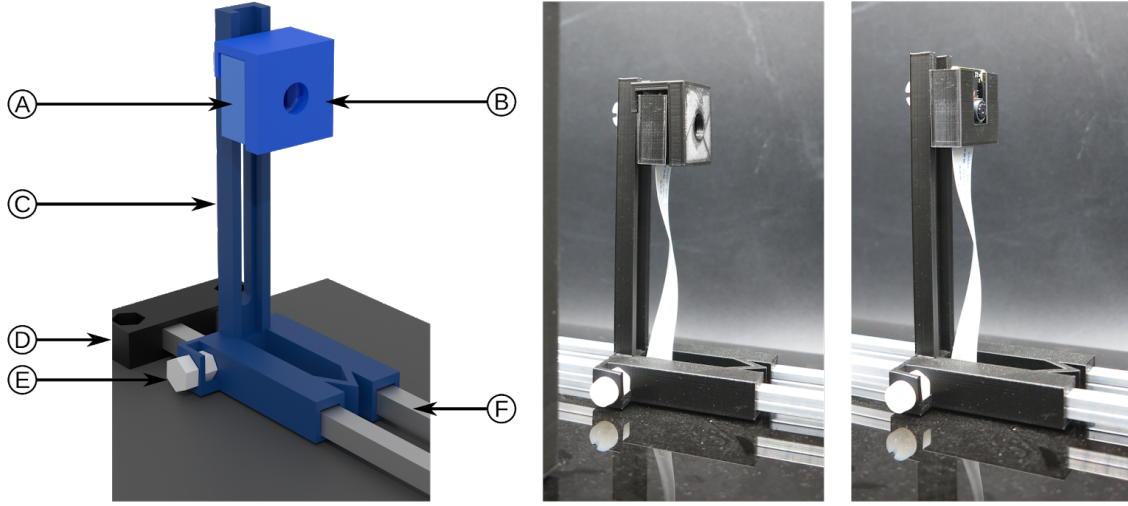


Figure 6: **The camera setup.** (Left) annotated 3D rendering; (middle) camera on a module carrying a filter; (right) camera on a module without the filter. A Raspberry Pi NoIR V2 module is enclosed in a 3D printed box (A) that could carry an optional filter (B). The box can be moved vertically on an L-shaped support (C) which itself is able to move along a double aluminum axis (F); The position of the L-shape support and of the camera box are secured by acrylic screw and nut (E for L-shaped support). The double aluminum axis is attached to the main board by a 3D printed part (D). The computer is a Raspberry Pi 3 model B with an additional camera multiplexer module (IVPort V2 Raspberry Pi Camera Module V2 Multiplexer, Ivmech Mekatronik & Inovasyon Ltd.) and allows to connect four cameras. It is boxed in a cage fixed under the main board. The computer runs the Raspbian GNU/Linux operating system and a python web application was developed to control the different components of the module. The camera multiplexer module is controlled through the `ivport_v2` library (<https://github.com/ivmech/ivport-v2>), the control of relay of the NIR backlight through pins 30 (ground) and 32 (GPIO 12) of the Raspberry Pi GPIO through the `RPi.GPIO` library (<https://pypi.org/project/RPi.GPIO/>). The web interface was developed using the Flask web framework (<https://flask.palletsprojects.com/>), uWSGI (<https://uwsgi-docs.readthedocs.io/>), Twitter Bootstrap for the frontend (<https://getbootstrap.com/>). The scheduling of the task was managed using the APScheduler library (<https://apscheduler.readthedocs.io/>). The source code of the application is available at <https://laforge.ips2.u-psud.fr/thomas.blein/kinematicrobot>.

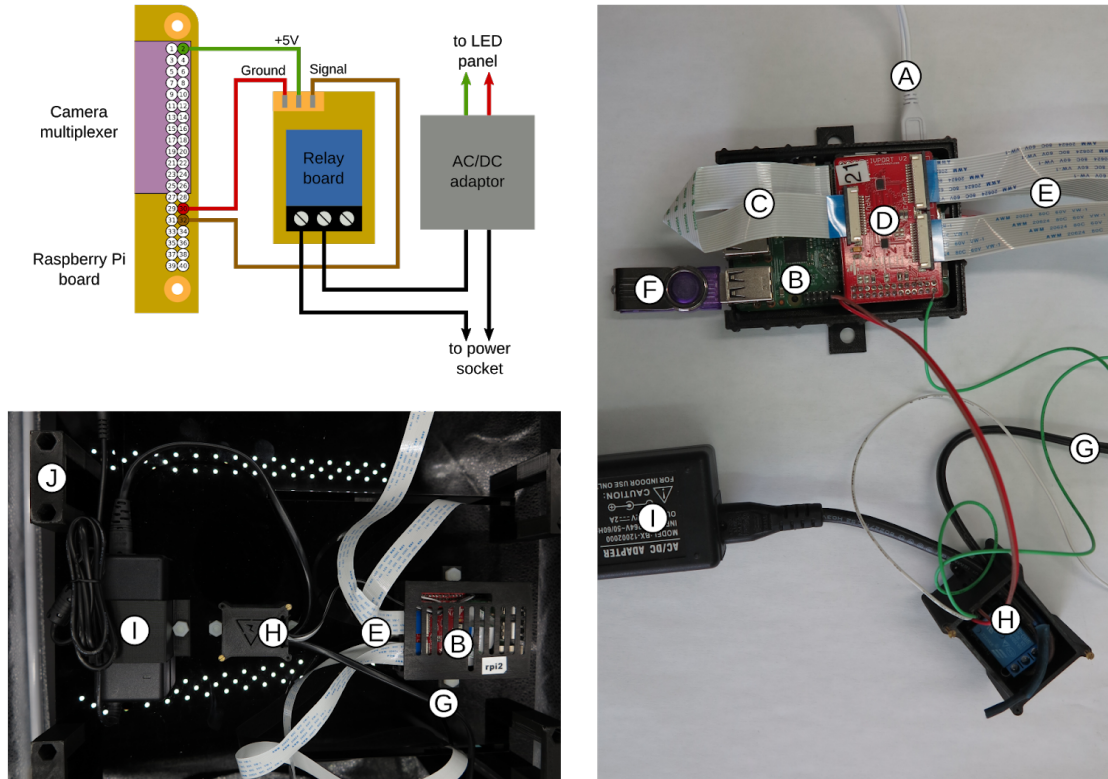


Figure 7: **Electronic connection of a module.** Schematic electric connection between the Raspberry Pi GPIO, the camera multiplexer, the relay board and AC/DC adaptor (up left). Real view of the connection with different part boxes open (right). The Raspberry Pi (B) is connected to a USB power adaptor (A). The camera multiplexer module (D) is plugged on the Raspberry Pi GPIO port and connected to the Raspberry Pi camera by a strip (C). Each camera is connected to the multiplexer by a strip (E). The pictures are saved on a USB stick (F) or a USB hard drive directly connected to the Raspberry Pi. A relay board (H) is connected to the Raspberry Pi GPIO and controls the main power (G) of the AC/Dc adaptor powering the LED strip. The final organization of the controlling part under the module main board raised by feet (J).

Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

Nicolás Gaggion¹, Lucas Mansilla¹, Diego H. Milone¹, and Enzo Ferrante¹

¹Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET, Universidad Nacional del Litoral, Santa Fe, Argentina

Abstract

In this work we address the problem of landmark-based segmentation for anatomical structures. We propose HybridGNet, an encoder-decoder neural architecture which combines standard convolutions for image feature encoding, with graph convolutional neural networks to decode plausible representations of anatomical structures. We benchmark the proposed architecture considering other standard landmark and pixel-based models for anatomical segmentation in chest x-ray images, and found that HybridGNet is more robust to image occlusions. We also show that it can be used to construct landmark-based segmentations from pixel level annotations. Our experimental results suggest that Hybrid-Net produces accurate and anatomically plausible landmark-based segmentations, by naturally incorporating shape constraints within the decoding process via spectral convolutions.

1 Introduction

Deep learning models based on convolutional neural networks have become the state-of-the-art for anatomical segmentation of biomedical images. The current practise is to employ standard convolutional neural networks (CNNs) trained to minimize a pixel level loss function, where dense segmentation masks are used as ground truth. Casting image segmentation as a pixel labeling problem is desirable in scenarios like lesion segmentation, where topology and location do not tend to be preserved across individuals. However, organs and anatomical structures usually present a characteristic topology which tends to be regular. Differently from dense segmentation masks, statistical shape models [15] and graph-based representations [4] provide a natural way to incorporate topological constraints by construction. Moreover, such shape representations make it easier to establish landmark correspondences among individuals, particularly important in the context of statistical shape analysis.

Since the early 1990's, variations of point distribution models (PDMs) have been proposed [8] to segment anatomical structures using landmarks. PDMs are flexible shape templates describing how the relative locations of important points can vary. Techniques based on PDMs, like active shape models (ASM) [8, 31] and active appearance models (AAM) [7] became the defacto standard to deal with anatomical segmentation at the end of the century. Subsequently, during the next decade, the development of more powerful and robust image registration algorithms [32] positioned deformable template matching algorithms as the choice of option for anatomical segmentation and atlas construction [11, 16, 25]. More recently, with the advent of deep fully convolutional networks [28, 27], great efforts were made to incorporate anatomical constraints into such models [17, 20, 24]. The richness of the image features learned by standard CNNs allowed them to achieve highly accurate results. However, most of these methods work directly on the pixel space, producing acceptable dense segmentations masks but without landmark annotations and connectivity structure. On the contrary, structured models like graphs appear as a natural way to represent landmarks, contours and surfaces. By defining the

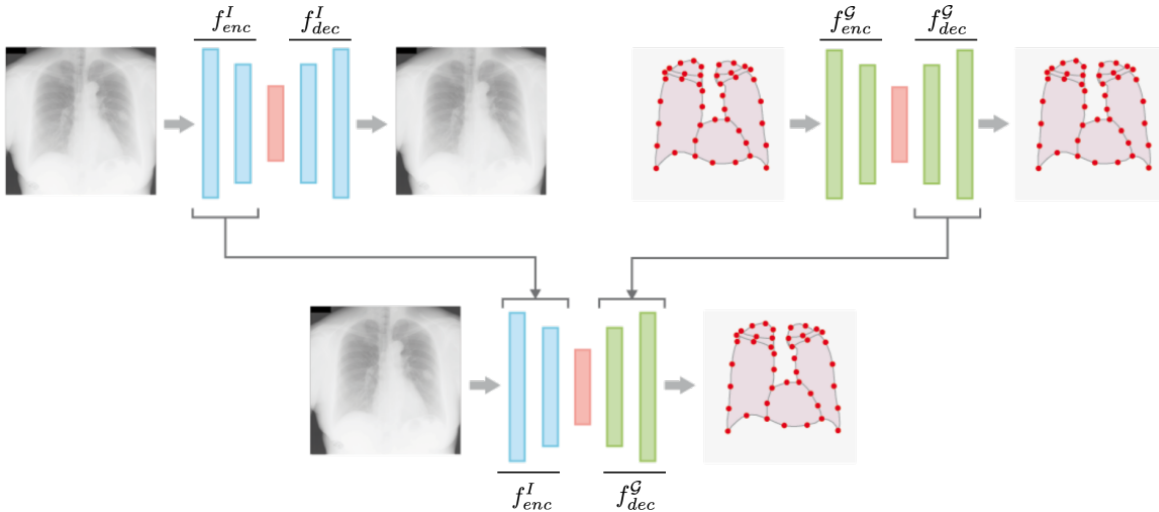


Figure 1: The proposed HybridGNet (bottom) is an encoder-decoder architecture which combines standard convolutions for image feature encoding (blue), with graph spectral convolutions (green) to decode plausible anatomical graph-based representations.

landmark position as a function on the graph nodes, and encoding the anatomical structure through its adjacency matrix, we can easily constrain the space of solutions and ensure topological correctness.

During the last years, the emerging field of geometric deep learning [5] extended the success of convolutional neural networks to non-Euclidean domains like graphs and meshes. While classical CNNs have been particularly successful when dealing with signals such as images or speech, more recent developments like spectral convolutions [6, 9] and neural message passing [12] enabled the use of deep learning on graphs. Recently, graph generative models were proposed [19, 26]. Of particular interest for this work is the convolutional mesh autoencoder proposed in [26]. The authors construct an encoder-decoder network using spectral graph convolutions, and train it in a variational setting using face meshes. By sampling the latent space, they are able to generate new expressive faces never seen during training. Inspired by this idea, we propose to exploit the generative power of convolutional graph autoencoders [26, 10] to decode plausible anatomical segmentations from images.

Contributions. In this work, we revisit landmark-based segmentation in light of the latest developments on deep learning for Euclidean and non-Euclidean data. We aim at leveraging the best of both worlds, combining standard convolutions for image feature encoding, with generative models based on graph convolutional neural networks (GCNN) to decode plausible representations of anatomical structures. Under the hypothesis that encoding connectivity information through the graph adjacency matrix will result in richer representations than standard landmark-based PDMs, we also compare our results with other statistical point distribution models which do not make explicit use of the graph connectivity. Our contributions are 4-fold: (1) we propose HybridGNet, an encoder-decoder architecture which combines standard convolutions with GCNNs to extract graph representations directly from images; (2) we showcase the proposed architecture in the context of anatomical landmark-based segmentation of chest x-ray images and benchmark the results considering other landmark and pixel-based segmentation models; (3) we propose to use HybridGNet to create graph-based representations from dense anatomical masks without paired images and (4) we show that HybridGNet is more robust to image occlusions than state-of-the-art pixel-level segmentation methods.

2 Hybrid graph convolutional neural networks

Problem setting. Let us have a dataset $\mathcal{D} = \{(I, \mathcal{G})_k\}_{0 < k < N}$, composed of N dimensional images I and their corresponding landmark-based segmentation represented as a graph $\mathcal{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$. V is the set of nodes (landmarks), $\mathbf{A} \in \{0, 1\}^{|V| \times |V|}$ is the adjacency matrix indicating the connectivity between pairs of nodes ($\mathbf{A}_{ij} = 1$ indicates an edge connecting vertices i and j , and $\mathbf{A}_{ij} = 0$ otherwise) and $\mathbf{X} \in \mathbb{R}^{|V| \times d}$ is a function (represented as a matrix) assigning a feature vector to every node. In our case, it assigns a d -dimensional spatial coordinate to every landmark. Without loss of generality, in this work we will showcase the proposed framework in 2D images (thus $\mathbf{X} \in \mathbb{R}^{|V| \times 2}$). However, note that extending our method to 3D images is straightforward, the only difference being that images I will be volumes and graphs $\mathcal{G}$, meshes.

In the context of landmark-based segmentation and point distribution models, it is common to have manual annotations with fixed number of points. Therefore, for all graphs, we assume that the set of nodes (landmarks) V and the connectivity matrices $\mathbf{A}$ are the same. The only difference among them is given by the spatial coordinates defined in $\mathbf{X}$. This assumption enables the use of spectral graph convolutions [9, 26] to learn latent representations of anatomy.

Spectral graph convolutions. We have previously mentioned that extending discretized convolutions to graph structures is not straightforward. A myriad of graph convolution formulations have been recently proposed [5, 12]. Here we adopt the localized spectral version since it has shown to be effective at learning powerful shape representations from graph structures with a fixed number of nodes [26]. Spectral convolutions are build using the eigendecomposition of the graph Laplacian matrix $\mathbf{L}$, exploiting the property that convolutions in the node domain are equivalent to multiplications in the graph spectral domain [30].

The graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$, and $\mathbf{A}$ is the adjacency matrix. The Laplacian $\mathbf{L}$ can be decomposed as $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{|V| \times |V|} = [u_0, u_1, \dots, u_{|V|-1}]$ is the matrix of eigenvectors (Fourier basis) and $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_{|V|-1})$ the diagonal matrix of eigenvalues (frequencies of the graph). By analogy with the classical Fourier transform for continuous or discrete signals, the graph Fourier transform of a function $\mathbf{X}$ defined on the graph domain is $\hat{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, while its inverse is given by $\mathbf{X} = \mathbf{U} \hat{\mathbf{X}}$. Based on this formulation, the spectral convolution between a signal $\mathbf{X}$ and a filter $\mathbf{g}_\phi = \text{diag}(\phi)$ is defined as $\mathbf{g}_\phi * \mathbf{X} = \mathbf{g}_\phi(\mathbf{L})\mathbf{X} = \mathbf{g}_\phi(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)\mathbf{X} = \mathbf{U}\mathbf{g}_\phi(\mathbf{\Lambda})\mathbf{U}^T\mathbf{X}$, where $\phi \in \mathbb{R}^n$ is a vector of coefficients parameterizing the filter. We follow the work of Defferrard et al [9] and restrict the class of filters to polynomial filters $\mathbf{g}_\phi = \sum_{k=0}^K \phi_k \mathbf{\Lambda}^k$. Polynomial filters are strictly localized in the vertex domain (a K -order polynomial filter considers K -hop neighborhoods around the node) and reduce the computational complexity of the convolutional operator. Such filters can be well approximated by a truncated expansion in terms of Chebyshev polynomials computed recursively. Following [9, 26] we adopt this approximation to implement the spectral convolutions. Note that a spectral convolutional layer will take feature matrices $\mathbf{X}^l$ as input, and produce filtered versions $\mathbf{X}^{l+1}$ akin to what standard convolutions do with images and feature maps.

Auto-encoding shape and appearance. Autoencoders are neural networks designed to reconstruct their input. They follow an encoder-decoder scheme, where an encoder $z = f_{enc}(x)$ maps the input x to a lower dimensional latent code z , which is then processed by a decoder $f_{dec}(z)$ to reconstruct the original input. The bottleneck imposed by the low-dimensionality of the encoding z forces the model to retain useful information, learning powerful representations of the data distribution. The model is trained to minimize a reconstruction loss $\mathcal{L}_{rec}(x, f_{dec}(f_{enc}(x)))$ between the input and the output reconstruction. To constrain the distribution of the latent space z , we add a variational loss term to the objective function, resulting in a variational autoencoder (VAE) [18]. We assume that the latent codes z are sampled from a distribution $Q(z)$ for which we will impose a unit multivariate Gaussian prior. In practise, during training, this results in the latent codes z being sampled from a distribution

$\mathcal{N}(\mu, \sigma)$ via the reparametrization trick [18], where μ, σ are deterministic parameters generated by the encoder $f_{enc}(x)$. Given a sample z , we can generate (reconstruct) the corresponding data point by using the decoder $f_{dec}(z)$. This model is trained by minimizing a loss function defined as:

$$\mathcal{L}_{vae} = \mathcal{L}_{rec}(x, f_{dec}(z)) + w\text{KL}(\mathcal{N}(0, 1) || Q(z|x)), \quad (1)$$

where the first term is the reconstruction loss, and the second term imposes a unit Gaussian prior $\mathcal{N}(0, 1)$ via the KL divergence loss. Depending on the type of data x that will be processed, here we will implement f_{enc} and f_{dec} using standard convolutional layers (to encode image appearance) or spectral convolutions (to encode graph structures).

Hybrid graph convolutional neural networks (HybridGNet). To construct the HybridGNet, we pre-train two independent VAEs. The first one, defined by the corresponding encoder f_{enc}^I and decoder f_{dec}^I , is trained to reconstruct images I . The second one, defined by f_{enc}^G and f_{dec}^G , is trained using graphs. In both cases, the bottleneck latent representation is modelled by fully connected neurons and have the same size. The encoder/decoder blocks are implemented using standard convolutional layers in the first case, and spectral convolutions in the later case. As depicted in Figure 1, once both models are trained, we decouple the encoders and decoders, keeping only the image encoder f_{enc}^I and graph decoder f_{dec}^G . The HybridGNet is thus constructed by connecting these two networks as $f_{HybridGNet}(x) = f_{dec}^G(f_{enc}^I(x))$. The coupled model is initialized with the pre-trained weights, and re-trained until convergence using the landmarks MSE as reconstruction loss, and the KL divergence term for regularization. Thus, HybridGNet takes images I as input and produce graphs $\mathcal{G}$ as output, combining standard with spectral convolutions in a single model.

Dual models. Based on the HybridGNet model, we implemented two more architectures incorporating a second decoder which outputs dense segmentation masks. The first one is the so called HybridGNet *Dual* model, which has two decoders: the graph f_{dec}^G and an extra f_{dec}^I whose output is a dense probabilistic segmentation. The model is trained to minimize a reconstruction loss $\mathcal{L}_{dual} = \mathcal{L}_{rec} + \mathcal{L}_{seg}$, combining the landmark localization error ($\mathcal{L}_{rec}$) with the dense segmentation error ($\mathcal{L}_{seg}$). The second model, entitled HybridGNet *Dual SC* incorporates skip connections (implemented as sums) between corresponding blocks from the image encoder f_{enc}^I and decoder f_{dec}^I . Even though in this paper we do not focus on dense segmentation models, these models were implemented to evaluate the effect of incorporating dense masks into the learning process. Since the dataset only contains contours, we derive the corresponding dense masks for training by filling the organ contours, assigning different labels to every organ.

Baseline models. Our work builds on the hypothesis that encoding connectivity information through graph structures will result in richer representations than standard landmark-based point distribution models. To assess the validity of this hypothesis, we construct standard point distribution models (like those used in ASM) from the graph representations. Differently from the graph structures which incorporate connectivity information, here the landmarks are treated as independent points. For a given graph $\mathcal{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$, we construct a vectorized representation S by concatenating the rows of $\mathbf{X}$ in a single vector as $S = (\mathbf{X}_{0,0}, \mathbf{X}_{0,1}, \mathbf{X}_{1,0}, \mathbf{X}_{1,1}, \dots, \mathbf{X}_{|V|-1,0}, \mathbf{X}_{|V|-1,1})$.

We implement two baseline models using the vectorized representation. In the first one, similar to [23, 3], we simply use principal component analysis (PCA) to turn the vectorized representations S into lower-dimensional embeddings. We then optimize the pre-trained image encoder f_{enc} to produce the coefficients of the PCA modes, reconstructing the final landmark-based segmentation as a linear combination of the modes. In the second model, we train a fully connected VAE, f^S , to reconstruct the vectorized representations S . Following the same methodology used to construct the HybridGNet, we keep only the decoder f_{dec}^S and connect it with the pre-trained image encoder f_{enc}^I . These baselines are also trained end-to-end computing the loss on the final location of the reconstructed landmarks, and using the pre-trained weights for initialization.

We also include a third baseline model for comparison, which uses a multi-atlas segmentation (MAS) approach [2, 1]. Given a test image for which we want to predict the landmarks, we take the 5 most similar images (in terms of mutual information) from the training set. We then perform pairwise non-rigid registration (with affine initialization) using Simple Elastix [21]. The final landmark-based segmentation is obtained by simply averaging the position of the corresponding transferred landmarks to the target space.

Detailed architectures and training details. The VAE image encoder f_{enc}^I consists of 5 residual blocks [14], alternated with 4 max pooling operations. The decoder f_{dec}^I is a mirrored version of f_{enc}^I . The graph autoencoder consists on 4 layers of Chebyshev convolutions for the encoder f_{enc}^G and 5 layers for the decoder f_{dec}^G , with 16 feature maps and a polynomial order of 6. In both models the latent space was fixed to 64 features and implemented as a fully connected layer, see Tables 1-4 in Sup. Mat. for more details. The autoencoders were trained using MSE as $\mathcal{L}_{rec}$ (for both the image and landmark reconstruction). The HybridGNet was initialized with the pre-trained weights and refined for 2000 epochs using the MSE to compute the landmark reconstruction error. The dual models also incorporated the $\mathcal{L}_{seg}$ term based on the average between soft Dice [22] and Cross Entropy. All models used the KL divergence for regularization (with $w = 1e - 5$, chosen by grid search) and were trained using Adam optimizer. All the models were implemented in PyTorch 1.7.1 and PyTorch Geometric 1.6.3. Our code is available at <https://github.com/ngaggion/HybridGNet>.

3 Experiments and discussion

We evaluated the proposed approaches using the Japanese Society of Radiological Technology (JSRT) Database [29], which consists on 247 high resolution X-Ray images, with expert annotations for lung, hearth and clavicles [13]. Annotations consist on 166 landmarks for every image. We constructed the graph adjacency matrix following a connectivity pattern as shown in Figure 2. This matrix was shared across all images. The dataset was divided in 3 folds, using 70% of the images for training, 10% for validation and 20% for testing. We showcase the potential of the proposed HybridGNet architecture in three different experiments.

Experiment 1: Anatomical landmark-based segmentation. We evaluated the proposed models using landmark specific metrics, including the mean square error (MSE, in pixel space) over the vectorized landmark location and the contour Hausdorff distance (HD, in millimeters). Figure 2 shows that the Hybrid models outperform the baselines in terms of both MSE and HD. However, incorporating the dual path with dense mask reconstruction does not seem to make a difference.

Experiment 2: Generating landmark-based representations from dense segmentations. The proposed HybridGNet offers a natural way to recover landmark based representations from dense segmentation masks. This could be useful in scenarios where dense segmentation masks are available, but we require matched landmarks for statistical shape analysis. We trained our HybridGNet using dense segmentation masks as input and compared the resulting model with the other approaches trained in the same conditions. Figure 3.d shows that proposed HybridGNet outperformed the three baselines, confirming that it can be used to construct statistical shape models with landmark correspondences from pixel level annotations (see Sup. Mat. for visual results).

Experiment 3: Robustness study. We assessed the robustness of the proposed model to image occlusions which may be due to anonymization reasons. We created artificial occlusions by overlapping random black boxes of different size on the test images. In order to highlight the advantages of the HybridGNet over standard dense segmentation models, we trained a UNet architecture (same as the HybridGNet Dual SC but without the graph decoder). Figure 3 shows the results when comparing the dense UNet with our HybridGNet variants. Since we cannot compute the landmark MSE for the dense segmentations, we compared the models evaluating the Dice coefficient and Hausdorff distance

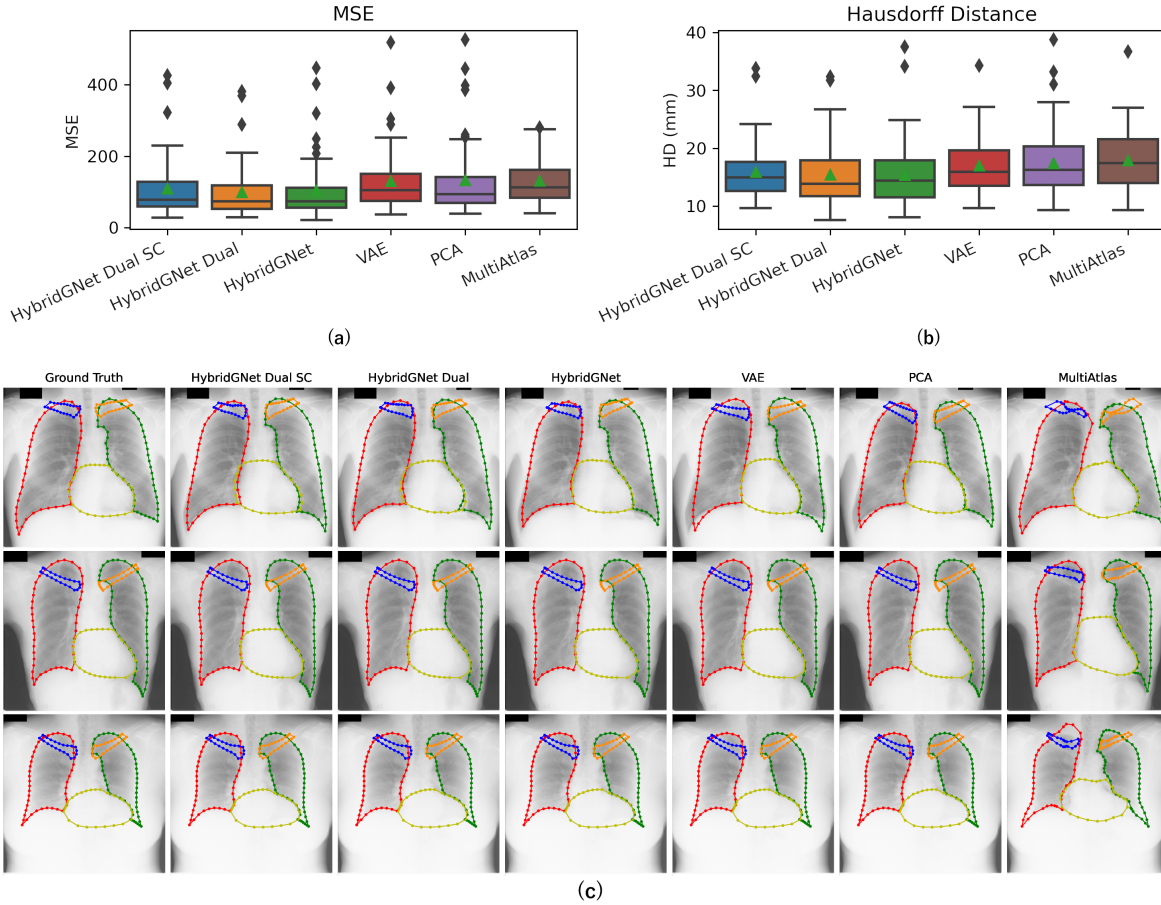


Figure 2: Quantitative and qualitative results. The difference between the means of the HybridGNet variants and the baselines for MSE (a) and HD (b) are statistically significant according to a Wilcoxon paired test (see the Sup. Mat. for details). We also include qualitative results (c) reflecting the improvement in anatomical plausibility obtained when using the HybridGNets (particularly in the clavicles which are the most challenging).

on the dense masks obtained from the UNet and the convolutional decoder of the dual models. We can see that not only the UNet, but also the dual models which incorporate dense segmentation masks during training, degrade the performance faster than the HybridGNet as we increase the size of the occlusion block.

4 Conclusions

In this work we proposed the HybridGNet architecture to perform landmark-based anatomical segmentation. We show that incorporating connectivity information through the graph adjacency matrix helps to improve the accuracy of the results when compared with other landmark-based models which only employ vectorized landmark representations. We showcased different application scenarios for the HybridGNet and confirm that it is robust to image occlusions, in contrast to standard dense segmentation methods which tend to fail in this task.

Acknowledgments. This work was supported by grants from ANPCyT (PICT 2018-3907 and

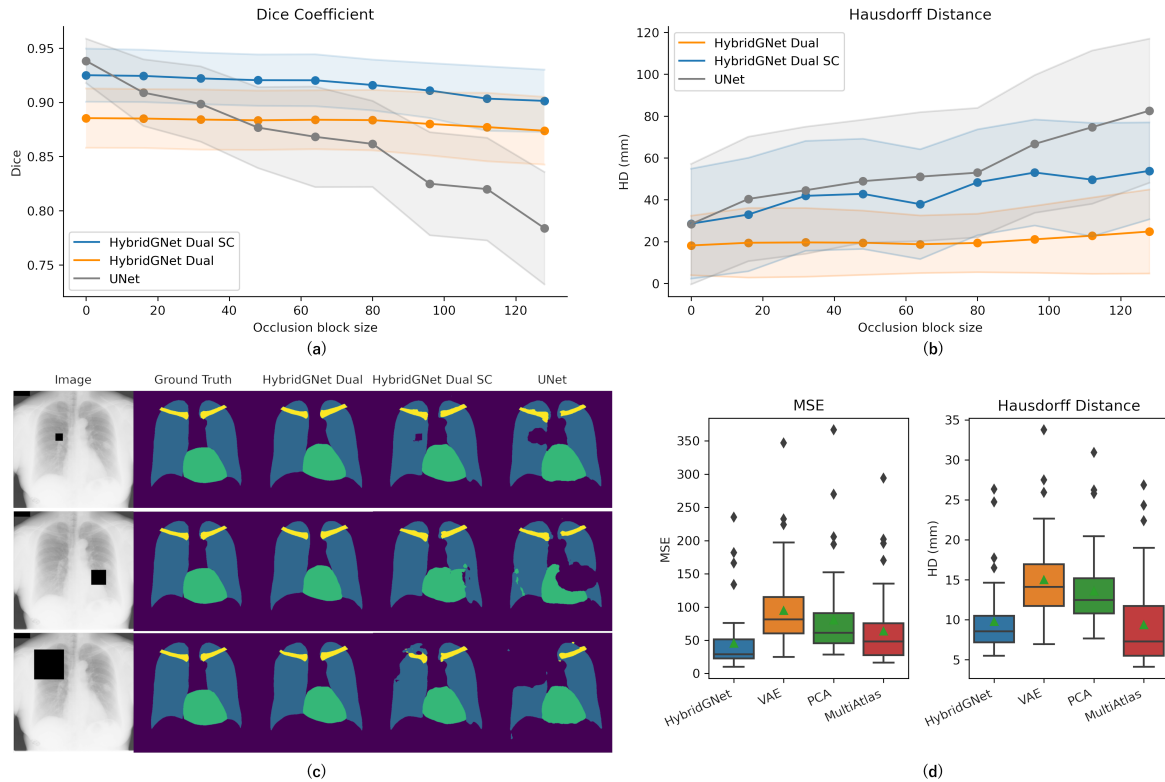


Figure 3: Figures (a), (b) and (c) show quantitative and qualitative results for the Experiment 3 (occlusion study). We can see that the HybridGNet Dual models are much more robust to missing parts than a standard UNet model. Figure (d) includes quantitative results for Experiment 2, showing that HybridGNet is more accurate at recovering landmark contours from dense segmentation masks.

3384), UNL (CAI+D 50220140100-084LI, 50620190100-145LI and 115LI) and The Royal Society (IES/R2/202165). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp used for this research.

References

- [1] Alvé, J., Kahl, F., Landgren, M., Larsson, V., Ulén, J.: Shape-aware multi-atlas segmentation. In: 2016 23rd International Conference on Pattern Recognition (Icpr). pp. 1101–1106. IEEE (2016)
- [2] Alvé, J., Kahl, F., Landgren, M., Larsson, V., Ulén, J., Enqvist, O.: Shape-aware label fusion for multi-atlas frameworks. *Pattern Recognition Letters* **124**, 109–117 (2019)
- [3] Bhalodia, R., Elhabian, S.Y., Kavan, L., Whitaker, R.T.: Deepssm: a deep learning framework for statistical shape modeling from raw images. In: International Workshop on Shape in Medical Imaging. pp. 244–257. Springer (2018)
- [4] Boussaid, H., Kokkinos, I., Paragios, N.: Discriminative learning of deformable contour models. In: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI). pp. 624–628. IEEE (2014)
- [5] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P.: Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* **34**(4), 18–42 (2017)

- [6] Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203 (2013)
- [7] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: European conference on computer vision. pp. 484–498. Springer (1998)
- [8] Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Training models of shape from sets of examples. In: BMVC92, pp. 9–18. Springer (1992)
- [9] Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. arXiv preprint arXiv:1606.09375 (2016)
- [10] Foti, S., Koo, B., Dowrick, T., Ramalhinho, J., Allam, M., Davidson, B., Stoyanov, D., Clarkson, M.J.: Intraoperative liver surface completion with graph convolutional vae. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis, pp. 198–207. Springer (2020)
- [11] Frangi, A.F., Niessen, W.J., Rueckert, D., Schnabel, J.A.: Automatic 3d asm construction via atlas-based landmarking and volumetric elastic registration. In: Biennial International Conference on Information Processing in Medical Imaging. pp. 78–91. Springer (2001)
- [12] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: International Conference on Machine Learning. pp. 1263–1272. PMLR (2017)
- [13] van Ginneken, B., Stegmann, M., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis* **10**(1), 19–40 (2006)
- [14] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
- [15] Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis* **13**(4), 543–563 (2009)
- [16] Heitz, G., Rohlfing, T., Maurer Jr, C.R.: Automatic generation of shape models using nonrigid registration with a single segmented template mesh. In: VMV. pp. 73–80 (2004)
- [17] Jurdia, R.E., Petitjean, C., Honeine, P., Cheplygina, V., Abdallah, F.: High-level prior-based loss functions for medical image segmentation: A survey. arXiv preprint arXiv:2011.08018 (2020)
- [18] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
- [19] Kipf, T.N., Welling, M.: Variational graph auto-encoders. arXiv preprint arXiv:1611.07308 (2016)
- [20] Larrazabal, A.J., Martinez, C., Ferrante, E.: Anatomical priors for image segmentation via post-processing with denoising autoencoders. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 585–593. Springer (2019)
- [21] Marstal, K., Berendsen, F., Staring, M., Klein, S.: Simpleelastix: A user-friendly, multi-lingual library for medical image registration. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 134–142 (2016)
- [22] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)

- [23] Milletari, F., Rothberg, A., Jia, J., Sofka, M.: Integrating statistical prior knowledge into convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 161–168. Springer (2017)
- [24] Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O’Regan, D.P., et al.: Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging* **37**(2), 384–395 (2017)
- [25] Paulsen, R., Larsen, R., Nielsen, C., Laugesen, S., Ersbøll, B.: Building and testing a statistical shape model of the human ear canal. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 373–380. Springer (2002)
- [26] Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018)
- [27] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
- [28] Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., Kokkinos, I.: Subcortical brain structure segmentation using f-cnn’s. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). pp. 269–272. IEEE (2016)
- [29] Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.i., Matsui, M., Fujita, H., Kodera, Y., Doi, K.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules. *American Journal of Roentgenology* **174**(1), 71–74 (2000)
- [30] Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine* **30**(3), 83–98 (2013)
- [31] Sozou, P.D., Cootes, T.F., Taylor, C.J., Di Mauro, E., Lanitis, A.: Non-linear point distribution modelling using a multi-layer perceptron. *Image and Vision Computing* **15**(6), 457–463 (1997)
- [32] Zitova, B., Flusser, J.: Image registration methods: a survey. *Image and vision computing* **21**(11), 977–1000 (2003)

1 Supplementary material for: Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

Convolutional Encoder	Feature Maps		Image Size
	In	Out	
Residual Block	1	8	512x512
Max Pooling			256x256
Residual Block	8	16	256x256
Max Pooling			128x128
Residual Block	16	32	128x128
Max Pooling			64x64
Residual Block	32	64	64x64
Max Pooling			32x32
Residual Block	64	64	32x32
Flatten			
Fully Connected (Mu)	65536	64	-
Fully Connected (Sigma)	65536	64	-

Table 1: Convolutional encoder f_{enc}^I detailed architecture.

Convolutional Decoder	Feature Maps		Image Size
	In	Out	
Fully Connected	64	65536	-
Unflatten	65536	64	32x32
Upsampling			64x64
Residual Block	64	64	64x64
Upsampling			128x128
Residual Block	64	32	128x128
Upsampling			256x256
Residual Block	32	16	256x256
Upsampling			512x512
Residual Block	16	8	512x512
Conv2D	8	4	512x512

Table 2: Convolutional decoder f_{dec}^I detailed architecture.

Encoder HybridGNet	Nodes x N ^o Kernels		Polynomial Order
	Input	Output	
Chebyshev Convolution	166x2	166x16	6
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x16	6
Fully Connected	166x16	64	-

Table 3: Graph encoder f_{enc}^g detailed architecture.

Decoder HybridGNet	Nodes x N ^o Kernels		Polynomial Order
	Input	Output	
Fully Connected	64	166x16	-
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x16	6
Chebyshev Convolution	166x16	166x2	6

Table 4: Graph decoder f_{dec}^g detailed architecture.

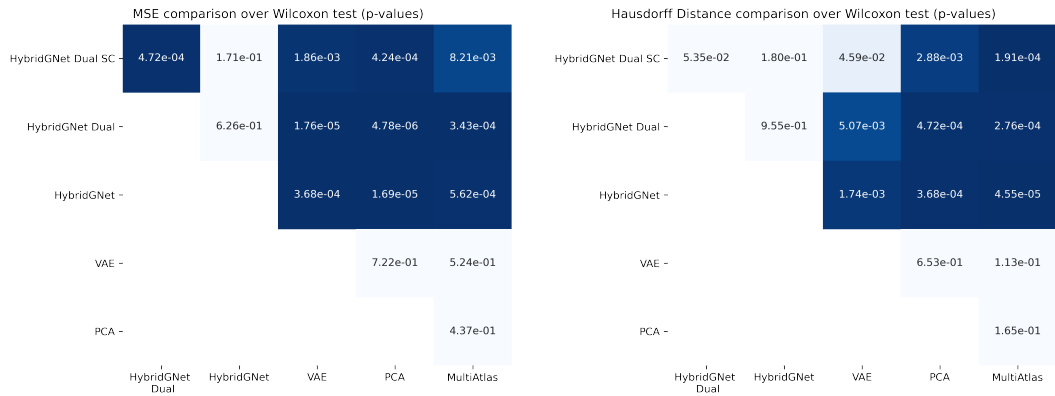


Figure 1: Statistical significance for model comparison in Figure 2.a and 2.b of the main manuscript, considering both MSE and Hausdorff Distance via Wilcoxon paired test. P-values < 0.05 implies significative differences between the means.

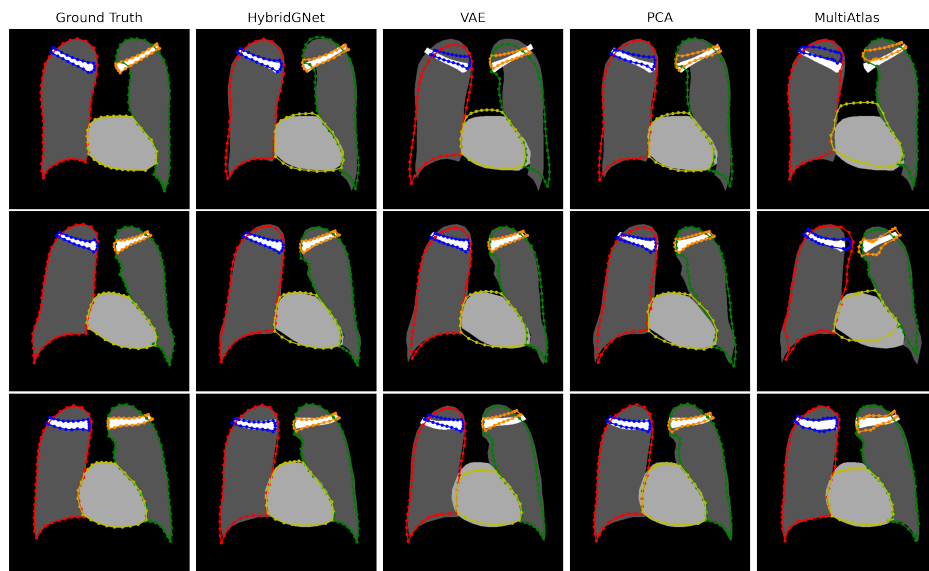


Figure 2: Examples of landmark-based representations generated from dense segmentation masks in the Experiment 2.

Improving anatomical plausibility in
medical image segmentation via
hybrid graph neural networks:
applications to chest x-ray analysis

Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis

Nicolás Gaggion¹, Lucas Mansilla¹, Candelaria Mosquera^{2,3}, Diego H. Milone¹, and Enzo Ferrante¹

¹Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET-UNL, Santa Fe, Argentina

²Health Informatics Department, Hospital Italiano de Buenos Aires

³Universidad Tecnológica Nacional, Buenos Aires, Argentina

Abstract

Anatomical segmentation is a fundamental task in medical image computing, generally tackled with fully convolutional neural networks which produce dense segmentation masks. These models are often trained with loss functions such as cross-entropy or Dice, which assume pixels to be independent of each other, thus ignoring topological errors and anatomical inconsistencies. We address this limitation by moving from pixel-level to graph representations, which allow to naturally incorporate anatomical constraints by construction. To this end, we introduce HybridGNet, an encoder-decoder neural architecture that leverages standard convolutions for image feature encoding and graph convolutional neural networks (GCNNs) to decode plausible representations of anatomical structures. We also propose a novel image-to-graph skip connection layer which allows localized features to flow from standard convolutional blocks to GCNN blocks, and show that it improves segmentation accuracy. The proposed architecture is extensively evaluated in a variety of domain shift and image occlusion scenarios, and audited considering different types of demographic domain shift. Our comprehensive experimental setup compares HybridGNet with other landmark and pixel-based models for anatomical segmentation in chest x-ray images, and shows that it produces anatomically plausible results in challenging scenarios where other models tend to fail.

1 Introduction

Deep convolutional neural networks (CNNs) have achieved outstanding performance in anatomical segmentation of biomedical images. Classical approaches employ standard encoder-decoder CNN architectures [1] that predict the desired segmentation at pixel-level by learning hierarchical features from annotated datasets. Casting image segmentation as a pixel labeling problem is desirable in scenarios where topology and location do not tend to be preserved across individuals, like lesion segmentation. However, organs and anatomical structures usually present a characteristic topology that tends to be regular. Since deep segmentation networks are typically trained to minimize pixel-level loss functions, such as cross-entropy or soft Dice [2], their predictions are not guaranteed to reflect anatomical plausibility, due to the inherent lack of sensitivity that these metrics have with respect to global shape and topology [3] (i.e. many different shapes can lead to the same score). Artifacts such as fragmented structures, topological inconsistencies and islands of pixels [4] are common for such models, especially when faced with challenging real-world clinical scenarios like image occlusions and inter-center domain

shift. Incorporating prior knowledge and shape constraints [5] to avoid these artifacts becomes fundamentally important when considering the downstream tasks where segmentation masks are used, like disease diagnosis, therapy planning and patient follow-up.

As an alternative to dense pixel-level masks, anatomical segmentation can be tackled using other approaches like statistical shape models [6] or graph-based representations [7], which provide a natural way to incorporate topological constraints by construction. Such representations make it easier to establish landmark correspondences among individuals, especially important in the context of statistical shape analysis. In particular, graphs appear as a natural way to represent landmarks, contours, and surfaces. By defining the landmark position as a function on the graph nodes, and encoding the anatomical structure through its adjacency matrix, we can easily constrain the space of solutions and encourage topological correctness. With the emergence of geometric deep learning [8], CNN extensions to non-euclidean domains like spectral graph convolutions [9, 10] and neural message passing [11] enabled the construction of deep learning models on graphs. This allowed for the creation of discriminative models that can make predictions based on graph data, as well as deep generative models [12, 13], which can be used to produce realistic graph structures under a certain distribution.

Contributions

In this work, we explore how landmark-based segmentation can be modeled by combining standard convolutions to encode image features, with generative models based on graph neural networks (GCNNs) to decode anatomically plausible representations of segmented structures. We introduce the HybridGNet architecture, which takes images as input, process them with standard convolutions and then generates landmark-based segmentations by sampling the bottleneck latent distribution, re-shaping and convolving in the graphs domain. We also present the “image-to-graph skip connections” (IGSC) module, which follows the same spirit of UNet skip connections, where feature maps at equivalent resolutions flow from encoder to decoder by-passing the model bottleneck. We propose to extract feature map patches from the image encoder path, and concatenate them with the node features in the GCNN graph decoder to improve accuracy by recovering details from localized high-resolution representations.

A preliminary version of this work was presented at MICCAI 2021 [14]. In this journal extension we include: 1) The novel IGSC module, which combined with graph unpooling operations, allows localized features at equivalent image/graph resolutions to flow from standard convolutional blocks to GCNN blocks. This is accompanied with new experiments which demonstrate the improvements produced by IGSC when compared to the vanilla HybridGNet and other baselines; 2) A new domain shift study based on two additional datasets, namely the Montgomery and Shenzhen datasets, showing the robustness of HybridGNet to multi-centric domain shift; 3) A new publicly available dataset of landmark annotations with node correspondences generated with HybridGNet for the databases which did not include this type of annotation originally; 4) A new robustness study which includes real X-ray image occlusion from the Padchest dataset caused by pacemakers; 5) A new experimental study to assess the impact of demographic domain shift (in particular, the train/test age distribution) in model performance; 6) Additional experiments to assess the behaviour of the model for pathological anatomy; and 7) A new clinical use-case study where we show how HybridGNet can be used to compute clinically meaningful indices like the cardiothoracic ratio.

2 Related work

Landmark-based segmentation

Since the early 1990’s, variations of point distribution models (PDMs) have been proposed [15] to segment anatomical structures using landmarks. PDMs are flexible shape templates describing how the relative location of important points can vary. Techniques based on PDMs, like active shape models (ASM) [15, 16] and active appearance models (AAM) [17] became the defacto standard to deal with

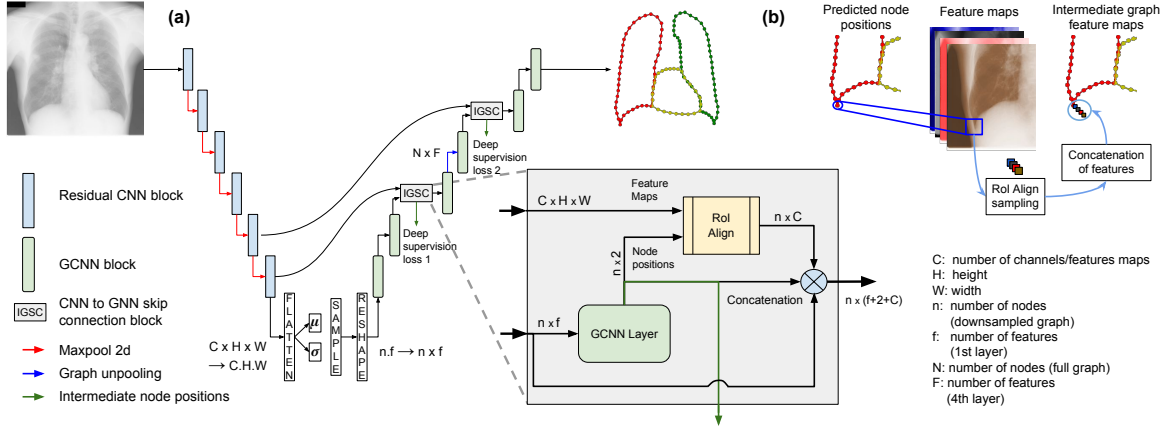


Figure 1: **HybridGNet architecture.** (a) The proposed HybridGNet architecture combines standard convolutions for image feature encoding (cyan) with graph spectral convolutions (green) to decode plausible anatomical graph-based representations. After the input image is processed by the image encoder, we sample a 1 dimensional vector from the VAE latent distribution (with $n.f$ components) which is then reshaped into a 2D matrix of size $n \times f$ representing the initial node features. Thus, the 1D sampled latent code must have $n.f$ components, so that it can be reshaped into a 2D node features table of dimensions $n \times f$. The Image-to-Graph skip-connection (IGSC) module provides localized features to the intermediate graph representations. (b) Illustrative visualization of the RoIAlign sampling and concatenation of features inside the IGSC module.

anatomical segmentation at the end of the century. Subsequently, the development of more powerful and robust image registration algorithms [18] positioned deformable template matching algorithms as the choice of option for anatomical segmentation and atlas construction [19–21]. In this scenario, contours (for 2D images) and meshes (for 3D images) have been used as deformable templates to solve landmark-based segmentation. However, these methods do not leverage the power of deep neural networks which have dominated image segmentation during the last decade.

More recently, with the advent of deep fully convolutional networks [1, 22], major efforts were made to incorporate anatomical constraints into such models [23–25]. The richness and robustness of the hierarchical features learned by CNNs allowed them to achieve highly accurate results. Unfortunately, most of these methods work directly on the pixel space, producing acceptable dense segmentation masks, but without landmark annotations and connectivity structure. On the contrary, structured models like graphs can easily represent landmarks, contours and surfaces. In line with this idea, recent studies [26–28] have integrated standard CNNs with different representations of landmark structures. These methods employ low-dimensional shape representations like Principal Component Analysis (PCA) decomposition of the original shape space [26, 27] or performed on more sophisticated particle distribution models [28]. In this work, inspired by previous studies on graph generative models [13], we propose to replace such embeddings by more powerful non-linear representations based on hierarchical graph convolutional [8] decoders.

Graph generative models

We want to exploit the generative power of graph variational autoencoders [29] to decode plausible anatomical segmentations from low dimensional embeddings. Of particular interest for our work is the convolutional mesh autoencoder proposed in [13]. The authors constructed an encoder-decoder network using spectral graph convolutions, and trained it in a variational setting using face meshes. By sampling the latent space, they are able to generate new expressive faces, never seen during training.

We build on top of this idea by keeping the graph convolutional *decoder*, but replacing the graph *encoder* with a standard CNN-based encoder that takes images as inputs. This hybrid architecture learns a variational distribution conditioned on image data, from which we can sample graphs representing anatomically plausible segmentations.

Image-to-graph localized skip connections

Last but not least, we are interested in producing accurate landmark-based segmentation for high-resolution 2D images. In that sense, propagating features learned at different hierarchical levels from encoder to decoder through skip connections has shown to be an effective mechanism not only to improve segmentation accuracy, but also to increase convergence speed and enable training of very deep networks [30]. Previous approaches incorporated different types of skip-connections in the context of mesh extraction from images. Pixel2Mesh [31] introduces a perceptual feature pooling layer designed to work with 3D meshes and 2D images, thus projecting 3D vertices to the image plane using camera intrinsics, which does not apply for our case where input image and output graph live in the same 2D space. Closer to our approach is Voxel2Mesh [32], a model designed to operate on images and graphs living in the same dimension. Voxel2Mesh employs a learned neighborhood sampling layer which pools image features in locations indicated by the node coordinates. However, both Voxel2Mesh and Pixel2Mesh build on the idea of deforming an initial sphere mesh template, thus limiting its applicability to certain topologies and single object segmentation. Another approach based on template refinement is Curve-GCN [33], an algorithm that can be used to perform automatic or interactive 2D landmark-based segmentation, which relies on the deformation of an initial ellipsoid template contour. Curve-GCN operates based on an iterative approach where the contour node displacements are successively inferred and used to deform an initial template. In this case, only localized image features are sampled from the corresponding node positions. On the contrary, our model employs a generative approach which samples a global embedding from a variational latent distribution, which is then used to directly decode the node positions in a single forward pass. Other approaches resort to refining meshes obtained from voxel predictions [34]. Here we adopt a different approach where output graphs (2D contours in our case) do not correspond to a deformed template, but instead are directly sampled from a latent distribution learnt during training. We also propose a new image-to-graph skip connection (IGSC) layer based on the well-known RoIAlign module [35], which enables end-to-end learning of localized features guided by intermediate node coordinates.

3 Anatomical segmentation via hybrid graph neural networks

3.1 Preliminaries

Problem setting

Let us have a dataset $\mathcal{D} = \{(\mathbf{I}, \mathbf{G})_n\}_{0 < n \leq N}$, composed of N images $\mathbf{I}$ and their corresponding landmark-based segmentation represented as a graph $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$ (see Section 4.1.1 for a detailed description of the graph construction for the heart and lungs structures used in this study). V is the set of nodes for M landmarks, $\mathbf{A} \in \{0, 1\}^{M \times M}$ is the adjacency matrix indicating the connectivity between pairs of nodes ($a_{ij} = 1$ indicates an edge connecting vertices i and j , and $a_{ij} = 0$ otherwise) and $\mathbf{X} \in \mathbb{R}^{M \times s}$ is a function (represented as a matrix) assigning a feature vector to every node. In our case, it assigns a 2-dimensional spatial coordinate (the landmark position) to every node ($s = 2$). In the context of landmark-based segmentation and point distribution models, it is common (and useful) to have manual annotations with a fixed number of points. Therefore, we assume that V and $\mathbf{A}$ are the same for all the images in the dataset, the only difference among them is given by the spatial coordinates defined in $\mathbf{X}$. This assumption enables us to follow the work of [10, 13] and use spectral graph convolutions to learn latent representations of anatomy. Please note that, following the literature of landmark-based segmentation in medical imaging [14, 36], we use the term *landmark* to denote points that can be

uniquely identified in a set of shapes. However, these are also called *pseudo-landmarks* as not all of them are actual anatomical points, but they are landmarks lying on the contour of a shape, determining its geometry [19].

Spectral graph convolutions

Spectral convolutions are built using the eigendecomposition of the graph Laplacian matrix $\mathbf{L}$, exploiting the property that convolutions in the node domain are equivalent to multiplications in the graph spectral domain [37]. The graph Laplacian is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the diagonal degree matrix with $d_{ii} = \sum_j a_{ij}$. The Laplacian can be decomposed as $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{M \times M} = [\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{M-1}]$ is the matrix of eigenvectors (Fourier basis) and $\mathbf{\Lambda} = \text{diag}(\lambda_0, \lambda_1, \dots, \lambda_{M-1})$ is the matrix of eigenvalues (frequencies of the graph). By analogy with the classical Fourier transform for continuous or discrete signals, the graph Fourier transform of a function $\mathbf{X}$ defined on the graph domain can be obtained as $\hat{\mathbf{X}} = \mathbf{U}^T \mathbf{X}$, while its inverse is given by $\mathbf{X} = \mathbf{U} \hat{\mathbf{X}}$. Based on this formulation, the spectral convolution between a signal $\mathbf{X}$ and a filter $\mathbf{g}_\phi = \text{diag}(\phi)$ is defined as $\mathbf{g}_\phi * \mathbf{X} = \mathbf{g}_\phi \mathbf{L} \mathbf{X} = \mathbf{g}_\phi (\mathbf{U} \mathbf{\Lambda} \mathbf{U}^T) \mathbf{X} = \mathbf{U} \mathbf{g}_\phi \mathbf{\Lambda} \mathbf{U}^T \mathbf{X}$, where $\phi \in \mathbb{R}^n$ is a vector of coefficients parameterizing the filter. We follow the work of Defferrard et al [10] and restrict the class of filters to polynomial filters with the form $\mathbf{g}_\phi = \sum_{k=0}^K \phi_k \mathbf{\Lambda}^k$. Polynomial filters are strictly localized in the vertex domain (a K-order polynomial filter considers K-hop neighborhoods around the node) and reduce the computational complexity of the convolutional operator. Such filters can be well approximated by a truncated expansion in terms of Chebyshev polynomials, computed recursively. Following [10, 13] we adopt this approximation to implement the spectral convolutions. Note that a spectral convolutional layer will take feature matrices $\mathbf{X}^\ell$ as input and produce filtered versions $\mathbf{X}^{\ell+1}$, similar to what standard convolutions do with images and feature maps.

3.2 HybridGNet: Image-to-graph extraction via hybrid convolutions

The proposed neural network takes images as input and produces graphs as output, combining standard with spectral convolutions in a single model that is trained end-to-end. The current HybridGNet formulation follows the same principles introduced in our original MICCAI publication [14], but incorporates new elements like image-to-graph skip connections, graph unpooling operations and variations in the training strategy, that will be later highlighted. Let us start by defining the basic architecture, which resembles a variational autoencoder (VAE) [38] (see Figure 1) in the sense that the latent space models a variational distribution parameterized as a multivariate Gaussian.

Autoencoders are neural networks designed to reconstruct their input. They follow an encoder-decoder scheme, where an encoder $\mathbf{z} = f_e(\mathbf{I})$ maps the input image $\mathbf{I}$ to a lower dimensional latent code $\mathbf{z}$, which is then processed by a decoder $f_d(\mathbf{z})$ to reconstruct the original input. The bottleneck imposed by the low-dimensionality of the encoding $\mathbf{z}$ forces the model to retain useful information, learning powerful representations of the data distribution. The model is trained to minimize a reconstruction loss $\mathcal{L}_r(\mathbf{I}, f_d(f_e(\mathbf{I})))$ between the input and the output reconstruction. To constrain the distribution of the latent space $\mathbf{z}$, we add a variational term to the loss function, resulting in a variational autoencoder (VAE) [38]. We assume that the latent codes $\mathbf{z}$ are sampled from a distribution $Q(\mathbf{z})$ for which we will impose a unit multivariate Gaussian prior. In practise, during training, this results in the latent codes $\mathbf{z}$ being sampled from a distribution $\mathcal{N}(\mu, \sigma)$ via the reparametrization trick [38], where μ, σ are deterministic parameters generated by the encoder $f_e(\mathbf{I})$. Given a sample $\mathbf{z}$, we can generate (reconstruct) the corresponding data point by using the decoder $f_d(\mathbf{I})$. This model is usually trained by minimizing a loss function defined as:

$$\mathcal{L}_a = \mathcal{L}_r(\mathbf{I}, f_d(\mathbf{z})) + w \text{KL}(\mathcal{N}(0, 1) || Q(\mathbf{z}|\mathbf{I})), \quad (1)$$

where the first term is the reconstruction loss, the second term imposes a unit Gaussian prior $\mathcal{N}(0, 1)$ via the KL divergence loss and w is a weighting factor.

In our previous work [14], HybridGNet was constructed by first pre-training two independent VAEs with the same latent dimension: one to reconstruct images using standard convolutions $f^I(\mathbf{I}) = f_d^I(f_e^I(\mathbf{I}))$ and another one to reconstruct graphs via spectral convolutions $f^G(\mathbf{G}) = f_d^G(f_e^G(\mathbf{G}))$. Once both models were trained, we decoupled their encoders and decoders, keeping only the image encoder $f_e^I(\mathbf{I})$ and graph decoder $f_d^G(\mathbf{z})$. The HybridGNet was then constructed by connecting these two pre-trained networks as $f^H(\mathbf{I}) = f_d^G(f_e^I(\mathbf{I}))$ and re-training until convergence by minimizing:

$$\mathcal{L}_H = \mathcal{L}_r(\mathbf{G}, f_d^G(\mathbf{z})) + w \text{KL}(\mathcal{N}(0, 1) || Q(\mathbf{z}|\mathbf{I})), \quad (2)$$

where $\mathcal{L}_r(\mathbf{G}, f_d^G(\mathbf{z}))$ is the graph-reconstruction loss computed as the mean squared error (MSE) of the predicted node positions, and Q is the variational distribution parameterized by $f_e^I(\mathbf{I})$.

Here we simplify the training strategy by eliminating the pre-training stage and directly training $f^H(\mathbf{I})$ from scratch, since we observed that pre-training only helps to achieve faster convergence, but does not produce significant improvements in terms of segmentation accuracy. This simplified end-to-end training process directly learns a single latent space relating images and graphs.

Graph unpooling

We included a fixed graph unpooling layer in the graph decoder $f_d^G(\mathbf{z})$, to learn representations at multiple resolutions [13]. We adopted a simple strategy where all graphs $\mathbf{G}$ in our dataset are pre-processed to produce lower resolution graphs $\mathbf{G}^k$ by reducing to half the number of nodes k times, replacing pairs of consecutive neighboring nodes with a single one, whose position is computed as their average. The unpooling layer is defined so that it reverses this operation by duplicating the number of nodes and interpolating the features between them. The unpooling layer was included after the 3rd GCNN layer of the decoder as shown in Figure 1.

Localized image-to-graph skip connections (IGSC) and deep supervision

Under the hypothesis that local image features may help to produce more accurate estimates of landmark positions, we designed a localized Image-to-Graph Skip Connection (IGSC) layer (see Figure 1.a,b). IGSC uses the well-known RoIAlign module [35] to sample localized features for each node from a specific encoder level given a certain location, which in our case is specified by intermediate node positions learned via deep supervision [39]. This layer is parameterized by a window size, indicating the area that will be sampled for every node. It receives a tensor of feature maps and a list of node positions which indicate the spatial location from where the feature map will be sampled, and returns the corresponding regions of interest (RoIs) of the given window size centered at the node positions. In our model, an internal GCNN layer learns intermediate node positions via deep-supervision, resulting in extra loss terms $\mathcal{L}_{DS}$ which compute the mean squared error between the ground truth node position (for both graph resolutions) and the intermediate predictions. The desired window input size was set to 3x3, while the output size was set to 1x1, so it only returns a single value per feature-map, which is calculated using average pooling. Then, this array of features is concatenated with the original node features and an augmented graph is obtained as shown in Figure 1.b).

4 Experimental setup

4.1 Database description

We evaluated the proposed model in a variety of tasks involving chest x-ray image segmentation. In what follows, we describe the databases used to perform these experiments.

4.1.1 JSRT Database

The Japanese Society of Radiological Technology (JSRT) Database [40] consists on 247 high resolution x-ray images, with expert landmark annotations (120 landmarks per image) for lung and heart [41].

Table 1: Landmark-based anatomical segmentation results for JSRT dataset. Mean (std). HD in millimeters.

Type	Model	MSE	Dice Lungs	HD Lungs	Dice Heart	HD Heart	
Landmark Methods	PCA	340.024 (243.549)	0.945 (0.014)	17.445 (9.669)	0.906 (0.037)	14.602 (5.400)	
	FC	332.197 (242.379)	0.945 (0.017)	17.535 (10.352)	0.910 (0.038)	15.020 (5.785)	
	MultiAtlas	492.262 (298.138)	0.944 (0.013)	20.317 (9.344)	0.886 (0.056)	16.780 (6.839)	
	HybridGNet (without IGSC)	294.621 (274.497)	0.952 (0.013)	15.642 (10.922)	0.913 (0.038)	13.658 (5.548)	
	1 IGSC	Layer 3	277.536 (298.725)	0.954 (0.014)	14.565 (11.441)	0.917 (0.037)	13.401 (5.376)
		Layer 4	288.597 (272.538)	0.956 (0.013)	16.054 (11.284)	0.916 (0.038)	14.153 (6.038)
		Layer 5	258.413 (245.724)	0.963 (0.010)	13.662 (11.107)	0.915 (0.039)	13.738 (5.181)
		Layer 6	250.123 (232.032)	0.960 (0.011)	14.378 (9.262)	0.924 (0.030)	12.339 (4.844)
	2 IGSC	Layers 4-3	263.973 (262.700)	0.963 (0.011)	14.942 (10.589)	0.921 (0.036)	13.198 (5.514)
		Layers 5-4	246.845 (230.235)	0.968 (0.009)	13.692 (10.984)	0.924 (0.040)	13.417 (6.144)
		Layers 6-5	200.748 (211.080)	0.974 (0.007)	12.089 (9.344)	0.933 (0.031)	11.613 (5.581)
Pixel-level Methods	UNet	—	0.981 (0.008)	21.839 (26.291)	0.942 (0.030)	25.176 (34.570)	
	UNet + Post-DAE	—	0.965 (0.010)	17.969 (14.457)	0.935 (0.029)	15.444 (14.283)	
	nnUNet	—	0.984 (0.005)	9.615 (7.874)	0.952 (0.023)	9.782 (5.006)	

The image resolution was 1024x1024 px, with a pixel spacing of 0.35x0.35 mm. The dataset was randomly split into 70%-10%-20% partitions for training, validation and test, respectively.

Given the mask contour for the structures of interest, several anatomical landmarks in the lung and heart were manually identified (e.g. the lung apex): 4 for the right lung, 5 for the left lung, and 4 for the heart [41]. The other intermediate points were interpolated across the mask contour providing a final set of 44, 50 and 26 points respectively. The number of interpolation points between anatomical landmarks was fixed for all subjects, resulting in a one-to-one correspondence of the distinctive points between subjects. The graphs described in Section 3.1 were then constructed by taking each landmark as a node, defining edges between neighboring nodes, and the (x, y) positions of each landmark were set as the node features. The adjacency matrix was then pre-computed to represent the connectivity structure of the aforementioned graph. By having the same number of landmarks with one-to-one correspondences, the adjacency matrix was the same for all graphs.

4.1.2 Montgomery County and Shenzhen Hospital x-ray sets

Two public chest x-ray datasets with dense lung segmentation masks were used as external test sets to evaluate inter-dataset DS. The Montgomery County dataset (138 images) [42] was acquired from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. The Shenzhen dataset (566 images) [43] was collected as part of the routine care at Shenzhen No.3 Hospital in Shenzhen, Guangdong province, China.

4.1.3 Padchest dataset

Consists of 160,868 chest x-ray images from 67,000 patients [44] including labels for 174 radiological findings, 19 diagnostic labels, and 104 anatomic locations. Although this dataset does not contain segmentation masks, a subset of 137 images with *cardiomegaly* diagnosis label were manually segmented by two radiologists who delineated the lungs and heart as dense masks, to evaluate our method in a real clinical task, namely cardiothoracic ratio estimation. From these images, 20 included pacemakers and 45 also included an *aortic elongation* label. The images with pacemakers were used to evaluate the robustness of the proposed model to occlusions produced by external artifacts.

4.2 Baselines models

Our work builds on the hypothesis that encoding connectivity information through graph structures can provide richer representations than standard landmark-based point distribution models. To evaluate this hypothesis, we build standard point distribution models from the graph representations by considering landmarks as independent points. For a given graph $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$, we

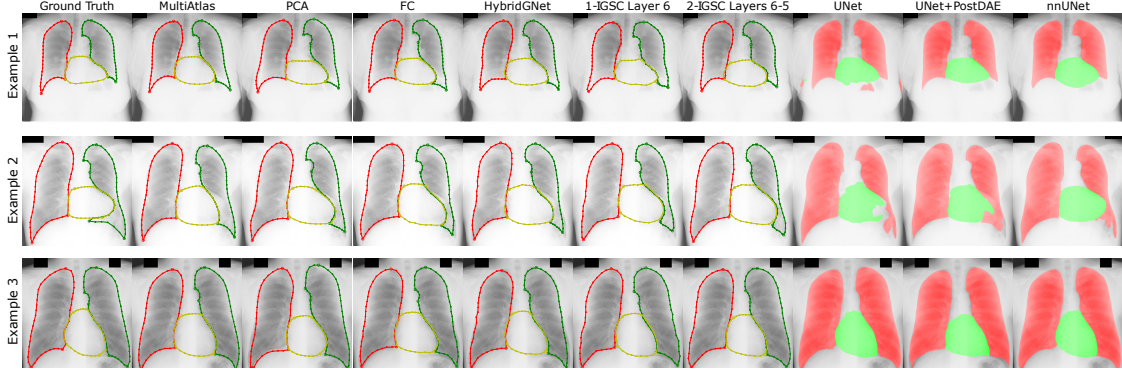


Figure 2: **Landmark-based anatomical segmentation.** Qualitative analysis for the JSRT test set. Results reflect the improvement in anatomically plausibility obtained when using the HybridGNet with IGSC.

construct a vectorized representation by concatenating the rows of $\mathbf{X}$ in a single vector as $\boldsymbol{\rho} = [x_{0,0}, x_{0,1}, x_{1,0}, x_{1,1}, \dots, x_{M-1,0}, x_{M-1,1}]$.

4.2.1 PCA

We first consider a single baseline similar to [26, 27], by performing principal component analysis (PCA) to transform the vectorized representation $\boldsymbol{\rho}$ into lower-dimensional embeddings. We then optimize the CNN encoder f_e^T to estimate the PCA coefficients, reconstructing the landmarks as a linear combination of the principal components.

4.2.2 FC

The second baseline combines the CNN encoder f_e^T with a fully connected (FC) decoder that directly reconstructs the vectored representations $\boldsymbol{\rho}$.

4.2.3 Multi-atlas

The third baseline implements a multi-atlas segmentation approach [45, 46], which employ several labeled atlases (i.e. pairs consisting of an image and its associated landmark-based segmentation) to delineate the structures of interest. Given a target image to be segmented, the 5 atlases most similar to the target image (based on the mutual information metric) are obtained from the training set. Then, we perform pairwise non-rigid registration (with affine initialization) using SimpleElastix [47]. Registration allows to transfer the landmarks of each selected image into the target space. The final landmark-based segmentation is obtained by averaging the position of the set of candidate landmarks.

4.2.4 UNet

Finally, a UNet [1] model was also included to benchmark our approach against a standard pixel-level segmentation method. We used the CNN encoder f_e^T and decoder f_d^T with standard skip-connections via concatenation, to guarantee comparable complexity.

4.2.5 Post-DAE Postprocessing:

UNet results were post-processed using a denoising autoencoder following [24], which was trained to obtain plausible segmentations from noisy dense segmentation masks.

4.2.6 nnUNet

a state-of the art nnUNet [48] was included, which consists of a self-configuring pipeline of automatic pre-processing, hyperparameter search, training, ensembling and post-processing following a 5-fold cross validation scheme. This 5-fold set is then treated as a ensemble model, averaging the predictions and incorporating standard post-processing techniques.

4.3 Implementation and training details

All models were implemented in PyTorch [49], using PyTorch Geometric [50] for the spectral GCNN layers¹. Every model and baseline shares the same CNN encoder f_e^I , with 6 residual blocks [51] interleaved with max-poolings as shown in Figure 1. For the GCNN decoders, we use 6 layers of Chebyshev convolutions with Layer Normalization [52] and ReLU nonlinearities. We set the k-hop neighbourhood parameter for the graph convolutions at 6. This hyperparameter was chosen performing an ablation study on the validation data, which is not included for space restrictions, but it is available in our repository. For HybridGNet models, we evaluated the inclusion of 1 and 2 IGSC modules, extracting features from layers 3 to 6 of the encoder. We also evaluated the incorporation of a third IGSC module (consequently adding another graph downsampling level), but it resulted in slightly worse performance (due to the low resolution of the downsampled graphs). Thus, we decided to stick to 2 IGSC modules.

4.3.1 Data augmentation

Online data augmentation was used to train all the models (i.e. baselines and HybridGNet) including: i) bright augmentation using a Gamma correction with random gamma between 0.60 and 1.40; ii) random image rotations between -3 and 3 degrees; iii) vertical and horizontal random scaling, ensuring that landmarks remain inside the visible area; iv) cropping or padding the images randomly if the shape was different to the expected input shape (1024×1024).

4.3.2 Model training

All models were trained for 3000 epochs using Adam optimizer, with a learning rate of $1e-4$, a batch size of 4, a weight decay of $1e-5$, and a KL divergence weight factor $w = 1e-5$. To prevent overfitting, learning rate decay was set to reduce it by 0.9 every 100 epochs (for IGSC models) and by 0.9 every 50 epochs (for HybridGNet model). We used the MSE in pixel space over the vectored landmark location as loss function for landmark models, and a combination of Dice and cross-entropy for the UNet. Checkpoints were selected based on validation loss.

5 Experiments and discussion

We performed a series of experiments to compare the proposed HybridGNet and its variants with the aforementioned baselines, and evaluate their performance in a variety of scenarios and tasks.

5.1 Model comparison

First, we compared HybridGNet with the baselines using the JSRT dataset, and assessed the effect of skip connections by evaluating alternative HybridGNet architectures. We used metrics that can be

¹Source code is publicly available at <https://github.com/ngaggion/HybridGNet>. The experiments are saved on Jupyter notebooks and extra information on statistical significance is also available in the repository. The Multi-atlas implementation is available at <https://github.com/lucasmansilla/multiatlas-landmark>.

Table 2: Inference times for each model per input image.

Model	Time (s)
HybridGNet 2-IGSC	0.53
UNet	0.60
UNet + Post-DAE	1.54
nnUNet	2.76

derived from graph representations, including landmark MSE and Hausdorff distance (HD, in millimeters). To benchmark our methods against pixel-level methods, we filled the organ contours to obtain dense masks from graph representations, and computed the Dice coefficient.

Table 1 reports metrics on the test set of JSRT dataset (bold numbers indicate the best performance for methods of the same type, i.e. landmark and pixel-level methods). All differences are significant according to the Wilcoxon test (except for the 1-IGSC Layer 6 model that has no significant difference in HD Heart with the 2-IGSC L6-5 model). First, it is worth noting that when comparing HybridGNet models with and without skip connections, there is a big difference in terms of MSE and HD in favor of the model with 2 IGSC (Layers 6-5), implying that localized features help to improve landmark prediction accuracy. Moreover, the HybridGNet 2 IGSC (Layers 6-5) outperforms the landmark-based baselines on MSE, Dice, and HD, confirming our main hypothesis that incorporating graph connectivity structure helps in producing more realistic segmentations.

For the sake of completeness, we also included pixel-level segmentation baselines. HybridGNet surpasses UNet and UNet+PostDAE baselines by a large margin in terms of HD and it is competitive with the nnUNet in that regard. On the contrary, the UNet model and the self-configuring nnUNet outperform the HybridGNet variants in Dice, what is somehow expected since dense predictions are not directly optimized in our models. In that vein, while Dice is agnostic to topological errors and islands of pixels (in the sense that wrong predictions are penalized independently of their location), due to its formulation HD is more sensitive to them, better reflecting anatomical plausibility, which is the main interest of this work (see [3] for a complete discussion about the limitations of image segmentation metrics). Figure 2 shows qualitative results for 3 exemplar cases and Table 2 shows inference times for HybridGNet and pixel-based models on an NVIDIA Titan Xp with 12GB RAM. The lower inference time of the HybridGNet is mainly due to the fact that the number of parameters in the graph decoder is much lower than that of the standard convolutional decoder.

As an important remark, although we include state-of-the-art pixel level segmentation methods like UNet, nnUNet and Post-DAE for completeness, we highlight the fact that our method is a landmark-based segmentation approach which produces structured predictions in the form of graphs (not pixel-level masks), enabling the extraction of additional information like anatomical landmark locations and inter-patient node correspondences, which cannot be obtained with standard pixel-level segmentation methods. Thus, for a more fair comparison of the HybridGNet results, the reader should focus on the landmark-based MSE which can only be computed for landmark-based segmentation methods like FC, PCA, Multi-atlas and the different variants of the HybridGNet model, which share the same output representation.

5.2 Generating landmark-based representations from dense segmentations

In this work we considered landmark-based segmentations with a fixed number of points, that enable establishing correspondences across images. This hinders the applicability of our method to tasks like brain tumor segmentation, where the shape of the structures of interest is not regular. Nonetheless, this is desirable in scenarios like population shape analysis, where we are interested in understanding how certain anatomical keypoints vary for different individuals. Unfortunately, in most segmentation datasets, only pixel-level annotations are available. In these cases, one could ask expert medical doctors to manually annotate specific landmarks in the mask contour (see Section 4.1.1 as an example) or

Table 3: Results for generating landmark annotations from dense segmentations in the jsrt dataset. Mean (std). HD in millimeters.

Model	MSE	Dice Lungs	HD Lungs	Dice Heart	HD Heart
PCA	77.2 (133.7)	0.978 (0.009)	6.02 (3.46)	0.97 (0.007)	4.37 (1.61)
FC	105.3 (173.2)	0.970 (0.014)	7.82 (3.96)	0.96 (0.014)	5.78 (2.94)
Multi-atlas	236.3 (244.8)	0.991 (0.004)	10.98 (8.53)	0.99 (0.006)	4.64 (2.48)
HybridGNet	96.9 (145.0)	0.970 (0.009)	7.65 (3.75)	0.96 (0.013)	6.02 (2.77)
1 IGSC: L6	70.5 (144.9)	0.983 (0.005)	5.54 (5.30)	0.97 (0.011)	4.02 (2.24)
2 IGSC: L6-5	55.1 (113.4)	0.991 (0.003)	3.92 (4.42)	0.99 (0.005)	2.58 (1.59)

Table 4: Domain shift results for landmark-based anatomical segmentation from JSRT dataset to Montgomery and Shenzhen. Mean (std). HD in pixels.

Model	Montgomery		Shenzhen	
	Dice Lungs	HD Lungs	Dice Lungs	HD Lungs
PCA	0.906 (0.082)	60.08 (36.89)	0.894 (0.054)	79.12 (47.73)
FC	0.897 (0.087)	60.02 (35.77)	0.895 (0.051)	77.11 (48.15)
Multi-alas	0.909 (0.080)	61.77 (31.62)	0.900 (0.054)	88.13 (48.94)
HybridGNet	0.909 (0.070)	55.97 (35.70)	0.901 (0.047)	72.13 (47.40)
1 IGSC: L6	0.930 (0.062)	48.22 (33.43)	0.914 (0.044)	67.39 (48.53)
2 IGSC: L6-5	0.954 (0.043)	45.50 (32.48)	0.935 (0.038)	64.46 (51.53)
UNet	0.944 (0.068)	127.72 (97.76)	0.933 (0.055)	220.89 (102.94)
UNet+PostDAE	0.907 (0.102)	119.53 (85.10)	0.906 (0.083)	135.05 (97.32)
nnUNet	0.955 (0.068)	44.79 (60.31)	0.949 (0.055)	61.31 (67.97)

perform automated estimation of landmarks from dense segmentations using HybridGNet. Our model can be trained to recover landmark-based representations from dense segmentation masks in a natural way. Thus, we trained our best performing models and baselines with dense segmentation masks as input (instead of images), to perform landmark estimation. Table 3 shows the results on the JSRT test set: the proposed HybridGNet 2 IGSC (Layers 6-5) outperforms the other baselines and architectures, proving useful in the building of shape models with landmark correspondences from pixel-level masks. Multi-atlas showed no differences in Dice with respect to our HybridGNet 2 IGSC (according to Wilcoxon’s test), but we observed that it loses track of the point-to-point correspondences as it is exhibited by the higher MSE error, which is computed for pairs of matching points.

HybridGNet was used to create landmark annotations for the Montgomery, Shenzhen and Padchest datasets, which originally did not include this type of segmentations. We are publicly releasing these new annotations² hoping that they will serve for future studies where point correspondences across individuals are required.

5.3 Domain shift (DS) evaluation

DS refers to a variation in the target (test) domain concerning the source (training) domain [53]. In most cases, such DS drops performance significantly as supervised learning assumes that training samples have the same distribution as the test samples. DS can be caused by multiple factors including changes in acquisition parameters, medical center or population demographics. We compared the effect of DS by measuring segmentation performance on datasets captured at different medical centers, i.e. training in the JSRT dataset and testing with Shenzhen and Montgomery. Table 4 shows how HybridGNet model greatly outperforms most baselines in terms of HD and Dice, and it is competitive with the self-configuring nnUNet (bold numbers indicate the best performance for methods of the same type, i.e. landmark and pixel-level methods). Differences between the two best performing methods (nnUNet and 2 IGSC: L6-5) and all other baselines are significant according to Wilcoxon test. On the contrary, differences between nnUNet and 2 IGSC: L6-5 are not significant, except for Dice Lung in the Shenzhen dataset. These results confirm the generalizability of the proposed model across medical centers.

²Annotations available at: <https://github.com/ngaggion/Chest-xray-landmark-dataset>

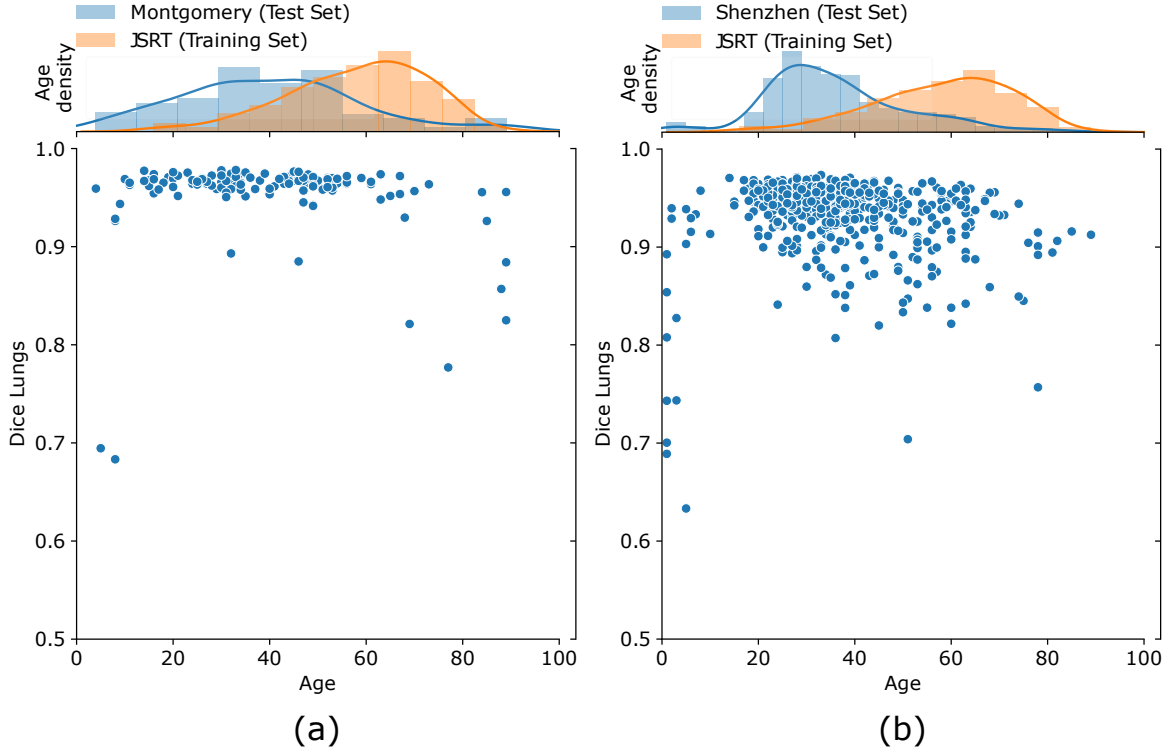


Figure 3: **Assessing the impact of domain shift by age distribution on lung segmentation.** Scatter plot of the lung Dice coefficient vs. age of patient for every individual in both (a) Montgomery and (b) Shenzhen datasets. Histograms show the age distribution for test (blue) and training sets (orange).

Moreover, recent studies on fairness in machine learning have shown that under-representation of certain demographic groups in the training data (e.g. in terms of gender [54] or ethnicity [55]) may result in biased models which present unequal performance in minority groups. Here we are interested in evaluating if the same holds for chest x-ray segmentation, in particular when considering age distribution shifts between training and test patients. To perform this analysis, we take our best performing model (HybridGNet 2 IGSC Layers 6-5) and build a scatter plot (see Figure 3) depicting the Dice coefficient for lung segmentation vs patients age. When observing the age histograms between training and test sets, we note that young patients are highly underrepresented in the training set. Interestingly, we found that model performance drastically drops for patients between 0-18 years old in both Montgomery and Shenzhen datasets, what can be attributed to the lack of young people on the JSRT database. Since the size of the organs tends to be smaller for younger patients (in particular we observed significant differences for the lung area), this can bias the learning process if this subpopulation is not well represented. This experiment highlights the importance of performing disaggregated analysis to detect potential subgroups where the model may under-perform.

5.4 Robustness to image occlusions (IO)

IO are common in chest x-rays, for example due to patient de-identification, that is covering protected information with black patches, or devices such as pacemakers, tubes, cables, or electrodes that can cover important parts of organs or other structures of interest. This is a common situation found in clinical centers, especially in images from hospitalized patients. We designed two experiments to assess

the robustness of HybridGNet to artificial and real IO that were not represented in the training set, and compare it with the other baselines.

First, we simulated artificial occlusions by overlapping a random black box on every image. We applied boxes of different sizes over the JSRT test set on random positions. Figures 4 (a) and (b) show Dice and HD distance for lungs and heart segmentation (averaged) as the occlusion block size increases. Although both nnUNet and UNet slightly outperform HybridGNet in Dice for very small occlusions, its performance drops with a steeper slope than HybridGNet as we increase the size of the occlusion block. Figure 4 (c) shows some qualitative results for three cases with different occlusion levels. Both quantitative and qualitative results show that HybridGNet is more robust to IO than pixel-level models.

Robustness to real occlusions produced by external devices was also assessed. To this end, we used 20 segmented images with pacemakers from Padchest as test set. To evaluate solely the occlusion effect on performance and alleviate DS issues due to intensity differences across different medical centers, we retrained the models (both HybridGNet and baseline) with an extended training dataset that includes Padchest images (without pacemakers). In Figure 5 we can see how our model outperforms the UNet both on Dice and Hausdorff distance, while having no significant differences when compared to the nnUNet but producing more anatomically plausible images qualitatively.

5.5 Model behaviour on pathological anatomy.

We are also interested in analyzing the behaviour of HybridGNet in the context of pathological anatomy. To this end, we followed the experimental setup introduced in [24] where a subset of patients from the Shenzhen database diagnosed with tuberculosis was considered. These patients have a collapsed lung and therefore a reduced air cavity. Every image was annotated by two expert radiologists following two different approaches to delineate the lungs (as discussed in [56]). The first approach was to segment only the air cavity of the lung field, i.e. segmenting only the dark areas (regions of lucency) and ignoring areas of increased attenuation (opacities), which correspond to infected lung tissue. Following [24] we call them *air* masks. In the second approach, annotators delineated the expected anatomy of the lung, including opaque areas following a comparative approach, mirroring the normal lung field onto the abnormal one. We call these *anatomy* masks.

We compared the segmentation performance of HybridGNet and UNet considering both types of annotations as ground-truth, when trained on JSRT (which contains only masks of non-pathological lungs). Quantitative results shown in Figures 6 (a) and (b) confirm that HybridGNet obtains results that are much closer to the *anatomy* masks than to the *air* masks, obtaining a higher Dice coefficient and a lower HD. Wilcoxon’s test showed that the difference between the means on both metrics was indeed significant. Conversely, this tendency is less pronounced for UNet predictions, suggesting that HybridGNet encourages more anatomically plausible predictions, while UNet focuses on local texture patterns. Figure 6 (c) shows examples of air and anatomy masks, and qualitative results of both methods for three different images. Regarding clinical utility, this opens the door for applications that combine both architectures: for example, the severity of tuberculosis infection could be estimated by measuring the difference between the UNet mask, representing the non-infected lung regions, and the HybridGNet mask, representing the healthy lung area if there was no lung collapse.

5.6 Cardiothoracic ratio estimation clinical use-case

A relevant intended use for lung and heart segmentation in chest x-rays is the detection of heart diseases, by identifying an enlargement of the cardiac silhouette. In radiology, this is done by measuring the CTR on a posteroanterior chest x-ray. This is calculated as the ratio between the (maximal) horizontal diameters of the heart and the thorax (inner edge of ribs/edge of pleura), which are manually measured by radiologists [57]. A normal CTR lays between 0.42 and 0.50, while a $CTR > 0.5$ is

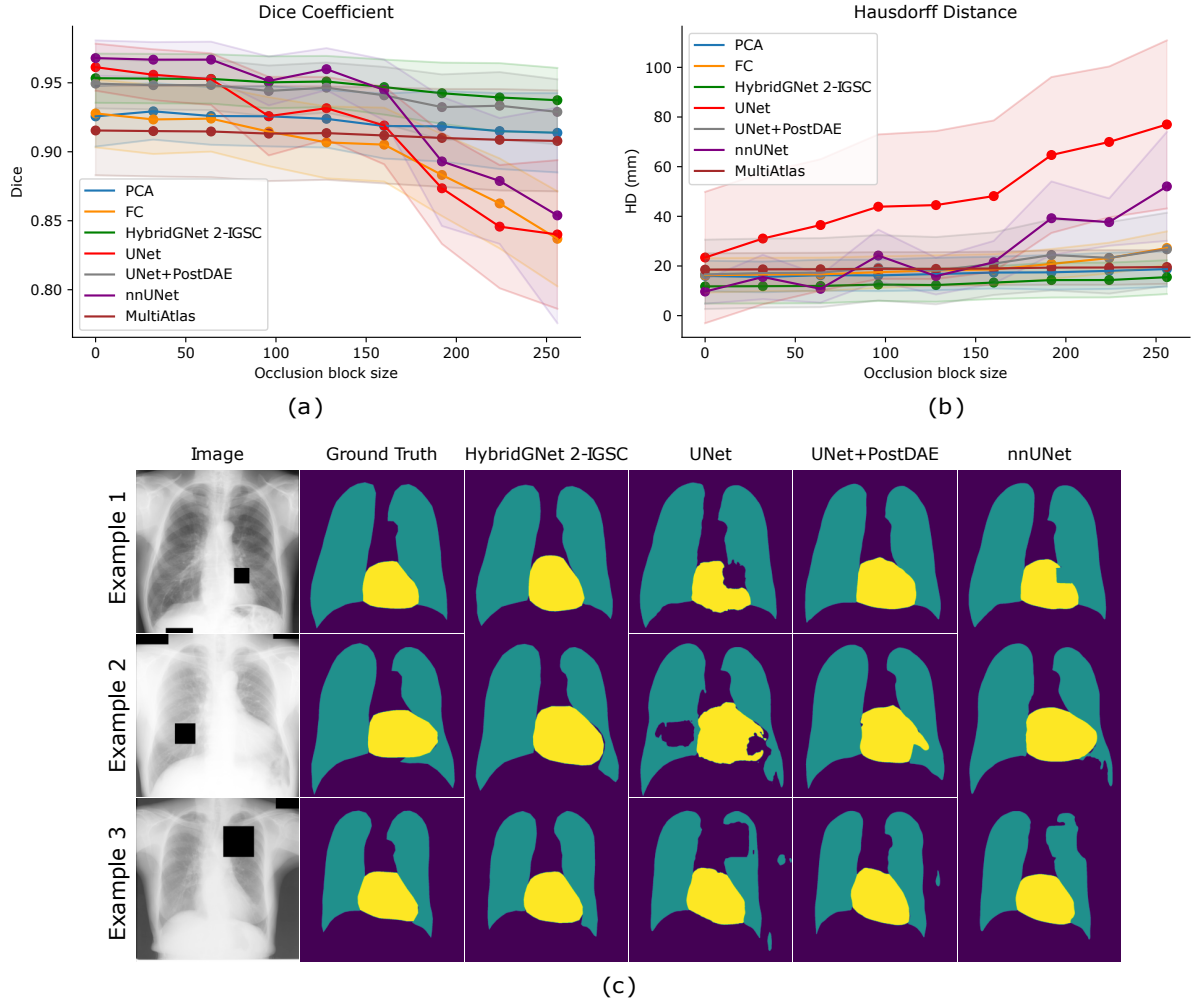


Figure 4: **Artificial occlusions study.** (a) Dice and (b) HD distance for increasing block size in artificial occlusions. (c) Shows qualitative results.

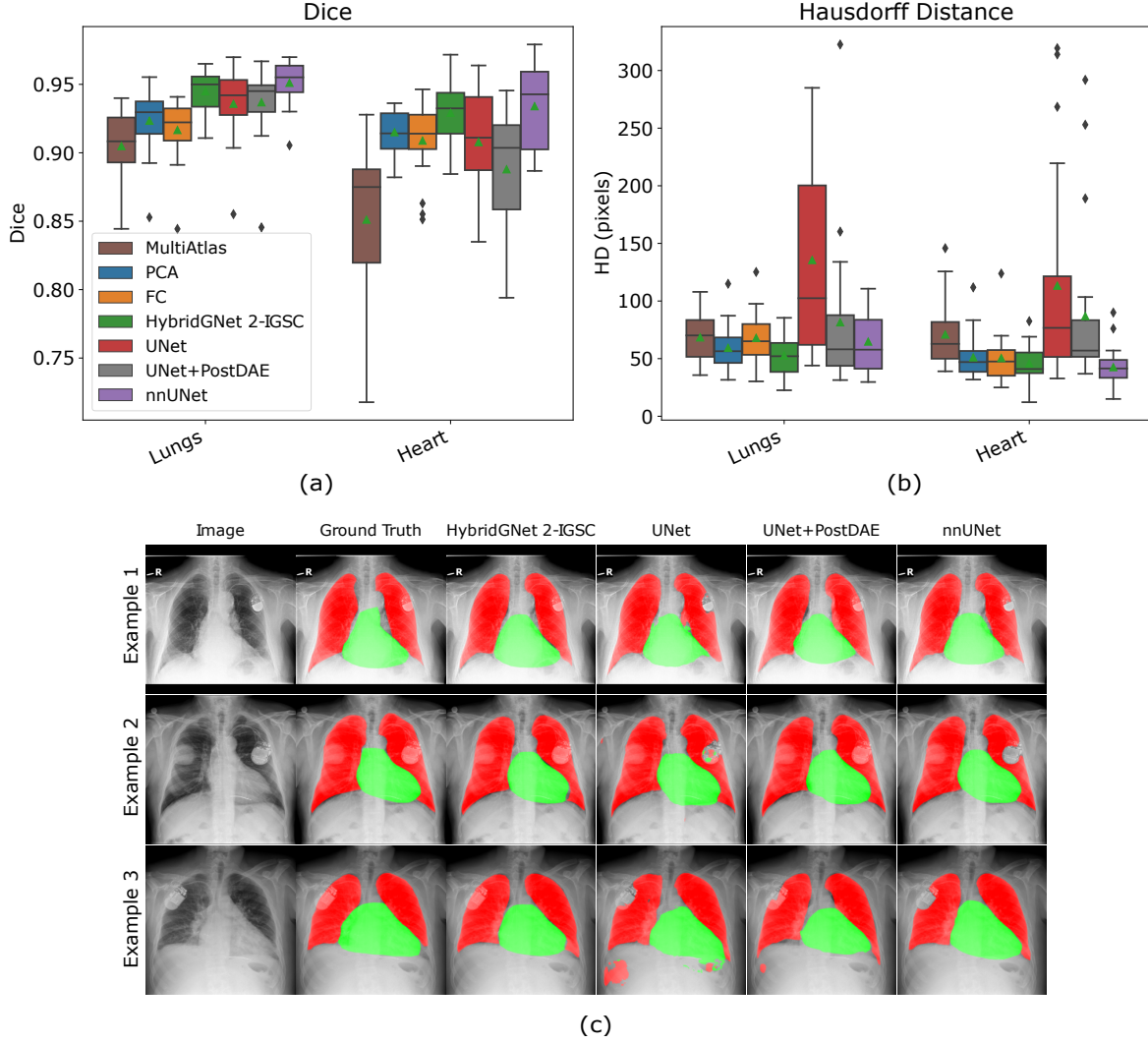


Figure 5: **Real occlusions study.** (a) Dice and (b) HD distances for the pacemaker Padchest subset. Wilcoxon test showed no significant differences between HybridGNet and nnUNet, the best two performing models. However, qualitative results (c) show how HybridGNet results in more anatomically plausible segmentations than nnUNet for complex cases.

considered an abnormal finding. For example, in young patients it might indicate a heart disease, such as cardiomegaly or pericardial effusion.

Manual calculation of CTR introduces observer variation and it is time consuming. Thus, here we evaluated the performance of HybridGNet for CTR estimation using a testing subset of 100 images from Padchest: 50 images with a *cardiomegaly* label, and 50 without this label. Two radiology specialists from Hospital Italiano de Buenos Aires collaborated in our study by manually calculating the CTR for this subset. The mean CTR among them was considered as ground-truth. To reduce the DS due to the change of medical center and the lack of pathological anatomy, we constructed an augmented training set by merging the JSRT images with a subset of 117 images from Padchest with *cardiomegaly* label. Since Padchest did not originally include landmark annotations, in this augmented set we used the ones generated from dense segmentations in the experiment described in section 5.2.

We compared model performance when training solely with JSRT images and when training with the augmented dataset. The predicted CTR was calculated automatically from HybridGNet outputs by measuring the maximum horizontal distance between lung borders and the maximum horizontal diameter of the heart mask. We found that the Pearson correlation coefficient between ground-truth CTR and predicted CTR increased when the model was trained with the augmented dataset. For the images with a ground-truth CTR < 0.5 (normal cardiac silhouette) correlation increased from 0.80 to 0.88 when target-domain images were included during training. This improvement was even stronger for abnormal cases (CTR > 0.5), increasing from 0.70 to 0.85. Figure 7 shows a scatter plot of the 100 test images as data points, where the diagonal represents a perfect agreement between the CTR measurement of HybridGNet and physicians. We can see how the model trained solely with normal cardiac silhouette cases (JSRT) tends to underestimate the CTR, while the model trained with target-domain cases improves CTR calculation on abnormal hearts. These results suggest that even when using models which encourage anatomically plausibility, the construction of diverse databases (i.e. including representative samples of the target population) is still needed so that performance is maintained in real clinical scenarios.

6 Conclusions

In this paper we introduced HybridGNet, a new method to perform landmark based anatomical segmentation via hybrid graph neural networks with image-to-graph localized skip connections. Our study confirms that incorporating connectivity information through the graph adjacency matrix helps to improve anatomical plausibility and accuracy of the results when compared with other landmark-based and pixel-level segmentation models. We also showcased several application scenarios for HybridGNet in the context of chest-x ray image analysis, and assessed its robustness with respect to different types of domain shift and image occlusions. When compared with dense pixel-level prediction models, we observed that HybridGNet achieves faster inference time, is much more robust to strong image occlusions and produces more anatomically plausible results in these contexts. We also evaluated the clinical utility of our model in the context of cardiothoracic ratio estimation and audited potential biases that may appear due to under-representation of certain demographic groups or pathologies. Our results go in line with the evidence reported in recent studies on fairness in biomedical image segmentation, highlighting the importance of constructing diverse databases which include representative demographic samples from the targeted population. In the future, we plan to extend the proposed HybridGNet model to volumetric images, where graphs can be used to represent meshes instead of contours.

7 Acknowledgements

The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-PIC50420150100098LI, CAID-PIC-50220140100084LI) and ANPCyT (PICT 2016-0651, PICT 2018-03907).

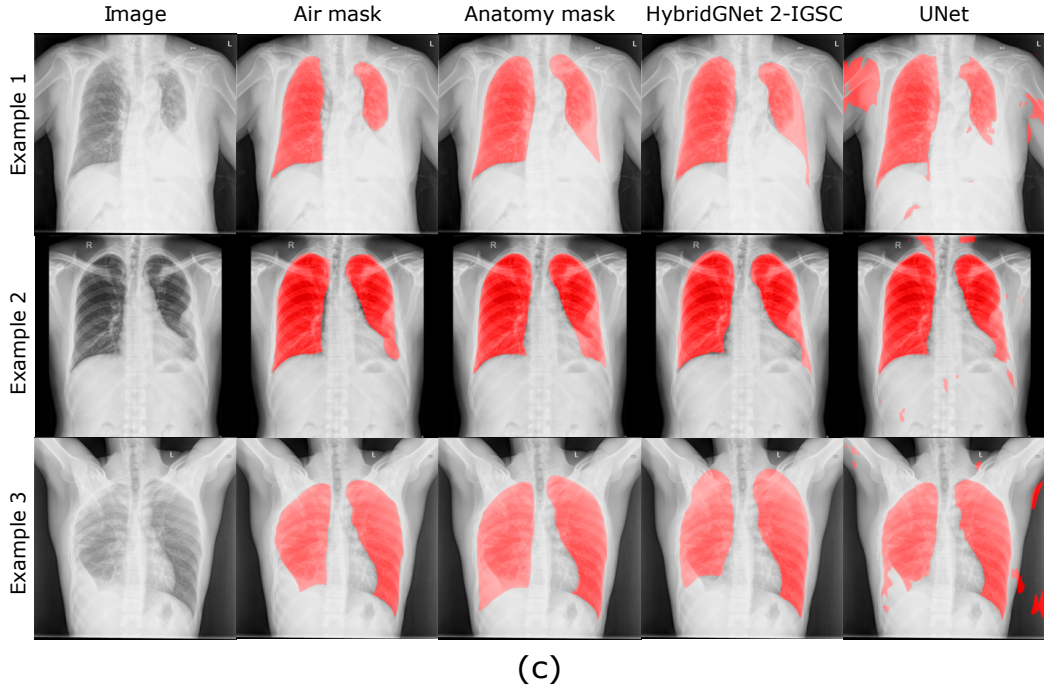
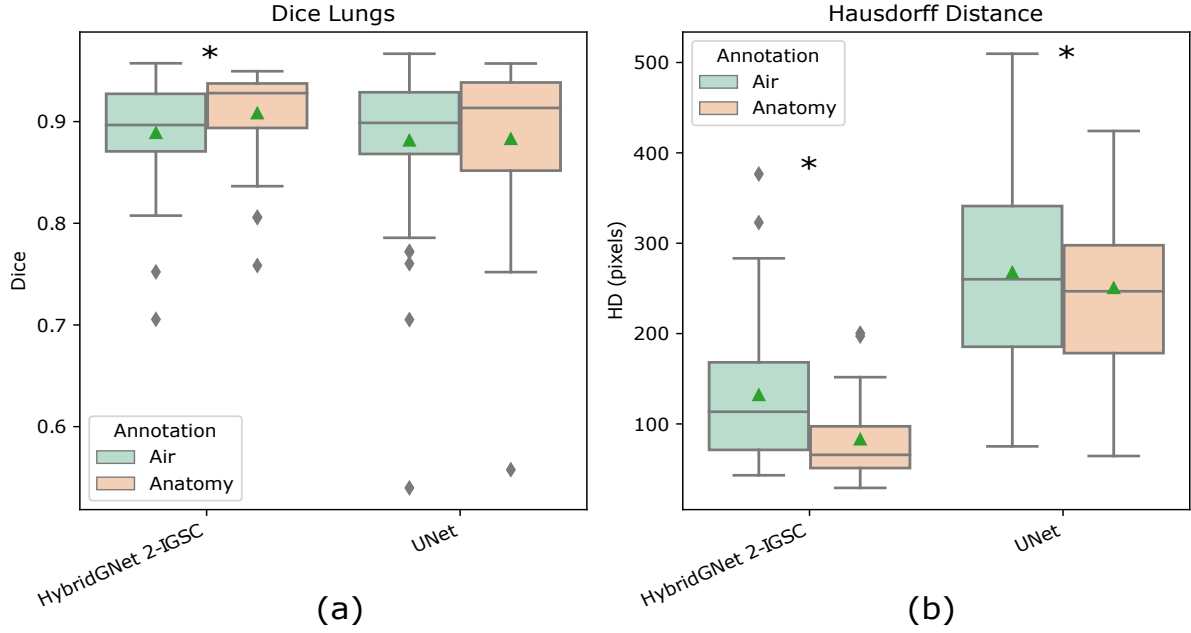


Figure 6: **Model behaviour on pathological anatomy.** Top boxplots show Dice coefficient (a) and HD (b), taking the air mask as ground truth (green) and taking the anatomy mask as ground truth (orange) for the HybridGNet and UNet models. * indicates significant differences between means according to Wilcoxon's test (p-value < 0.05). (c) Visual examples for the air and anatomy masks, and outputs given by both models. Results shows that our model tends to predict masks that follow the expected shape of the organs, while UNet predictions resample the visible air section of the lungs.

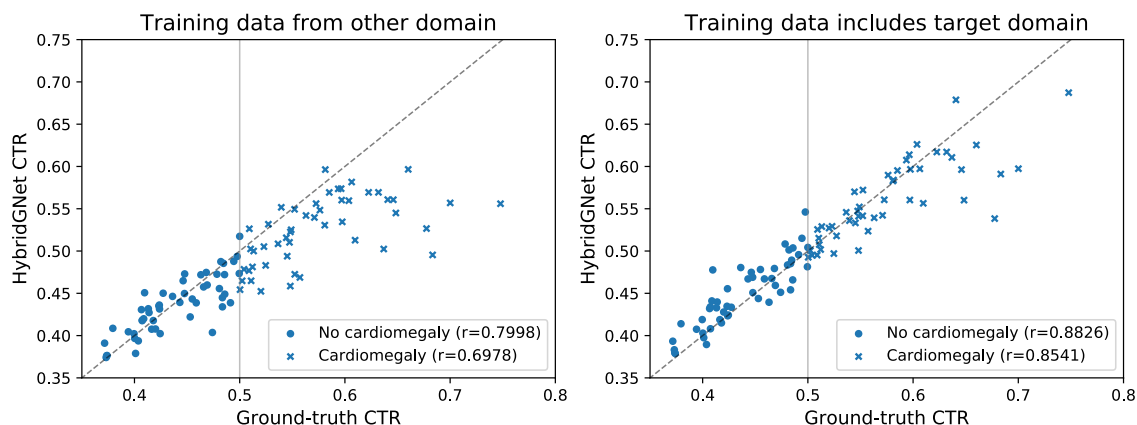


Figure 7: **CTR study.** Ground-truth CTR vs HybridGNet CTR when training with JSRT dataset only (left) and the dataset augmented with cardiomegaly images from Padchest (right). The vertical line indicates the boundary between normal and abnormal CTR.

We thank Alexandros Karargyris, Sema Candemir and Stefan Jaeger for sharing the segmentation masks used in the pathological anatomy experiment. We also thank Facundo Diaz and Martina Aineseder -specialists from the Radiology Department at Hospital Italiano de Buenos Aires- for their collaboration in the annotation of the Padchest images.

References

- [1] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [2] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *4th international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.
- [3] A. Reinke *et al.*, “Common limitations of image processing metrics: A picture story,” 2021.
- [4] S. Bohlender, I. Oksuz, and A. Mukhopadhyay, “A survey on shape-constraint deep learning for medical image segmentation,” *arXiv preprint arXiv:2101.07721*, 2021.
- [5] R. El Jurdi, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah, “High-level prior-based loss functions for medical image segmentation: A survey,” *Computer Vision and Image Understanding*, vol. 210, p. 103248, 2021.
- [6] T. Heimann and H.-P. Meinzer, “Statistical shape models for 3d medical image segmentation: a review,” *Medical image analysis*, vol. 13, no. 4, pp. 543–563, 2009.
- [7] H. Boussaid, I. Kokkinos, and N. Paragios, “Discriminative learning of deformable contour models,” in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2014, pp. 624–628.
- [8] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: going beyond euclidean data,” *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

- [9] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” *arXiv*, 2013.
- [10] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” *arXiv preprint arXiv:1606.09375*, 2016.
- [11] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1263–1272.
- [12] T. N. Kipf and M. Welling, “Variational graph auto-encoders,” *arXiv preprint arXiv:1611.07308*, 2016.
- [13] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, “Generating 3d faces using convolutional mesh autoencoders,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 704–720.
- [14] N. Gaggion, L. Mansilla, D. H. Milone, and E. Ferrante, “Hybrid graph convolutional neural networks for landmark-based anatomical segmentation,” in *Medical Image Computing and Computer Assisted Intervention*. Cham: Springer International Publishing, 2021.
- [15] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, “Training models of shape from sets of examples,” in *BMVC92*. Springer, 1992.
- [16] P. D. Sozou, T. F. Cootes, C. J. Taylor, E. Di Mauro, and A. Lanitis, “Non-linear point distribution modelling using a multi-layer perceptron,” *Image and Vision Computing*, vol. 15, no. 6, 1997.
- [17] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” in *European conference on computer vision*. Springer, 1998.
- [18] B. Zitova and J. Flusser, “Image registration methods: a survey,” *Image and vision computing*, vol. 21, no. 11, pp. 977–1000, 2003.
- [19] A. F. Frangi, W. J. Niessen, D. Rueckert, and J. A. Schnabel, “Automatic 3d asm construction via atlas-based landmarking and volumetric elastic registration,” in *Biennial International Conference on Information Processing in Medical Imaging*. Springer, 2001, pp. 78–91.
- [20] G. Heitz, T. Rohlfing, and C. R. Maurer Jr, “Automatic generation of shape models using nonrigid registration with a single segmented template mesh,” in *VMV*, 2004, pp. 73–80.
- [21] R. Paulsen, R. Larsen, C. Nielsen, S. Laugesen, and B. Ersbøll, “Building and testing a statistical shape model of the human ear canal,” in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2002, pp. 373–380.
- [22] M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, and I. Kokkinos, “Sub-cortical brain structure segmentation using f-cnn’s,” in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 269–272.
- [23] R. E. Jurdia, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah, “High-level prior-based loss functions for medical image segmentation: A survey,” *arXiv preprint arXiv:2011.08018*, 2020.
- [24] A. J. Larrazabal, C. Martinez, and E. Ferrante, “Anatomical priors for image segmentation via post-processing with denoising autoencoders,” in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 585–593.
- [25] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O’Regan *et al.*, “Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation,” *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384–395, 2017.

- [26] F. Milletari, A. Rothberg, J. Jia, and M. Sofka, “Integrating statistical prior knowledge into convolutional neural networks,” in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017.
- [27] R. Bhalodia, S. Y. Elhabian, L. Kavan, and R. T. Whitaker, “Deepssm: a deep learning framework for statistical shape modeling from raw images,” in *International Workshop on Shape in Medical Imaging*. Springer, 2018, pp. 244–257.
- [28] R. Bhalodia, S. Elhabian, J. Adams, W. Tao, L. Kavan, and R. Whitaker, “Deepssm: A blueprint for image-to-shape deep learning models,” *arXiv preprint arXiv:2110.07152*, 2021.
- [29] S. Foti, B. Koo, T. Dowrick, J. Ramalhinho, M. Allam, B. Davidson, D. Stoyanov, and M. J. Clarkson, “Intraoperative liver surface completion with graph convolutional vae,” in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*. Springer, 2020, pp. 198–207.
- [30] M. Drozdal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, “The importance of skip connections in biomedical image segmentation,” in *Deep learning and data labeling for medical applications*. Springer, 2016, pp. 179–187.
- [31] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang, “Pixel2mesh: Generating 3d mesh models from single RGB images,” 2018. [Online]. Available: <http://arxiv.org/abs/1804.01654>
- [32] U. Wickramasinghe, E. Remelli, G. Knott, and P. Fua, “Voxel2mesh: 3d mesh model generation from volumetric data,” in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020.
- [33] H. Ling, J. Gao, A. Kar, W. Chen, and S. Fidler, “Fast interactive object annotation with curve-gcn,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5257–5266.
- [34] G. Gkioxari, J. Malik, and J. Johnson, “Mesh R-CNN,” *CoRR*, vol. abs/1906.02739, 2019. [Online]. Available: <http://arxiv.org/abs/1906.02739>
- [35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [36] A. Besbes and N. Paragios, “Landmark-based segmentation of lungs while handling partial correspondences using sparse graph-based priors,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 989–995.
- [37] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” *IEEE signal processing magazine*, vol. 30, 2013.
- [38] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [39] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Artificial intelligence and statistics*, 2015, pp. 562–570.
- [40] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

- [41] B. van Ginneken, M. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, 2006.
- [42] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014.
- [43] S. e. a. Jaeger, "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014.
- [44] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical image analysis*, vol. 66, p. 101797, 2020.
- [45] J. Alvé, F. Kahl, M. Landgren, V. Larsson, J. Ulén, and O. Enqvist, "Shape-aware label fusion for multi-atlas frameworks," *Pattern Recognition Letters*, vol. 124, pp. 109–117, 2019.
- [46] J. Alvé, F. Kahl, M. Landgren, V. Larsson, and J. Ulén, "Shape-aware multi-atlas segmentation," in *2016 23rd International Conference on Pattern Recognition (Icpr)*. IEEE, 2016, pp. 1101–1106.
- [47] K. Marstal, F. Berendsen, M. Staring, and S. Klein, "Simpleelastix: A user-friendly, multi-lingual library for medical image registration," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 134–142.
- [48] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.
- [49] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Proc Systems*, 2019.
- [50] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.
- [52] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [53] A. Choudhary, L. Tong, Y. Zhu, and M. D. Wang, "Advancing medical imaging informatics by deep learning-based domain adaptation," *Yearbook of medical informatics*, vol. 29, no. 01, pp. 129–138, 2020.
- [54] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, 2020.
- [55] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, and A. P. King, "Fairness in cardiac mr image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," in *MICCAI 2021*. Springer, 2021, pp. 413–423.
- [56] A. Karargyris, J. Siegelman, D. Tzortzis, S. Jaeger, S. Candemir, Z. Xue, K. C. Santosh, S. Vajda, S. Antani, L. Folio, and G. R. Thoma, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 1, pp. 99–106, 2016.

- [57] C. S. Danzer, “The cardiothoracic ratio: an index of cardiac enlargement.” *The American Journal of the Medical Sciences (1827-1924)*, vol. 157, no. 4, p. 513, 1919.

A multi-view hybrid graph convolutional architecture for Volume-to-Mesh extraction in cardiovascular MR images

A multi-view hybrid graph convolutional architecture for Volume-to-Mesh extraction in cardiovascular MR images

Nicolás Gaggion¹, Benjamin Matheson², Yan Xia², Rodrigo Bonazzola², Nishant Ravikumar², Diego H. Milone¹, Alejandro F. Frangi², and Enzo Ferrante¹

¹Institute for Signals, Systems and Computational Intelligence, sinc(i),
CONICET-UNL, Santa Fe, Argentina

²Centre for Computational Imaging and Simulation Technologies in Biomedicine
(CISTIB), School of Computing, University of Leeds, Leeds, UK

Abstract

Recent advancements in cardiac imaging have revolutionized heart characterization, with cardiovascular magnetic resonance (CMR) imaging emerging as a crucial tool for scrutinizing cardiac morphology and function. Essential to this endeavor are 3D surface and volumetric meshes derived from CMR images, facilitating computational anatomy, biomarker discovery, and in-silico simulations. However, conventional surface mesh generation methods like active shape models and multi-atlas segmentation tend to be highly time-consuming and require complex processing pipelines to generate simulation-ready 3D meshes. In response, we introduce HybridVNet, a novel architecture for direct image-to-mesh extraction seamlessly integrating standard convolutional neural networks with graph convolutions, which we prove is capable of efficiently handling both surface and volumetric meshes by encoding them as graph structures. To further enhance accuracy, we extend the proposed architecture to a multi-view setting and show it can increase performance in the context of cardiac MR mesh generation. Our model combines traditional convolutional networks with variational graph generative models, deep supervision vision, and mesh-specific regularization. Experiments on a comprehensive dataset from the UK Biobank confirms HybridVNet’s potential to significantly advance cardiac imaging and computational cardiology by efficiently generating high-fidelity meshes from CMR images.

1 Introduction

Cardiac imaging has undergone significant advancements in recent years, becoming an indispensable tool in the diagnosis, treatment planning, and management of various cardiovascular diseases. One of the pivotal components for such advancements is the extraction of accurate 3D meshes from CMR images, which serve as the foundation for computational simulations [7], biomarkers discovery [5], and analyzing hearth deformation and dynamics [4].

Despite the undeniable importance of cardiac mesh extraction, the task is fraught with challenges. Traditional methods, such as active shape models [21] and multi-atlas segmentation [3], often require extensive computational resources and can be time-consuming. The variability in heart shapes, sizes, and pathologies further complicates the extraction process, necessitating a method that is both robust and adaptable.

Furthermore, the transition from 2D image slices to a cohesive 3D representation demands a seamless integration of spatial information, specially when modeling tetrahedral meshes. Current methodologies often require intricate post-processing steps to refine the meshes [7, 20], making them suitable for simulations, which can introduce additional sources of error and prolong the overall processing time.

In light of these challenges, there is a pressing need for innovative approaches that can streamline the cardiac mesh extraction process, enhance accuracy, and reduce the time and computational overhead. This paper introduces a novel method for generating high-quality surface and volumetric meshes, immediately suitable for computational models, directly from cardiovascular magnetic resonance (CMR) imaging.

Generating detailed and anatomically plausible ventricular meshes from CMR images is challenging and time-consuming if performed manually. Automatic methods for mesh extraction mostly depend on a concatenation of several stages, usually including voxel-level segmentation, surface mesh extraction, and a final volumetric mesh generation step [7, 20]. One of the main limitations of these approaches is that voxel-level segmentation techniques (usually deep learning models like the U-Net [31] or the volumetric V-Net [18]) are prone to introduce errors due to the local support of the convolutional models, resulting in unrealistic masks with holes or spurious segmentations [16]. Other approaches are based on deforming an initial template. In that regard, previous studies have proposed to either estimate mesh node displacements [27, 22] or directly deform the space enclosing a simulation-ready whole heart template [15]. However, the accuracy of these approaches is bounded by the quality of the estimated deformations.

Closer to our work, in [33, 35] a neural network is trained end-to-end to estimate parameterised shapes directly from images. These methods infer the parameters of a Principal Component Analysis (PCA) shape model using convolutional neural networks. Despite the utility of PCA-based models in mesh extraction from volumetric images, their expressiveness is inherently limited by the linearity of PCA. To address this limitation, we enhance the expressiveness of parametric shapes by leveraging a graph-convolutional decoder, capable of handling both surface and volumetric meshes. Notably, previous studies have employed generative models based on geometric deep learning [10, 9] for extracting contours from 2D images. Our contribution builds upon this foundation, introducing the HybridVNet architecture and its multi-view variant, which combines standard 3D convolutions for volumetric image encoding, with a decoder based on spectral graph convolutions for cardiac mesh generation.

Contributions: Our primary contributions encompass the development of a multi-view volumetric hybrid graph convolutional model capable of seamlessly integrating multiple CMR views within a jointly-learned latent space, directly yielding meshes from images. Our model exhibits a natural ability to produce both cardiac surfaces and tetrahedral meshes, suited for finite element simulations. We further explore classic regularization techniques for surface meshes and introduce a novel differentiable regularization term specifically tailored for tetrahedral meshes, markedly enhancing element quality. Importantly, while previous works often relied on cropped regions of volumetric images, our model demonstrates exceptional performance in both cropped regions and complete images, with only a marginal decrease in performance when compared to competing methods. The model’s performance is rigorously evaluated using the UK Biobank CMR dataset [26], underscoring the clinical relevance of our results as we compare cardiac function indices extracted from our resulting meshes with those derived from ground-truth annotations.

2 Volume-to-Mesh extraction in cardiovascular MR

2.1 Reference CMR meshes and images

The reference cohort of 3D surface meshes used in this study was introduced in Xia et al (2022) [35]. It was created by registering a high resolution atlas of the human heart [30] to manually delineated 2D contours at end-diastole (ED) and end-systole (ES). The rationale behind the selection of the subjects chosen for manual segmentation, as well as the methodology followed, can be found in [24]. This atlas consists of a mesh model that includes six separate structures: the left ventricle (LV), the right ventricle (RV), the left atrium (LA), the right atrium (RA), and the ascending aorta. The reference cohort was pre-processed to obtain a unified mesh structure, since the meshes did not share any faces between different parts of the heart, such as between the LV and RV. Therefore, graph-based models

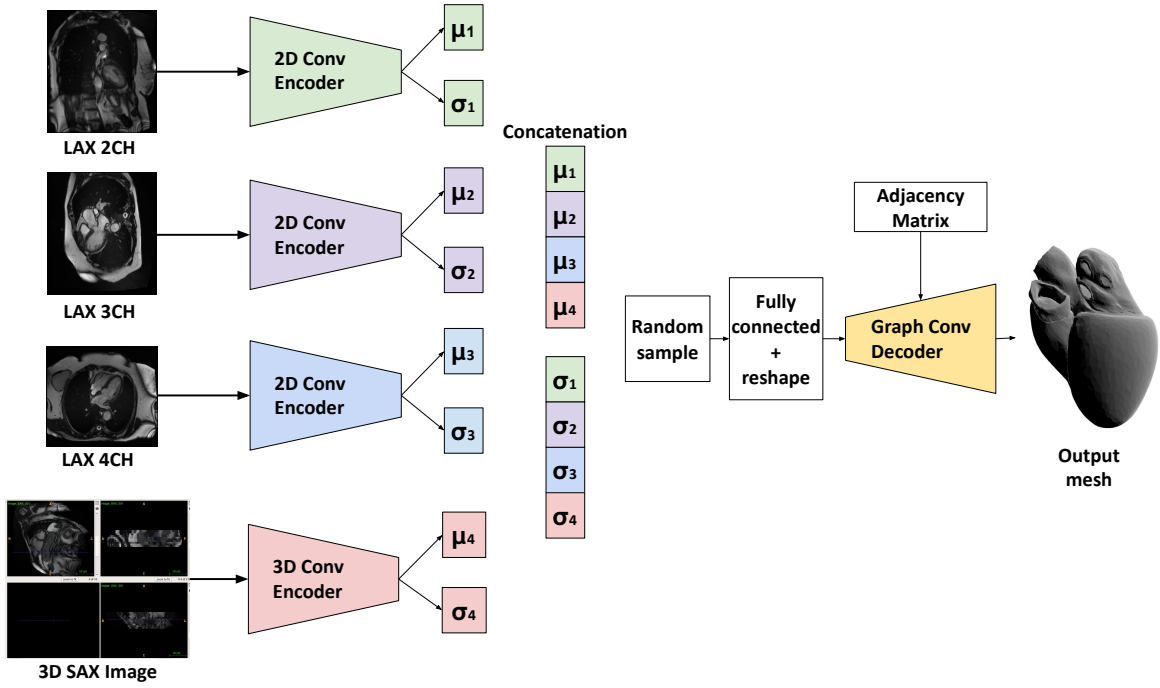


Figure 1: **Multi-view HybridVNet model architecture:** The proposed model uses a variational encoder-decoder architecture to generate a graph representation of a desired organ from multi-view input images. The encoder consists of independent branches for each input view, which are concatenated to obtain a joint latent space. The latent code is then passed through a fully connected layer and reshaped to obtain initial node features for the graph convolutional decoder. This decoder uses the initial node features to generate a final graph representation of the organ.

were not able to share information between different parts of the heart as there was no connection between them. To obtain watertight surface meshes and to make the best use of the full heart mesh connectivity, we had to look into the original atlas to find which nodes were originally connected prior to the registration procedure, which ended up disconnecting them as each part of the heart was registered individually. Then, we removed the repeated nodes and generated a new set of faces for the full dataset. Since the cohort was obtained by registering an atlas, note that each of the final ground-truth meshes has the same number of nodes and the same set of faces, which gives the same adjacency matrix. More details about the image and surface mesh dataset are provided in Section 3.1.

Volumetric mesh dataset generation

We derived volumetric mesh ground-truth annotations from the surface meshes, using the one-to-one correspondence between surface nodes to register a volumetric atlas to the cardiac surface mesh dataset. We used Simpleware’s ScanIP (Simpleware Ltd, Exeter, UK) [32] to construct the volumetric atlas mesh. We imported heart structures from the human heart atlas [30] as individual closed surface meshes of triangular elements, then populated the hollow surface meshes with tetrahedral elements, setting the elements at the interfaces between different cardiac structures to share nodes. This resulted in a mesh of 408,764 elements. We registered the volumetric atlas to the whole surface dataset using mesh-to-mesh thin plate spline warping, using the Vedo library [19].

2.2 HybridVNet formulation

As shown in Figure 1, our HybridVNet model receives multiple CMR views as input: the short-axis view (SAX), which is a 3D cross-sectional view of the heart acquired perpendicular to the long axis; and three different 2D long-axis views (LAX), for two, three and four chambers of the heart (LAX 2CH, LAX 3CH and LAX 4CH, respectively), providing 2D cross-sectional views acquired parallel to the long axis. Given these four images (one volumetric and three 2D), we aim at generating a (surface or tetrahedral) mesh representing the structures of interest.

Consider a dataset $\mathcal{D} = \{(\mathbf{I}, \mathbf{G})_n\}_{0 < n \leq N}$, composed of N samples of multi-view CMR images $\mathbf{I} = (\mathbf{I}^{\text{LAX 2CH}}, \mathbf{I}^{\text{LAX 3CH}}, \mathbf{I}^{\text{LAX 4CH}}, \mathbf{I}^{\text{SAX}})$ and their associated cardiac meshes as graphs $\mathbf{G} = \langle V, \mathbf{A}, \mathbf{X} \rangle$. V is the set of M nodes or vertices ($|V| = M$), $\mathbf{A} \in \{0, 1\}^{M \times M}$ is the adjacency matrix indicating the connectivity between pairs of nodes ($a_{ij} = 1$ indicates an edge connecting vertices i and j , and $a_{ij} = 0$ otherwise) and $\mathbf{X} \in \mathbb{R}^{M \times s}$ is a function (represented as a matrix) assigning a feature vector to every node. In our case, it assigns a 3-dimensional spatial coordinate (the mesh vertex position, $s = 3$). In this situation, since our dataset includes meshes with the same number of nodes and same connectivity by construction, we can use spectral graph convolutions to decode meshes from a latent space [28, 5].

The proposed model consists of a hybrid variational encoder-decoder architecture with multiple inputs. An image convolutional encoder, f_e^I , learns a latent representation of the input images, and a spectral graph convolutional decoder, f_d^G , generates a graph representation of the organ. Since our input consists of four different images with varying shapes and views, we use a multi-view encoder to handle it. To this end, independent encoder branches are defined for each image view, and a joint latent space is constructed by concatenating their outputs. For all types of LAX images, we use 2D convolutional encoders, $f_e^{\text{LAX 2CH}}$, $f_e^{\text{LAX 3CH}}$ and $f_e^{\text{LAX 4CH}}$, with residual convolutions [11]. For the 3D SAX image, we use a 3D convolutional encoder, f_e^{SAX} , consisting of 3D residual blocks interleaved by max-pooling operations. Then, the multi-view input encoder is defined as:

$$f_e^I = [f_e^{\text{LAX 2CH}}, f_e^{\text{LAX 3CH}}, f_e^{\text{LAX 4CH}}, f_e^{\text{SAX}}]. \quad (1)$$

In consequence, our model uses a variational encoder-decoder architecture to generate a graph representation of a desired organ from multi-view input images. The encoder maps the input to a lower-dimensional embedding which represents the parameters of a latent distribution, $\mathbf{z} = f_e^I(\mathbf{I}^{\text{LAX 2CH}}, \mathbf{I}^{\text{LAX 3CH}}, \mathbf{I}^{\text{LAX 4CH}}, \mathbf{I}^{\text{SAX}})$. This latent distribution is then sampled using the reparametrisation trick [14], passed through a fully connected layer and reshaped to obtain initial node features for the graph convolutional decoder, f_d^G . Following the variational autoencoder formulation, the latent code is assumed to be sampled from a multivariate Gaussian posterior, $Q(\mathbf{z}|\mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}))$. The distribution is parameterised by the concatenation of outputs from the joint multi-view encoder, $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = f_e(\mathbf{I})$. Given a sample of the latent code, $\mathbf{z}$, the graph representation of the organ can be obtained through the decoder $f_d^G(\mathbf{z})$.

The model is trained by minimizing a loss function defined as:

$$\mathcal{L} = \mathcal{L}_r(f_d(f_e(\mathbf{I})), \mathbf{G}) + \lambda_{KL} \mathcal{L}_{KL}(Q(\mathbf{z}|\mathbf{I})||\mathcal{N}(0, 1)), \quad (2)$$

where the first term is the reconstruction loss based on the mean squared error (MSE) of the vertex positions, the second term imposes a unit Gaussian prior $\mathcal{N}(0, 1)$ for the latent posteriors via the KL divergence loss ($\mathcal{L}_{KL}$) and λ_{KL} is a weighting factor.

Deeply-supervised spectral graph decoder

To generate the graph representation of the target organ, we employ a decoder constructed using spectral graph convolutional neural networks (GCNN). Spectral convolutions are founded on the eigen-decomposition of the graph Laplacian matrix. In this context, we adopt the spectral convolutions introduced by Defferrard et al. (2016) [6], which constrain the filters to polynomial filters. This constraint arises from the observation that polynomial filters exhibit strict localization in the vertex

domain, consequently reducing the computational complexity of the convolutional operation. For an in-depth understanding of spectral convolutions, please refer to [6].

In essence, a spectral convolutional layer operates much like standard convolutions applied to images and feature maps. It takes an input feature matrix $\mathbf{X}^\ell$ and produces filtered versions $\mathbf{X}^{\ell+1}$ as output. Our spectral decoder architecture comprises five graph-convolutional layers, each complemented by ReLU nonlinearities with previous Layer Normalization [2]. These layers are strategically interleaved with four fixed graph unpooling layers, allowing the network to learn representations at multiple resolutions.

To obtain these multiple resolutions, we implement the technique outlined by Ranjan et al. (2018) [28] to construct pairs of pooling and unpooling layers. The process begins by estimating the pooling matrix, achieved through an iterative contraction of vertex pairs while maintaining precise surface error approximations using quadric matrices into the atlas surface mesh. Simultaneously, the unpooling matrix is derived to enable the reversal of the pooling transformation. This process is repeated four times (on the previously pooled version of the atlas), resulting in four sets of pooling and unpooling layers, each reducing and increasing the number of nodes by a factor of two, respectively. Importantly, this set of pooling and unpooling matrices remains fixed during the training process, as they are estimated only once for the atlas surface mesh.

To increase our model’s performance, we apply the concept of deep supervision [17], which involves supervising the network at various resolution levels. During training, we utilize the estimated pooling operation to obtain down-sampled versions of the ground-truth meshes, enabling us to minimize the reconstruction error at each resolution level. Ultimately, we employ a final graph-convolutional layer, without bias and identity activation function, to predict the final vertex positions.

The incorporation of deep supervision terms leads to the following loss function:

$$\mathcal{L} = \mathcal{L}_r + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{DS} \sum_{i=1}^4 \mathcal{L}_r^i, \quad (3)$$

where $\mathcal{L}_{KL}$ is the previously defined KL term, λ_{DS} is a weighting factor, and the index i indicates the resolution level of the graph.

Mesh regularization loss functions

To ensure smooth meshes, state-of-the-art approaches to surface mesh generation often use regularizers such as normal regularization, edge length regularization, and Laplacian smoothing ($\mathcal{L}_{lap}$), as introduced in [34], which we also incorporate. However, these existing metrics were originally designed for triangular surface meshes, and therefore do not take into account the structure of tetrahedral elements in a volumetric mesh [22]. To address this limitation, we propose a new regularization loss function that is specifically designed for directly generating tetrahedral volumetric meshes. We introduce our new *tetrahedral element regularization* loss as follows:

$$\mathcal{L}_{ter} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{1}{6} \sum_{j=1}^6 \left(\|\mathbf{e}_j^i\|_2 - \frac{1}{6} \left(\sum_{k=1}^6 \|\mathbf{e}_k^i\|_2 \right) \right)^2, \quad (4)$$

where N_t is the number of tetrahedra, i represents the i^{th} tetrahedron, and $\mathbf{e}_j^i$ and $\mathbf{e}_k^i$ represent edges of that tetrahedron. This term encourages tetrahedron regularity by penalizing elements with edge lengths too different from the average edge length for each tetrahedron. Thus, the final loss function used to train the model results in:

$$\mathcal{L} = \mathcal{L}_r + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{DS} \sum_{i=1}^4 \mathcal{L}_r^i + \lambda_{reg}\mathcal{L}_{reg}, \quad (5)$$

where $\mathcal{L}_{reg}$ can be any of the previously mentioned regularization losses: $\mathcal{L}_{lap}$ for the surface case or $\mathcal{L}_{ter}$ for the volumetric case, and λ_{reg} is the corresponding weighting factor.

3 Experimental Setup

	Metrics	MCSI-Net SAX	HybridVNet		MCSI-Net SAX-LAX	MV-HybridVNet	
		Cropped	Full Image	Cropped	Cropped	Full Image	Cropped
LV Endo	DC $\uparrow$	0.87 (0.05)	0.89 (0.05)	0.90 (0.04)	0.88 (0.05)	0.90 (0.04)	0.91 (0.04)
	HD $\downarrow$	5.13 (1.97)	4.48 (1.32)	4.08 (1.22)	4.74 (1.75)	4.22 (1.22)	3.89 (1.18)
	MCD $\downarrow$	1.93 (0.83)	1.67 (0.55)	1.49 (0.49)	1.86 (0.79)	1.55 (0.51)	1.39 (0.46)
LV Myo	DC $\uparrow$	0.76 (0.09)	0.80 (0.06)	0.83 (0.05)	0.78 (0.08)	0.81 (0.05)	0.84 (0.04)
	HD $\downarrow$	5.31 (1.98)	4.71 (1.36)	4.23 (1.27)	4.75 (1.76)	4.40 (1.26)	3.96 (1.23)
	MCD $\downarrow$	1.97 (0.95)	1.71 (0.56)	1.49 (0.51)	1.86 (0.82)	1.57 (0.52)	1.35 (0.46)
RV Endo	DC $\uparrow$	0.85 (0.06)	0.85 (0.05)	0.86 (0.05)	0.85 (0.06)	0.86 (0.05)	0.87 (0.05)
	HD $\downarrow$	7.11 (2.78)	6.97 (2.31)	6.44 (2.19)	7.06 (2.64)	6.79 (2.23)	6.13 (2.23)
	MCD $\downarrow$	2.34 (0.98)	2.10 (0.64)	1.90 (0.57)	2.27 (0.95)	1.99 (0.59)	1.76 (0.59)

Table 1: Quantitative ventricle segmentation results for surface meshes. An up arrow ($\uparrow$) indicates that higher values are better, while a down arrow ($\downarrow$) indicates that lower values are better.

3.1 Data and annotations

The data for this study was collected from the UK Biobank (UKB) under access applications number 2964 and 11350. The study adhered to the guidelines set forth in the Declaration of Helsinki and received ethical approval from the National Health Service National Research Ethics Service on 17th June 2011 (Ref 11/NW/0382) and extended on 10th May 2016 (Ref 16/NW/0274). Informed consent was obtained from all participants. The UKB resource is available for researchers to use for health-related research in the public interest. The rationale behind the UKB imaging study is explained in Petersen et al. (2013) [25], and the CMR acquisition protocol is detailed in Petersen et al. (2015)[26].

We performed our experiments on train/test splits from 4525 UKB subjects. To ensure fair comparison with previous work and facilitate reproducibility of our results, we used the same train/test splits as Xia et al. [35], in which 600 subjects were reserved as a separate test split. This allowed us to consistently evaluate the model performance and compare with previous studies.

Image and mesh pre-processing

CMR images were pre-processed by normalizing intensities to the range [0, 1]. SAX images had dimensions ranging from (100, 100, 6) to (200, 200, 16) and voxel spacing of [1.82, 1.82, 10] mm, while LAX images had varying dimensions depending on the associated SAX image. To handle varying size of SAX images across subjects, we evaluated our model under two settings: (1) *Full image input*, where we padded all SAX images to (210, 210, 16), and (2) *Cropped input*, where we followed prior work [1, 35] and cropped SAX images to (100, 100, 16), padding slices as needed. In all cases, LAX images were zero-padded to have a square shape of size (224, 224).

Inspired by classic object detection approaches, we align the vertex positions of the mesh with their relative position inside the SAX image, which has been shown to be effective when using graph generative models for landmark detection [10]. We first remove the origin of the SAX image and divide each direction by the corresponding voxel spacing to obtain positions in voxel space. For the full-image pipeline, we then add the padding applied to the positions and divide by the image size. For the cropped-image pipeline, we subtract the origin of the bounding box and divide by the image size. With this, we obtain a "relative positional space", that we use for training the models, with a value of (0.5, 0.5, 0.5) indicating a node in the center of the SAX image. To evaluate the results, we reverse this operation and recover the original positions in millimeters.

Data augmentation

All models were trained using online data augmentation, including intensity augmentation, random rotations of the SAX images (between -10 and 10 degrees), and random scaling on the x and y axes. The LAX images were scaled to match the scaling performed in the associated SAX image using each LAX image’s respective direction vector. For the cropped model, we added a step to randomly choose the cropping center, ensuring that the entire heart is always inside the region. This helps the model to avoid dependence on a perfectly centered crop and also acts as an extra data augmentation step.

3.2 Model implementation and training details

All models were implemented in Python using the PyTorch framework [23]. The PyTorch Geometric library [8] was employed for spectral Graph Convolutional Neural Network (GCNN) layers. Hyperparameters were selected via grid search, with a k -hop neighborhood parameter [6] set to 6. We conducted training for 600 epochs using the Adam optimizer with a learning rate of 10^{-4} . Batch size was set to 4, and weight decay was applied at 10^{-5} . A KL divergence weight factor of $\lambda_{KL} = 10^{-5}$ was introduced, and learning rate decay with a factor of 0.99 occurred after each epoch. Both 2D and 3D Convolutional Neural Network (CNN) encoders consisted of six residual blocks [12]. In 2D encoders, maxpooling layers were interleaved with these blocks. In 3D encoders, maxpooling was applied on the X and Y axes between each residual block, with Z-axis maxpooling at the third layer. The latent representations were obtained by using fully connected layers in the encoders, with a dimension of 32 for the 3D encoder and 8 for all 2D encoders, after a grid search hyper-parameter selection. GCNN decoders, in both 2D and 3D models, comprised six layers of Chebyshev convolutions with Layer Normalization [2] and ReLU nonlinearities. Classic surface regularization losses from the PyTorch3D library [29] were utilized. These losses included edge length, normal vector, and Laplacian regularization terms.

3.3 Model comparison

We implemented different single and multi-view variants of the HybridVNet architecture. We also compared our approach to the results obtained by the Multi-Cue Shape Inference Network (MCSI-Net) [35] for the ventricle dense segmentation task, which constitutes the state-of-the-art for point distribution models in this particular dataset. MCSI-Net combines two different networks. The first one is a position inference network that predicts the central coordinates of the mesh and a rotation vector. The second one is a shape inference network that uses CNN layers to infer the parameters of a point distribution model (PDM) based on PCA. This model uses the same SAX and multiple LAX views as our model, but also incorporates patient metadata information to the PDM learning process, while our model does not require patient metadata.

4 Results and Discussion

We conducted a comprehensive series of experiments to evaluate the performance of the proposed HybridVNet model alongside baseline models and its own various configurations. These experiments covered both surface mesh and tetrahedral volumetric mesh scenarios, including a sensitivity analysis of the proposed regularization losses. All evaluations were carried out on the same dataset comprising 600 subjects, as presented in Xia et al. 2022 [35], with respect to the ground truth meshes associated with this dataset.

4.1 Surface mesh extraction

To evaluate the quality of the cardiac meshes, we used mesh (2) and mask based (Table 1) metrics. First, to enable a direct comparison with the state-of-the-art Multi-Cue Shape Inference Network (MCSI-Net) [35], which was evaluated directly on the segmentation masks generated by the model in

Subpart	Metric	Full SAX Image		Cropped SAX Image	
		HybridVNet	MV-HybridVNet	HybridVNet	MV-HybridVNet
Full Mesh	MAE ↓	2.56 (0.62)	2.26 (0.55)	2.43 (0.59)	2.18 (0.54)
	MSE ↓	12.20 (7.11)	9.29 (5.48)	11.27 (6.69)	8.80 (5.31)
	RMSE ↓	3.38 (0.89)	2.95 (0.76)	3.25 (0.86)	2.87 (0.76)
LV	MAE ↓	1.90 (0.57)	1.79 (0.55)	1.75 (0.54)	1.70 (0.54)
	MSE ↓	6.23 (4.28)	5.60 (4.03)	5.35 (3.83)	5.11 (3.67)
	RMSE ↓	2.39 (0.73)	2.26 (0.71)	2.21 (0.70)	2.15 (0.70)
RV	MAE ↓	2.18 (0.64)	2.08 (0.60)	2.00 (0.58)	1.97 (0.59)
	MSE ↓	8.39 (5.64)	7.69 (4.93)	7.12 (4.84)	7.04 (4.72)
	RMSE ↓	2.78 (0.82)	2.66 (0.78)	2.56 (0.76)	2.54 (0.78)
LA	MAE ↓	2.90 (1.00)	2.37 (0.78)	2.84 (0.99)	2.30 (0.77)
	MSE ↓	15.40 (13.73)	10.07 (9.74)	14.88 (13.29)	9.58 (9.24)
	RMSE ↓	3.69 (1.33)	3.00 (1.02)	3.63 (1.31)	2.92 (1.02)
RA	MAE ↓	3.07 (0.96)	2.57 (0.76)	2.98 (0.93)	2.51 (0.80)
	MSE ↓	17.46 (13.65)	12.00 (9.42)	16.67 (13.13)	11.75 (10.16)
	RMSE ↓	3.97 (1.32)	3.30 (1.05)	3.87 (1.30)	3.24 (1.11)
AORTA	MAE ↓	2.66 (0.93)	2.37 (0.84)	2.56 (0.89)	2.34 (0.83)
	MSE ↓	13.17 (11.05)	10.24 (8.71)	12.38 (10.52)	10.04 (8.43)
	RMSE ↓	3.41 (1.23)	3.01 (1.09)	3.31 (1.20)	2.97 (1.09)

Table 2: Quantitative mesh evaluation results for surface meshes. An up arrow (↑) indicates that higher values are better, while a down arrow (↓) indicates that lower values are better.

the SAX image space, we derived dense segmentation masks from the surface meshes and evaluated classic segmentation metrics like Dice coefficient, Hausdorff distance, and the average distance between the reference and predicted contours at each slice, as summarized in Table 1.

In our initial comparison, we evaluated our HybridVNet against the SAX-only MCSI-Net, both with full images and cropped versions centered on the structure of interest (Table 1). Remarkably, HybridVNet outperforms the SAX MCSI-Net for all metrics and structures. Next, we compared our MV-HybridVNet with the standard MCSI-Net which also incorporates multiple-views and is the current state-of-the-art for this dataset. The results clearly demonstrate the superiority of our MV-HybridVNet, as it outperforms the standard MCSI-Net across all segmentation metrics for both the left and right ventricle segmentation tasks.

It is worth noting that our *full image* variant of the model achieves superior results compared to the baselines, all while eliminating the need for an extra step to detect the region of interest (ROI) during the segmentation process. Moreover, the MV-HybridVNet model on *cropped images* beats the results with significant differences with respect to the full image.

To account for structures that may not be visible in SAX images, and to provide more insight on how the incorporation of long-axis views in our model helps the model learn more details about the complete heart structure, we conducted a thorough evaluation of our proposed models directly on various subparts of the output mesh. Standard mesh evaluation metrics, including vertex mean squared error (MSE), mean average error (MAE), and root mean squared error (RMSE) in millimeters, were employed. Table 2 summarizes the results in our models, comparing the HybridVNet with its multi-view version, for *cropped images* and *full images* versions independently. The evaluation was performed at the nodes of the left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), and aorta.

Comparing the performance of the HybridVNet with and without the inclusion of LAX images, we observe a significant improvement in accuracy for all parts of the mesh. This improvement is particularly pronounced for the left and right atria (LA and RA) and the aorta, structures that are not fully visible in SAX images. The base HybridVNet model demonstrates the ability to approximate the positions of these structures, with further refinement achieved through the integration of LAX

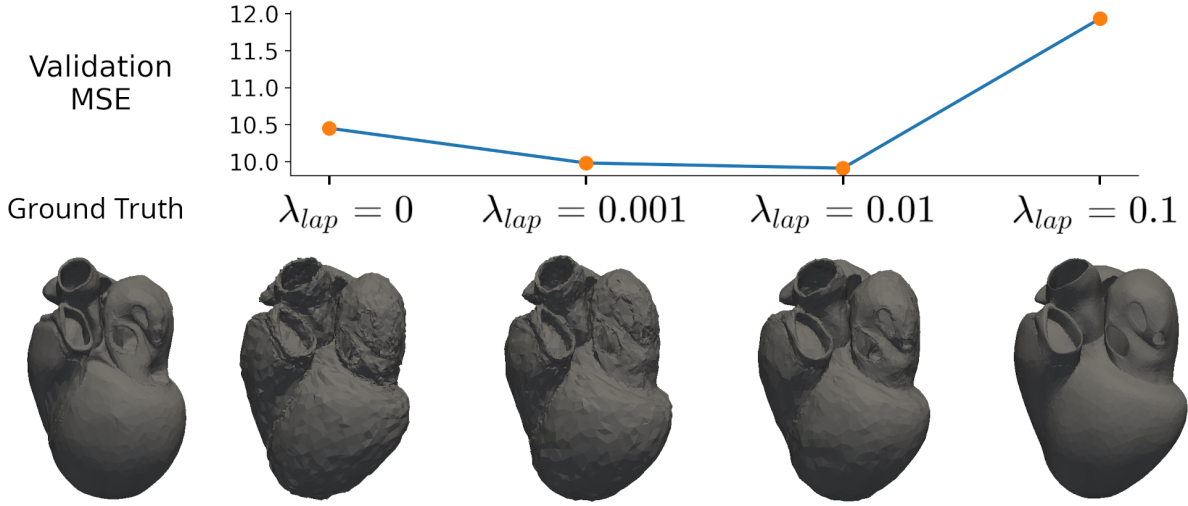


Figure 2: This qualitative analysis explores the impact of the Laplacian regularization term on surface mesh smoothness. It demonstrates the influence of adjusting the regularization parameter on mesh quality. Notably, the best quantitative results in terms of MSE for the validation split were achieved when $\lambda_{lap} = 0.01$.

images.

Surface mesh regularization effect

In the context of the surface mesh experiment, we undertook a comprehensive evaluation of various surface regularization loss functions to enhance the performance of our HybridVNet model. Specifically, we investigated the efficacy of three distinct regularization approaches: normal regularization, edge length regularization, and Laplacian smoothing. For more information about these regularizers see [34].

It is noteworthy that normal regularization and edge length regularization, while commonly employed in mesh regularization tasks, did not yield significant improvements in our model’s performance. This observation aligns with the intuitive understanding that these metrics are better suited for meshes with varying node counts and highly irregular target shapes, which is not the case in our dataset. In contrast, the incorporation of Laplacian smoothing produced notably smoother surface meshes. This can be visually appreciated in Figure 2, which presents a qualitative analysis of the obtained meshes as the regularization parameter for the Laplacian regularization loss was increased. The figures clearly illustrate the enhanced smoothness and quality of the meshes as the regularization strength is adjusted.

To quantitatively assess the impact of different loss terms during the training process, we refer to Figure 3. This figure provides a comparison of the MSE values throughout both training and validation phases. Notably, due to the resource-intensive nature of the validation process, we adjusted the intervals when recording loss values, with smaller intervals as more training time elapsed.

Significantly, the red curve in Figure 3 illustrates that the best performance is achieved when combining both deep supervision and Laplacian regularization losses. This combination not only eases the training process but also leads to improved model performance. The optimal regularization strength for Laplacian smoothing, resulting in the best performance in terms of mean squared error (MSE) for both the full image and cropped models, was determined to be $\lambda_{lap} = 0.01$. This finding was consistent with both qualitative and quantitative evaluations, as over-smoothed meshes appeared when using high values of the regularization term.

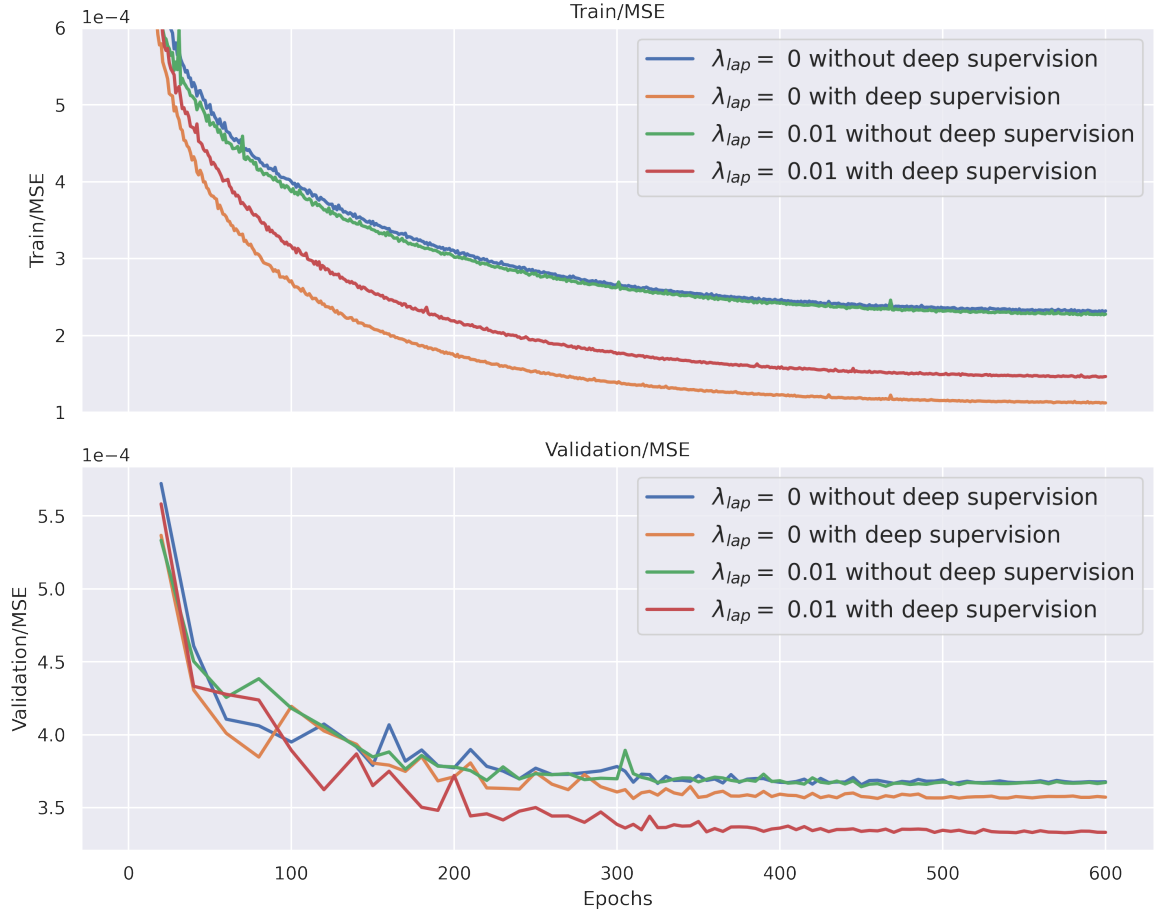


Figure 3: MSE values over the course of training and validation for different configurations of hyperparameters, measured in the "relative positional space". The red curve highlights the significant impact of combining Deep Supervision and Laplacian regularization losses on model performance. Smaller intervals were used for loss recording as training progressed.

4.2 Towards simulation-ready tetrahedral meshes

Our second experiment focused on the creation of simulation-ready tetrahedral meshes. We evaluated various weighting factors (λ_{ter}) for the tetrahedral element regularization term defined in equation 4, aiming to understand its influence on both mesh quality and ventricle segmentation performance. Table 3 presents the results, including metrics for mesh quality and ventricle segmentation, across different λ_{ter} values.

Our exploration reveals a nuanced relationship between λ_{ter} and model performance. Notably, for segmentation metrics, the finest outcomes emerge when $\lambda_{ter} = 1e - 4$. In this configuration, results closely resemble those of the non-regularized model, especially in terms of LV Endo metrics. However, for metrics related to mesh prediction performance, the optimal choice shifts slightly, with $\lambda_{ter} = 1e - 3$ yielding the best outcomes. While this setting leads to a minor dip in ventricle segmentation performance, it significantly reduces the mesh error.

A closer examination of the training dynamics, as illustrated in Figure 4, reinforces the benefits of using small λ_{ter} values. These values result in improved validation performance without substantial fluctuations in the training curves. Conversely, the highest regularization strength ($\lambda_{ter} = 1e - 2$)

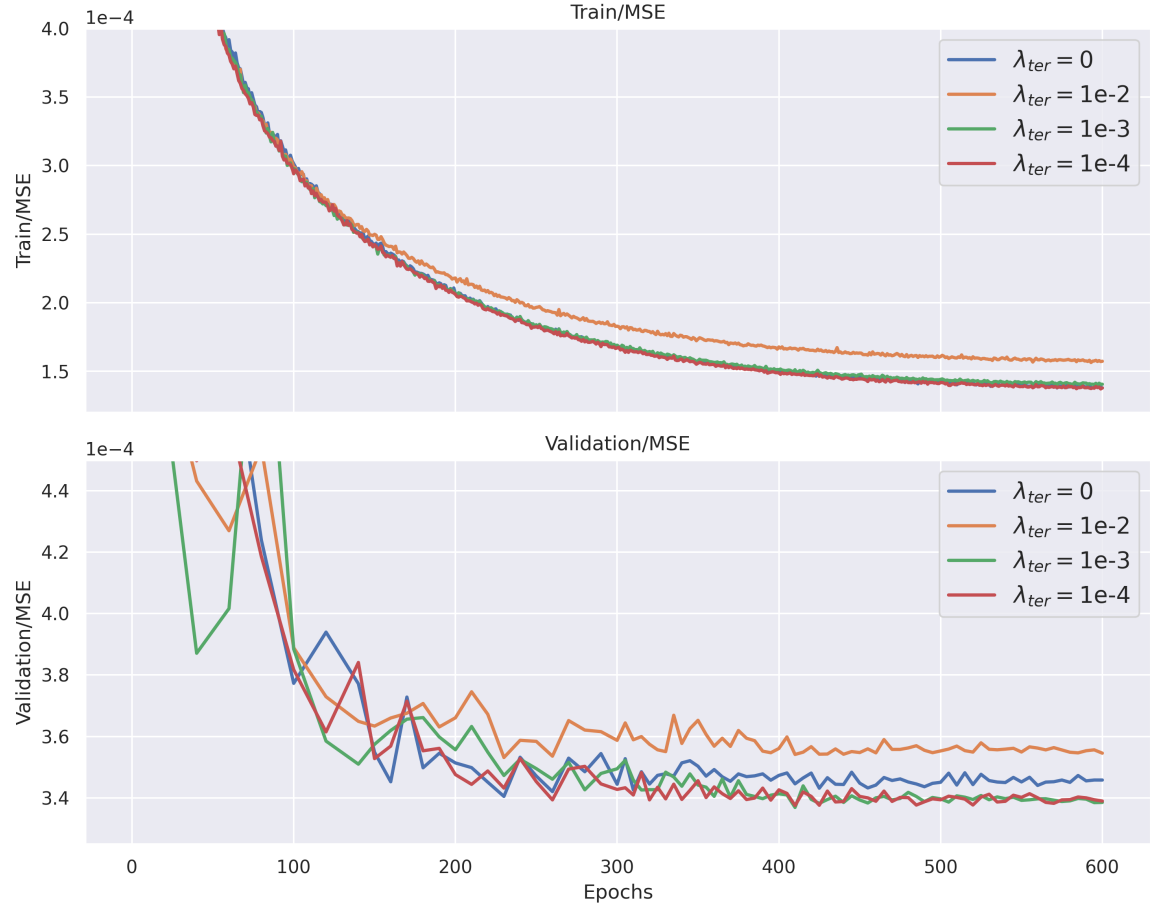


Figure 4: MSE values over the course of training and validation for volumetric meshes, exploring different configurations of λ_{ter} , with values measured in the "relative positional space". The red and green curves highlight an increase of performance when using $\lambda_{ter} = 1e - 4$ (Red) and $\lambda_{ter} = 1e - 3$ (Green). Noticeably, $\lambda_{ter} = 1e - 2$ (Yellow) shows both a high decay of performance for both train and validation curves. Smaller intervals were used for loss recording as training progressed.

	Metrics	MV-HybridVNet			
		$\lambda_{ter} = 0$	$\lambda_{ter} = 1e - 4$	$\lambda_{ter} = 1e - 3$	$\lambda_{ter} = 1e - 2$
Mesh	MAE ↓	2.08 (0.63)	2.07 (0.64)	2.04 (0.61)	2.11 (0.61)
	MSE ↓	8.25 (6.14)	8.22 (6.12)	7.93 (5.63)	8.39 (6.00)
	RMSE ↓	2.74 (0.88)	2.73 (0.88)	2.69 (0.84)	2.77 (0.84)
LV Endo	DC ↑	0.90 (0.04)	0.90 (0.04)	0.90 (0.05)	0.88 (0.05)
	HD ↓	4.36 (1.22)	4.32 (1.24)	4.41 (1.35)	5.21 (1.42)
	MCD ↓	1.52 (0.46)	1.51 (0.49)	1.58 (0.54)	1.89 (0.62)
LV Myo	DC ↑	0.78 (0.04)	0.78 (0.04)	0.76 (0.05)	0.74 (0.06)
	HD ↓	5.27 (1.47)	4.98 (1.40)	5.17 (1.50)	5.30 (1.57)
	MCD ↓	1.86 (0.61)	1.81 (0.64)	1.95 (0.72)	1.96 (0.77)
RV Endo	DC ↑	0.85 (0.06)	0.86 (0.05)	0.85 (0.05)	0.85 (0.06)
	HD ↓	7.22 (2.76)	6.97 (2.54)	7.38 (2.67)	7.55 (2.80)
	MCD ↓	2.05 (0.64)	2.02 (0.63)	2.09 (0.64)	2.13 (0.69)

Table 3: Quantitative results for segmentation metrics in volumetric meshes. (↑) indicates that higher results are better, while (↓) indicates that lower results are better. Bold results in two columns indicate there are no significant differences between these two.

		mean	std	min	max	1%	5%	25%	50%	75%
Reference Meshes	Atlas	0.491	0.174	0.092	0.984	0.115	0.194	0.367	0.494	0.617
	Ground Truth	0.355	0.156	-0.207	0.838	0.04	0.103	0.238	0.353	0.47
	Simpleware	0.524	0.185	0.064	0.992	0.128	0.202	0.387	0.535	0.667
MV-HybridGNet3D	$\lambda_{ter} = 0$	0.222	0.225	-0.759	0.876	-0.327	-0.144	0.065	0.219	0.384
	$\lambda_{ter} = 1e - 4$	0.229	0.23	-0.771	0.871	-0.337	-0.151	0.068	0.231	0.397
	$\lambda_{ter} = 1e - 3$	0.433	0.206	-0.719	0.904	-0.138	0.059	0.307	0.457	0.585
	$\lambda_{ter} = 1e - 2$	0.501	0.309	-0.931	0.943	-0.681	-0.298	0.434	0.577	0.688

Table 4: Quantitative results for the quality of elements in volumetric meshes. The values correspond to the scaled Jacobian, and higher values imply a better quality of the tetrahedra.

leads to decreased performance in both training and validation.

Note that the impact of the regularization term on the reported ventricle segmentation metrics is not drastic. However, as demonstrated in Figure 5 and elaborated upon in the following paragraphs, it substantially elevates the quality of tetrahedral elements, a critical consideration for simulations.

To validate the quality of tetrahedral elements, we used the widely adopted scaled Jacobian metric. The Jacobian of a tetrahedron is a matrix that describes how the tetrahedron’s shape changes under deformation. The scaled Jacobian is a quantitative measure of regularity and symmetry, falling within the range $[-1, 1]$ and unaffected by scale or units. A high scaled Jacobian value implies high regularity, low distortion, and therefore high quality [13]. Table 4 provides a comprehensive overview of element qualities under different λ_{ter} conditions. The table includes statistics such as average, standard deviation, minimum, maximum, and various percentiles of tetrahedron quality across all test subjects. We compare the results with volumetric atlases, ground-truth meshes, our approaches, and a subset of surface meshes converted to volumetric meshes using ScanIP [32].

Importantly, our findings demonstrate that our regularized models surpass ground-truth elements in terms of quality, beginning from the 25% quartile and onwards, for $\lambda_{ter} = 1e - 3$ and higher. This observation underscores our hypothesis that the regularization loss significantly enhances mesh quality. Figure 5 visually encapsulates this improvement, positioning our method competitively with Simpleware meshes, with the exception of a small number of elements, potentially due to the original low quality of the ground truth.

In comparison to the conventional approach of directly converting surface to volumetric meshes, our

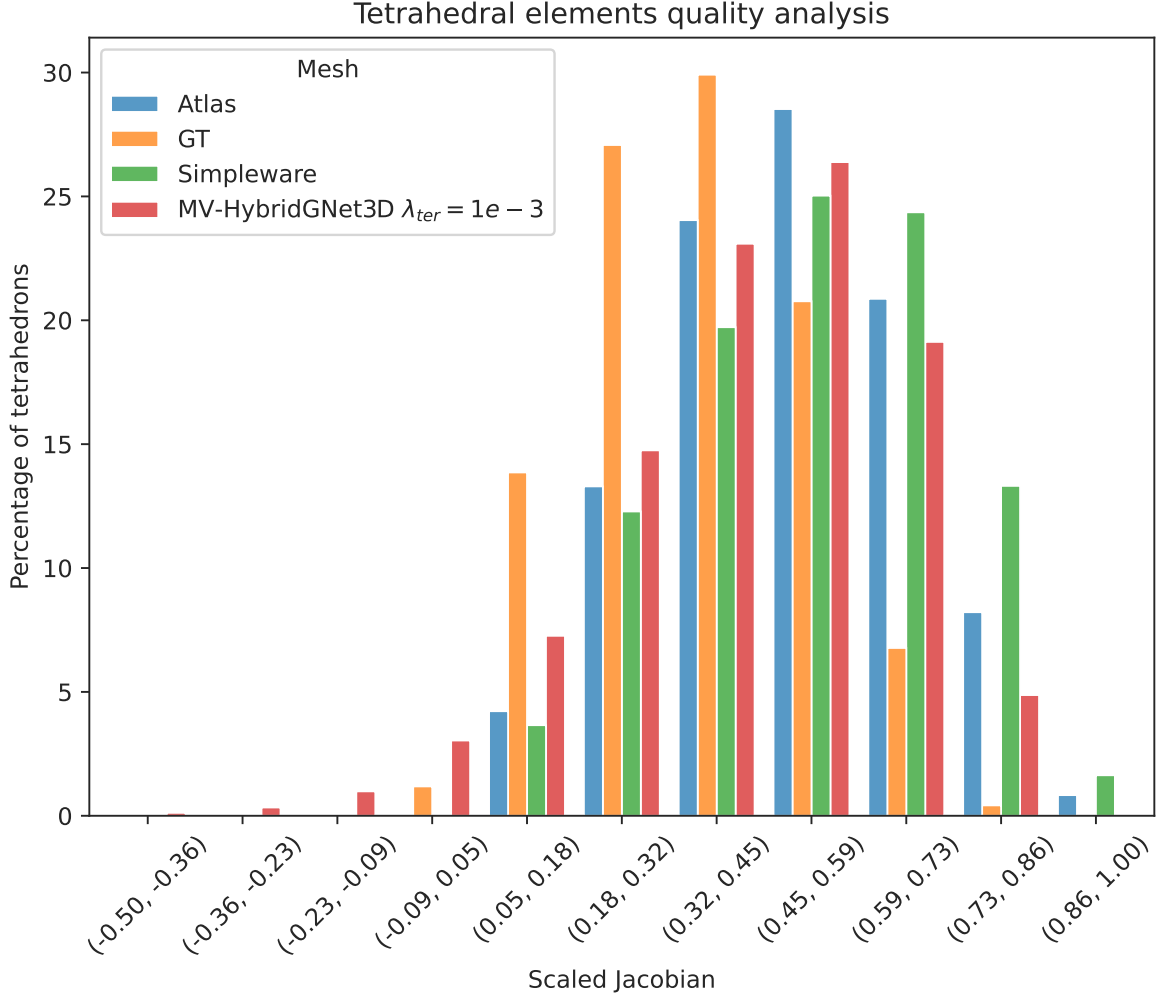


Figure 5: Histogram of tetrahedral mesh quality using scaled Jacobian values. The x-axis represents the scaled Jacobian values and the y-axis shows the percentage of tetrahedral elements within each range.

model demonstrates competitive results. Moreover, it addresses a challenge posed by direct conversion, where degenerate triangles can obstruct volumetric mesh creation, affecting approximately 10% of cases in our experiments. Additionally, ScanIP’s procedure, a typical alternative, consumes about 6 minutes per mesh on average. In contrast, our approach can generate the vertex set for a volumetric mesh in just around 0.04 seconds per set of CMR images when running on a NVIDIA A100-SXM4 GPU, resulting in a remarkable speed-up.

5 Conclusions

In this work we propose HybridVNet, a novel method for generating both surface and simulation-ready tetrahedral meshes directly from images. Our approach outperforms the state-of-the-art on multiple standard mesh evaluation metrics, demonstrating its potential for efficiently generating high-quality surface and tetrahedral meshes. For future work we consider extending HybridVNet to other organs and image modalities. So far, our model was only applied to CMR images. However, its generic formulation

makes it simple to adapt it for other anatomical structures, like cortical surface reconstruction from brain MR images, which are left as future work. Additionally, future work will focus on improving the quality of the volumetric atlas used as ground truth, to more accurately evaluate the potential of our method for use in in-silico simulation-based studies.

References

- [1] Attar, R., et al.: Quantitative CMR population imaging on 20,000 subjects of the UK biobank imaging study: Lv/rv quantification pipeline and its evaluation. *Medical image analysis* **56**, 26–42 (2019)
- [2] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
- [3] Bai, W., Shi, W., Ledig, C., Rueckert, D.: Multi-atlas segmentation with augmented features for cardiac mr images. *Medical image analysis* **19**(1), 98–109 (2015)
- [4] Beetz, M., Acero, J.C., Banerjee, A., Eitel, I., Zacur, E., Lange, T., Stiermaier, T., Evertz, R., Backhaus, S.J., Thiele, H., Bueno-Orovio, A., Lamata, P., Schuster, A., Grau, V.: Mesh u-nets for 3d cardiac deformation modeling. In: Camara, O., Puyol-Antón, E., Qin, C., Sermesant, M., Suinesiaputra, A., Wang, S., Young, A. (eds.) *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers*. pp. 245–257. Springer Nature Switzerland, Cham (2022)
- [5] Bonazzola, R., Ravikumar, N., Attar, R., Ferrante, E., Syeda-Mahmood, T., Frangi, A.F.: Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in CMR images. In: *MICCAI*. pp. 699–708. Springer (2021)
- [6] Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. *arXiv preprint arXiv:1606.09375* (2016)
- [7] Fedele, M., Quarteroni, A.: Polygonal surface processing and mesh generation tools for the numerical simulation of the cardiac function. *International Journal for Numerical Methods in Biomedical Engineering* **37**(4), e3435 (2021)
- [8] Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019)
- [9] Gaggion, N., Mansilla, L., Milone, D.H., Ferrante, E.: Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In: *MICCAI*. Springer International Publishing (2021)
- [10] Gaggion, N., Mansilla, L., Mosquera, C., Milone, D.H., Ferrante, E.: Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Medical Imaging* (2022). <https://doi.org/10.1109/tmi.2022.3224660>
- [11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778 (2016)
- [12] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770–778 (2016). <https://doi.org/10.1109/CVPR.2016.90>
- [13] Johnen, A., Geuzaine, C., Toulorge, T., Remacle, J.F.: Efficient computation of the minimum of shape quality measures on curvilinear finite elements. *Computer-Aided Design* **103**, 24–33 (2018)

- [14] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
- [15] Kong, F., Shadden, S.C.: Whole heart mesh generation for image-based computational simulations by learning free-from deformations. In: MICCAI. pp. 550–559. Springer (2021)
- [16] Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. IEEE TMI (2020)
- [17] Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. pp. 562–570 (2015)
- [18] Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. IEEE (2016)
- [19] Musy, M., others.: vedo: 2022.4.1 (Oct 2022). <https://doi.org/10.5281/zenodo.7222019>
- [20] Neic, A., Gsell, M.A., Karabelas, E., Prassl, A.J., Plank, G.: Automating image-based mesh generation and manipulation tasks in cardiac modeling workflows using meshtool. SoftwareX **11**, 100454 (2020)
- [21] Ordas, S., Oubel, E., Leta, R., Carreras, F., Frangi, A.F.: A statistical shape model of the heart and its application to model-based segmentation. In: Medical Imaging 2007: Physiology, Function, and Structure from Medical Images. vol. 6511, pp. 490–500. SPIE (2007)
- [22] Pak, D.H., Liu, M., Kim, T., Liang, L., McKay, R., Sun, W., Duncan, J.S.: Distortion energy for deep learning-based volumetric finite element mesh generation for aortic valves. In: MICCAI. pp. 485–494. Springer International Publishing (2021)
- [23] Paszke, A., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (2019)
- [24] Petersen, S.E., Aung, N., Sanghvi, M.M., Zemrak, F., Fung, K., Paiva, J.M., Francis, J.M., Khanji, M.Y., Lukaschuk, E., Lee, A.M., et al.: Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (cmr) in caucasians from the uk biobank population cohort. Journal of cardiovascular magnetic resonance **19**(1), 1–19 (2017)
- [25] Petersen, S.E., et al.: Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. Journal of Cardiovascular Magnetic Resonance **15**(1), 46 (Dec 2013)
- [26] Petersen, S.E., et al.: UK biobank’s cardiovascular magnetic resonance protocol. Journal of cardiovascular magnetic resonance **18**(1), 1–7 (2015)
- [27] Puyol-Anton, E., et al.: A multimodal spatiotemporal cardiac motion atlas from mr and ultrasound data. Medical image analysis **40**, 96–110 (2017)
- [28] Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3D faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 704–720 (2018)
- [29] Ravi, N., et al.: Accelerating 3D deep learning with pytorch3d. arXiv:2007.08501 (2020)
- [30] Rodero, C., et al.: Linking statistical shape models and simulated function in the healthy adult human heart. PLOS Computational Biology **17**(4), 1–28 (04 2021). <https://doi.org/10.1371/journal.pcbi.1008851>

- [31] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
- [32] Synopsis: Simpleware (2021), <https://www.synopsys.com/simpleware.html>, Last opened on 2022-07-11
- [33] Tóthová, K., Parisot, S., Lee, M., Puyol-Antón, E., King, A., Pollefeys, M., Konukoglu, E.: Probabilistic 3D surface reconstruction from sparse MRI information. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 813–823. Springer (2020)
- [34] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3D mesh models from single rgb images. In: Proceedings of the European conference on computer vision (ECCV). pp. 52–67 (2018)
- [35] Xia, Y., et al.: Automatic 3D+t four-chamber CMR quantification of the UK biobank: integrating imaging and non-imaging data priors at scale. *Medical Image Analysis* **80**, 102498 (2022). <https://doi.org/https://doi.org/10.1016/j.media.2022.102498>

Multi-center anatomical segmentation with heterogeneous labels via landmark-based models

Multi-center anatomical segmentation with heterogeneous labels via landmark-based models

Nicolás Gaggion¹, Maria Vakalopoulou², Diego H. Milone¹, and Enzo Ferrante¹

¹Research Institute for Signals, Systems and Computational Intelligence, sinc(i),
CONICET-UNL, Argentina

²MICS, CentraleSupélec, Université Paris-Saclay, Inria Saclay, France

Abstract

Learning anatomical segmentation from heterogeneous labels in multi-center datasets is a common situation encountered in clinical scenarios, where certain anatomical structures are only annotated in images coming from particular medical centers, but not in the full database. Here we first show how state-of-the-art pixel-level segmentation models fail in naively learning this task due to domain memorization issues and conflicting labels. We then propose to adopt HybridGNet, a landmark-based segmentation model which learns the available anatomical structures using graph-based representations. By analyzing the latent space learned by both models, we show that HybridGNet naturally learns more domain-invariant feature representations, and provide empirical evidence in the context of chest X-ray multiclass segmentation. We hope these insights will shed light on the training of deep learning models with heterogeneous labels from public and multi-center datasets.

1 Introduction

Anatomical segmentation is one of the pillar problems in medical image analysis, required by several downstream tasks like radiotherapy planning [1] or shape variability analysis in computational anatomy [2]. Fully convolutional neural networks such as UNet [3], and its self-configuring variant nnUNet [4], have become the state-of-the-art for this task. In this work, we are interested in addressing two common situations encountered when training anatomical segmentation models in real clinical scenarios: multi-center image databases and heterogeneous labels. On one hand, multi-center databases may lead to domain shift problems [5] due to changes in intensity distribution caused by differences in acquisition device or protocol parameters. On the other hand, heterogeneous or missing labels [6] make it difficult to train a single segmentation model for all regions of interest (ROIs), as missing labels in different images may send contradictory training signals. Notably, when we face both issues at the same time, the problem is far from trivial as it lies in the intersection of multi-task learning, domain adaptation and weakly supervised learning [7]. As we will show in this work, when different organs are annotated in images coming from various centers, commonly used pixel-level segmentation methods like UNet and nnUNet trained with standard procedures tend to associate certain labels to specific domains.

Several methods have been proposed to independently address the problems of domain shift [5, 8, 9] and heterogeneous labels [10, 6, 11] in medical image segmentation. As for the joint problem, Dorent and coworkers [7] proposed a framework which combines a variational formulation to cope with heterogeneous labels, with conventional techniques based on data augmentation, adversarial learning, and pseudo-healthy image generation to address domain shift. In this work, we argue that landmark based segmentation methods like the HybridGNet [12, 13] can naturally handle these scenarios, as they incorporate prior knowledge about the expected anatomy, without additional burden related to

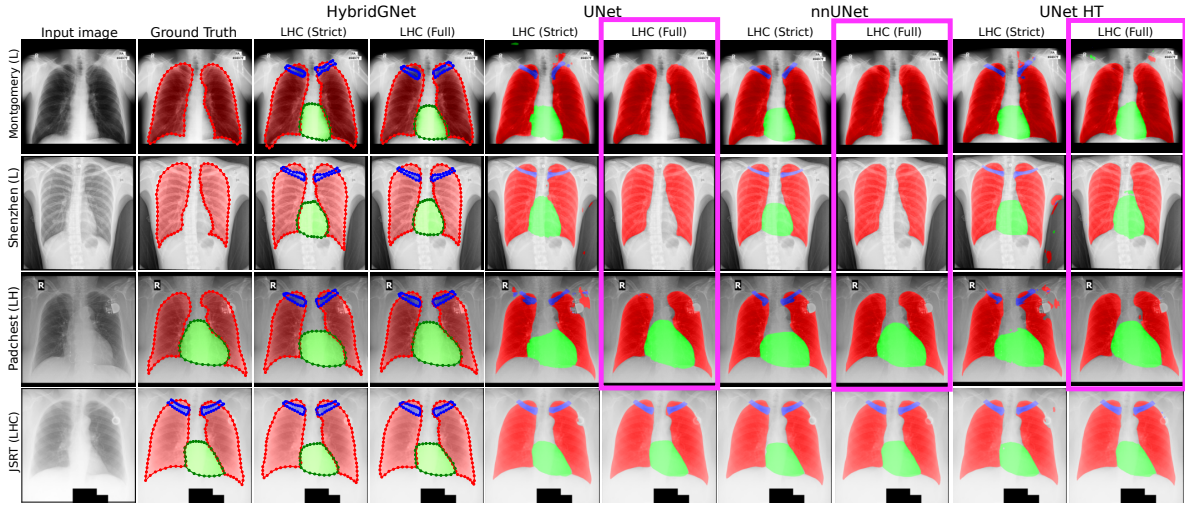


Figure 1: **Qualitative examples.** Examples from 4 datasets (rows) segmented with different methods and annotation settings (columns). Note how pixel UNet and nnUNet (in pink) fail completely to segment structures that are not presented in the corresponding dataset when trained with heterogeneous labels (i.e. LHC (Full)). It is not the case for the HybridGNet, which always provides segmentations of all structures. Note also that the UNet HT (trained using the same heterogeneous setup as HybridGNet) produces heart segmentation for all datasets, as there are no conflicting labels. However, it only segments clavicles for JSRT, due to the conflicting annotations in the overlapping area between lung and clavicles (see Section 3 for more details).

data augmentation, adversarial training, or image generation. We first provide empirical evidence showing how widely used pixel-level approaches drastically fail to learn robust segmentation models using heterogeneous labels from multicentric datasets, while HybridGNet can naturally handle this problem, avoiding memorization issues. Further analysis of the latent spaces learned by the different architectures, indicates that generative landmark-based approaches like HybridGNet tend to learn more invariant representations, which helps to improve the robustness with respect to domain memorization.

2 Methods and experiments

Problem statement and experimental setup: We explore anatomical segmentation of lung, heart and clavicles in a multi-center database of chest X-ray images, with heterogeneous labels. The database is composed of 4 publicly available datasets (JSRT [14], Padchest [15], Montgomery [16] and Shenzhen [17]), which originally provide pixel level annotations. Since the proposed method employs landmark-based annotations, we adopted the publicly available *Chest X-ray landmark dataset*, which provides landmarks for 3 different organs from the aforementioned databases (github.com/ngaggion/Chest-xray-landmark-dataset). Images from JSRT (246 subjects) include annotations for lungs, heart and clavicle (LHC); Padchest (137 subjects) include lungs and heart (LH); while Montgomery (138 subjects) and Shenzhen (390 subjects) include only lung (L). To evaluate each domain separately, we divide the datasets into 80% train/val and 20% test partitions.

HybridGNet: HybridGNet is a landmark-based segmentation model, where the ROIs are encoded as anatomical graphs representing the organ contour. It follows an encoder-decoder architecture that combines standard convolutions for image encodings, with graph generative models to extract anatomically plausible representations directly from images. The model is trained to minimize the mean squared error (MSE) between the predicted node positions and the ground truth coordinates. Pixel-level masks

Model	Trained in	Montgomery			Shenzhen			Padchest						JSRT					
		Lungs			Lungs			Lungs			Heart			Lungs			Heart		
		MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD
HybridGNet	L (Both)	128.4	0.97	27.4	142.9	0.97	32.5	152.1	0.96	36.6	-	-	-	120.2	0.97	33.0	-	-	-
	LH (Strict)	295.6	0.95	38.5	264.4	0.95	43.1	201.8	0.95	41.0	351.7	0.94	36.8	155.7	0.97	36.3	382.4	0.94	34.5
	LH (Full)	104.4	0.97	27.8	145.9	0.97	32.8	187.0	0.96	38.8	342.0	0.94	36.0	124.8	0.97	31.0	375.8	0.94	33.8
	LHC (Strict)	571.5	0.94	49.2	486.0	0.93	51.1	480.2	0.92	56.8	1317.0	0.87	70.9	139.4	0.97	33.9	413.1	0.93	35.7
UNet	LHC (Full)	110.7	0.97	26.2	148.7	0.96	33.2	172.0	0.96	37.1	235.0	0.94	30.5	122.2	0.97	31.1	390.8	0.94	34.2
	L (Both)	-	0.98	53.3	-	0.98	53.3	-	0.96	85.2	-	-	-	-	0.95	99.9	-	-	-
	LH (Strict)	-	0.96	73.4	-	0.96	131.2	-	0.96	89.4	-	0.93	80.5	-	0.95	101.1	-	0.94	51.7
	LH (Full)	-	0.97	70.0	-	0.96	74.0	-	0.96	94.8	-	0.89	75.2	-	0.95	103.8	-	0.94	51.8
nnUNet	LHC (Strict)	-	0.90	182.4	-	0.91	217.6	-	0.89	227.1	-	0.87	220.8	-	0.97	67.5	-	0.94	67.6
	LHC (Full)	-	0.97	72.5	-	0.97	72.7	-	0.96	106.1	-	0.91	76.7	-	0.98	63.0	-	0.94	54.5
	L (Both)	-	0.98	26.1	-	0.97	32.3	-	0.96	40.2	-	-	-	-	0.98	31.1	-	-	-
	LH (Strict)	-	0.97	45.1	-	0.96	53.3	-	0.96	40.5	-	0.95	32.9	-	0.98	26.3	-	0.95	29.0
UNet HT	LH (Full)	-	0.98	34.0	-	0.97	35.2	-	0.96	37.4	-	0.94	34.5	-	0.98	28.6	-	0.95	30.9
	LHC (Strict)	-	0.93	120.5	-	0.92	121.0	-	0.93	83.6	-	0.89	80.4	-	0.98	35.0	-	0.95	29.5
	LHC (Full)	-	0.98	25.3	-	0.97	35.2	-	0.96	39.3	-	0.95	33.3	-	0.98	35.6	-	0.95	31.1
	L (Both)	-	0.97	46.9	-	0.97	78.7	-	0.96	70.0	-	-	-	-	0.95	106.1	-	-	-
UNet HT	LH (Strict)	-	0.96	74.8	-	0.96	131.6	-	0.96	74.9	-	0.94	80.2	-	0.95	105.6	-	0.94	58.4
	LH (Full)	-	0.98	60.5	-	0.97	57.8	-	0.96	87.9	-	0.93	125.6	-	0.95	100.2	-	0.94	59.4
	LHC (Strict)	-	0.91	168.5	-	0.91	204.8	-	0.90	223.6	-	0.87	199.0	-	0.98	78.1	-	0.94	82.2
	LHC (Full)	-	0.97	72.6	-	0.97	77.1	-	0.96	66.1	-	0.93	87.8	-	0.98	55.2	-	0.94	47.6

Table 1: **Quantitative results for both landmark and pixel-based baselines:** Results show an increase in performance when combining heterogeneous labels (Full) compared to those cases where only images with all the required annotations (Strict) are available. **Blue:** images from the target dataset are present at training time. **Red:** images from the target dataset are not present during training. **Green:** heterogeneous setting, all datasets are present during training.

are then generated by filling in the contours. See [12, 13] for more details about HybridGNet.

Training landmark-based models with heterogeneous labels: HybridGNet provides a natural way to learn with heterogeneous labels, which only relies on indexation. The model outputs a $D \times 2$ matrix, where the number of nodes D is fixed and sequentially ordered: first lung nodes, t_L , then heart nodes, t_H , and finally clavicles nodes, t_C , as in the ground-truth $target = [t_L, t_H, t_C]$. The length of every subset is D_L , D_H , D_C , respectively. For training, batches composed of images *from a single database at a time* are randomly chosen at every iteration, thus constraining the type of annotation to those available for that dataset. For example, if the input batch only includes lungs, the loss function for that gradient descent iteration is only evaluated for the first D_L nodes, and errors are not back-propagated for heart and clavicle. The same is done for the LH task, where the loss is evaluated in the first $D_L + D_H$ nodes, ignoring the rest of the output. This is implemented via slicing operations, and constitutes the only modification made to HybridGNet.

Baselines and heterogeneous UNet training:

We propose to compare the HybridGNet with two pixel-level segmentation models: a UNet [3] with residual convolutional blocks, trained with a compound cross entropy and soft Dice loss [18]; and a nnUNet [4] trained with its self-configuring method. We also propose to train the UNet in the same heterogeneous training (HT) setup as the HybridGNet for fair comparison. In UNet HT, each training batch contains a specific set of labels, and we avoid back-propagating the gradient loss for unseen labels in the batch. We implement this method simply treating each anatomical structure as an independent binary segmentation problem. The UNet HT model thus has one independent output feature map per anatomical structure, akin to a multi-label classification problem. We apply a sigmoid non-linearity to each output segmentation map. We use binary cross-entropy and a modified soft Dice loss function to allow using a single feature map, instead of the standard one-hot encoding used when classes are mutually exclusive. At test time, the sigmoid outputs are just thresholded at 0.5, obtaining an independent binary map for each structure.

Training details: As not all labels are available for each dataset, we took into account the label availability and devised two different training settings: **Strict** and **Full**. While **Strict** indicates that only datasets annotated with the particular listed labels were used, **Full** indicates that all datasets were used for training (resulting in an heterogeneous label setting). Thus, we trained models in the following settings:

- **Strict training setting:**
 - **LH:** Models were trained to predict lung and heart, using only images that had both lung and heart annotations (i.e. JSRT and Padchest datasets).
 - **LHC:** Models were trained to predict lung, heart and clavicles, using only images that had all annotations available (thus, just JSRT dataset).
- **Full training setting:**
 - **L:** Models were trained with all datasets to predict only lungs. This case can also be considered **Strict**, since all datasets have lung annotations available.
 - **LH:** Models were trained to predict lungs and heart using all datasets in an heterogeneous annotation setting, i.e. some images had only lung annotations (Montgomery and Shenzhen) and others had lung and heart (JSRT and Padchest).
 - **LHC:** Models were trained to predict lungs, heart and clavicles using all datasets in an heterogeneous setting, i.e. some images had only lung annotations (Montgomery and Shenzhen), some had lung and heart (Padchest), while some had the 3 structures (JSRT).

Artificial removal of labels: As we will discuss in the next section, our experiments show that naively trained pixel-level approaches learn to map anatomical structures to datasets where they were annotated, failing to segment them in datasets where these structures are not labeled. However, as we do not have ground-truth for these structures, we cannot show this quantitatively. To overcome this limitation and provide quantitative support for our claims, we took the two datasets with more than one annotated structure (JSRT and Padchest), and artificially removed one of the organs during training. This resulted in 4 different experiments, where we can compute quantitative results for labels that were not seen during training: removing lungs from JSRT (**Exp 1**), removing heart from JSRT (**Exp 2**), removing lungs from Padchest (**Exp 3**) and removing heart from Padchest (**Exp 4**).

3 Results, discussion and conclusions

Figure 1 shows one of our main findings: when trained with heterogeneous labels associated to different datasets (i.e. LH (Full) and LHC (Full)), naive pixel-based methods segment only those ROIs that were annotated in the corresponding dataset (see first two cases highlighted in pink). Meanwhile, the UNet HT model which is aware of the heterogeneous labels by ignoring annotations not present in every specific dataset, only segments classes that are not in conflict (i.e. we say a pixel class is in conflict if it is considered to be part of one class for a specific dataset, but also part of a different class in other datasets, like clavicles). Instead, the HybridGNet always predicts the complete set of anatomically plausible segmentations. This effect is present on all the samples for each dataset. UNet and nnUNet clearly memorize which structure was annotated in what dataset, and use the multi-centric distribution shift as a shortcut to decide what ROIs should be segmented in every test dataset. UNet HT overcomes this issue for the heart masks, as we make it aware of heterogeneous labels by ignoring classes that are not annotated in specific datasets, but still fails with the clavicles, as they conflict with lung labels. On the contrary, HybridGNet is forced to predict all ROIs by construction, using the anatomical priors encoded in the learned latent space to infer the organ position, even if it was not present in that particular dataset.

To better understand the reasons behind domain memorization, we performed dimensionality reduction on both the latent space of the HybridGNet model and the bottleneck features of the UNets. We analyzed the **LHC (Full)** models, and used UMAP[19] for dimensionality reduction. Figure 2 (Left column) shows the 2D projection of images from all datasets for each model. UNet and UNet HT clearly clusterize samples per dataset. Since JSRT images tend to have much bigger lung area than the other datasets, we also experimented scaling all images to have the same organ’s bounding box area, discarding the potential clustering effect associated to the lung size and not to the multi-center intensity shift. This is shown in Figure 2 (Right column), where the clustering was completely removed on the HybridGNet latent space, but it did not affect the UNet and UNet HT. nnUNet results could not



Figure 2: **Latent space inspection via UMAP embeddings:** Different datasets are shown in colors, allowing to see how both UNet and UNet HT models tend to clusterize images per dataset, while HybridGNet doesn't, explaining the improved robustness to domain-label memorization.

be obtained due to the encapsulation of the training and test framework, but are expected to behave similarly to the naive UNet since both models share the same underlying architecture. This clustering effect explains why memorization issues happen in UNet and UNet HT, but not in HybridGNet.

Table 1 shows quantitative results for the different training scenarios, models and test sets. In general, we see that incorporating the Full dataset through heterogeneous labels improve performance when compared to the Strict case, which only employs images where all annotations are available. Performance is only evaluated for ROIs available as ground-truth in every dataset. As shown in Figure

Model	Trained in	JSRT				Padchest			
		Lungs		Heart		Lungs		Heart	
		Dice	HD	Dice	HD	Dice	HD	Dice	HD
HybridGNet	Exp 1	0.931	47.2	0.941	31.2	0.952	37.8	0.942	31.4
	Exp 2	0.970	34.0	0.910	58.2	0.953	40.5	0.935	33.7
	Exp 3	0.970	34.3	0.941	32.3	0.905	56.3	0.946	31.4
	Exp 4	0.975	32.6	0.939	31.9	0.957	38.1	0.899	69.7
UNet HT	Exp 1	0.968	90.3	0.937	80.0	0.960	96.9	0.928	141.9
	Exp 2	0.980	57.3	0.860	340.6	0.960	69.6	0.923	116.6
	Exp 3	0.979	44.0	0.941	46.7	0.932	202.0	0.931	102.7
	Exp 4	0.979	53.4	0.941	66.7	0.960	63.3	0.872	182.5
UNet	Exp 1	0.034	-	0.937	56.5	0.956	59.3	0.935	82.3
	Exp 2	0.979	65.8	0.027	-	0.959	98.5	0.936	101.5
	Exp 3	0.977	56.1	0.938	71.4	0.035	-	0.921	153.7
	Exp 4	0.980	49.0	0.944	68.4	0.961	95.0	0.039	-
nnUNet	Exp 1	0.000	-	0.952	30.2	0.963	43.4	0.947	31.6
	Exp 2	0.980	33.4	0.000	-	0.965	38.8	0.945	37.8
	Exp 3	0.979	47.4	0.948	30.1	0.000	-	0.945	31.7
	Exp 4	0.981	32.1	0.951	29.6	0.964	38.2	0.000	-

Table 2: **Artificially removed labels experiment.** In red: label was not present during training. In blue: label was present during training.

1, for the heterogenous settings (**LHC (Full)** and **LH (Full)**) UNet and nnUNet drastically fail at segmenting structures that were not present during training in the corresponding dataset. This is due to the memorization issues and conflicting background/organ labels in different domains. This would imply a Dice of 0 for UNet and nnUNet if ground-truth were available for evaluation.

To quantify these results, we artificially removed labels from datasets with more than one annotated structure (JSRT and Padchest). Table 2 shows quantitative results for each model, on the same test sets that Table 1. While naive segmentation models (UNet and nnUNet) drastically fail to segment missing structures in the corresponding datasets (showing Dice of around 0), the methods that are aware of heterogeneous labels (HybridGNet and UNet HT) slightly reduce their performance, but can still recover the anatomical structure. More importantly, note that HybridGNet largely outperforms UNet HT in terms of HD distance for the missing labels (in red), while it is competitive in terms of Dice coefficient.

Conclusions. Here we show how HybridGNet can deal with heterogeneous labels in multi-center scenarios, where state-of-the-art UNet and nnUNet drastically fail. Moreover, the UNet HT model trained to be aware of heterogeneous labels by ignoring missing structures proved to be useful in avoiding contradictory signals between missing annotations and the background. However, this is not enough when there are contradictory signals between organs. In these cases, landmark-based models like HybridGNet offer a simple framework which can easily handle overlapping structures, particularly heterogeneous labels in multi-center scenarios.

4 Compliance with ethical standards

This is a numerical simulation study for which no ethical approval was required.

5 Acknowledgements

The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-PIC50420150100098LI, CAID-PIC-50220140100084LI) and ANPCyT (PICT 2016-0651, 2018-03907, PRH- 2019-00009). This work is supported by the Google Award for Inclusion Research (AIR) Program.

References

- [1] Enzo Ferrante, Puneet Kumar Dokania, Rafael Marini Silva, and Nikos Paragios, “Weakly supervised learning of metric aggregations for deformable image registration,” *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1374–1384, 2018.
- [2] Juan J Cerrolaza et al., “Computational anatomy for multi-organ analysis in medical imaging: A review,” *Medical image analysis*, vol. 56, pp. 44–67, 2019.
- [3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *MICCAI*. Springer, 2015.
- [4] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, “nnunet: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature methods*, vol. 18, 2021.
- [5] Hao Guan and Mingxia Liu, “Domain adaptation for medical image analysis: a survey,” *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, 2021.
- [6] Jana Kemnitz et al., “Combining heterogeneously labeled datasets for training segmentation networks,” in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 276–284.
- [7] Reuben Dorent, Thomas Booth, Wenqi Li, Carole H Sudre, Sina Kafiabadi, Jorge Cardoso, Sebastien Ourselin, and Tom Vercauteren, “Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets,” *Medical image analysis*, vol. 67, pp. 101862, 2021.
- [8] Julián Alberto Palladino, Diego Fernandez Slezak, and Enzo Ferrante, “Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter mr images,” in *16th International Symposium on Medical Information Processing and Analysis*. SPIE, 2020, vol. 11583, p. 1158302.
- [9] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy, “Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology,” in *MICCAI – MICCAI 2022*. 2022, pp. 120–129, Springer Nature Switzerland.
- [10] Olivier Petit et al., “Handling missing annotations for semantic segmentation with deep convnets,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 20–28. Springer, 2018.
- [11] Gregory Filbrandt et al., “Learning from partially overlapping labels: Image segmentation under annotation shift,” in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 123–132. Springer, 2021.
- [12] Nicolás Gaggion, Lucas Mansilla, Diego H. Milone, and Enzo Ferrante, “Hybrid graph convolutional neural networks for landmark-based anatomical segmentation,” in *MICCAI*. 2021, Springer International Publishing.

- [13] Nicolás Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone, and Enzo Ferrante, “Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: Applications to chest x-ray analysis,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 2, pp. 546–556, 2023.
- [14] Junji Shiraishi et al., “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules,” *American Journal of Roentgenology*, vol. 174, no. 1, 2000.
- [15] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá, “Padchest: A large chest x-ray image dataset with multi-label annotated reports,” *Medical image analysis*, vol. 66, pp. 101797, 2020.
- [16] Sema Candemir and others., “Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014.
- [17] Stefan et al Jaeger, “Automatic tuberculosis screening using chest radiographs,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014.
- [18] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *2016 fourth international conference on 3D vision*. IEEE, 2016.
- [19] Leland McInnes, John Healy, and James Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images

CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images

Nicolás Gaggion¹, Candelaria Mosquera^{2, 3}, Lucas Mansilla¹, Martina Aineseder²,
Diego H. Milone¹, and Enzo Ferrante¹

¹Institute for Signals, Systems and Computational Intelligence, sinc(i)
CONICET-UNL, Santa Fe, S3002, Argentina

³Health Informatics Department at Hospital Italiano de Buenos Aires, Buenos Aires,
CP, Argentina

³Universidad Tecnológica Nacional, Buenos Aires, CP, Argentina

Abstract

The development of successful artificial intelligence models for chest X-ray analysis relies on large, diverse datasets with high-quality annotations. While several databases of chest X-ray images have been released, most include disease diagnosis labels but lack detailed pixel-level anatomical segmentation labels. To address this gap, we introduce an extensive chest X-ray multi-center segmentation dataset with uniform and fine-grain anatomical annotations for images coming from six well-known publicly available databases: CANDID-PTX, ChestX-ray8, Chexpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR, resulting in 676,803 segmentation masks. Our methodology utilizes the HybridGNet model to ensure consistent and high-quality segmentations across all datasets. Rigorous validation, including expert physician evaluation and automatic quality control, was conducted to validate the resulting masks. Additionally, we provide individualized quality indices per mask and an overall quality estimation per dataset. This dataset serves as a valuable resource for the broader scientific community, streamlining the development and assessment of innovative methodologies in chest X-ray analysis. The CheXmask dataset is publicly available at: <https://physionet.org/content/chexmask-cxr-segmentation-data/>.

Background & Summary

Chest radiography is a pivotal imaging technique used to diagnose a variety of lung diseases, including pneumonia, tuberculosis, and lung cancer. The significant role of chest X-rays (CXR) in clinical practice is ascribed to their non-invasive nature, relatively low cost, and rapid diagnostic potential. However, the interpretation of these images poses a considerable challenge due to the intricate and overlapping structures within the thoracic cavity, and the subtle manifestations of certain pathological conditions. The high demand for chest radiography and the global shortage of radiologists accentuate the need for efficient and reliable automated analysis systems.

In recent years, methods based on deep learning (DL) have demonstrated exceptional prowess in interpreting medical images, rivaling and occasionally surpassing expert human performance [29, 17]. Convolutional neural networks (CNN) have been particularly instrumental in facilitating such computer-aided diagnosis (CADx) systems [33, 25]. Nonetheless, the success of these algorithms is closely tethered to the availability of accurately annotated data, with sufficient quantity and diversity, to train the models.

An essential task within this framework is segmentation – the delineation of specific anatomical structures or pathological lesions within an image. In the context of CXR, this might involve the

demarcation of anatomical structures such as lungs or heart, or the location of disease abnormalities [32]. Accurate and robust segmentation can serve as a precursor to other downstream tasks, for example providing significant information about the location and size of specific organs or detected abnormalities. However, manual segmentation is a time-consuming process, demanding substantial expertise, and thus, does not scale well to the size of large databases required for DL model training [31].

HybridGNet, a deep learning model for realistic organ contouring, offers a solution for the generation of anatomically plausible CXR segmentations [12, 14]. Utilizing a hybrid approach, it combines conventional convolution operations for image encoding with graph generative models for the anatomically-guided delineation of organ contours. The HybridGNet model was initially introduced with a small CXR landmark dataset to demonstrate its efficacy. In this work, we leverage this model to accomplish our main objective: introducing a large-scale segmentation dataset, named CheXmask, which provides anatomical masks with their corresponding quality index, for 6 extensive chest X-ray databases: CANDID-PTX[9], Chest x-ray8[36], Chexpert[17], MIMIC-CXR-JPG[20], Padchest[6] and VinDr-CXR[28]. These databases collectively represent a wide variety of geographical locations, patient demographics, and disease spectra, enabling the development of a broad, diverse segmentation dataset.

As the original databases lack manually curated ground-truth segmentations, we perform quality control by implementing our own Reverse Classification Accuracy (RCA) framework [35]. RCA allows to estimate the accuracy of a segmentation method for an individual image with no ground-truth (GT) masks, which is particularly valuable for large-scale image analysis studies like ours. The fundamental concept behind RCA involves training an auxiliary model (known as the reverse classifier) solely on the individual image, using its predicted segmentation as pseudo-GT. This model is then evaluated on a reference database that contains GT data to obtain a performance metric, which is expected to correlate with the performance that would be measured for the individual image if its GT was available. We validated this method by comparing it to traditional performance evaluation on a subset of test images with masks manually segmented by an expert physician. Additionally, since large-public CXR databases built from automatic analysis of electronic health records (EHR) are subject to errors both in image selection and image annotation, we found that RCA is a useful tool to detect out-of-distribution samples (e.g. poor-quality images). Thus, the RCA metrics for HybridGNet segmentations stand out as a powerful quality metric to handle large databases for downstream tasks, by detecting not only low quality segmentation masks, but also images that should be filtered out.

Our comprehensive analysis underscores the capacity of the HybridGNet model to generate high-quality segmentations of lungs and heart structures in CXR, and presents the RCA method as a way to use these segmentations for detection of poor-quality images in large volumes of data. CheXmask dataset provides a key resource to the medical imaging community and represents a significant stride towards democratizing access to diverse, large-scale segmentation datasets, thereby propelling the advancement of automated CXR analysis research.

Methods

Data Preparation

Image datasets

In this study, we utilized six extensive CXR datasets: CANDID-PTX (available at Figshare [10]), ChestX-ray8 (available in a dedicated website [1]), CheXpert (available in a dedicated website [2]), MIMIC-CXR-JPG (available at PhysioNet[19]), Padchest (available in a dedicated website [5]), and VinDr-CXR (available at PhysioNet[27]). Figure 1 provides an overview of the complete study, which was repeated six times, one for each dataset individually. The CANDID-PTX dataset [9], which encompasses 19,237 anonymized X-ray images, was collected at Dunedin Hospital in New Zealand between 2010 and 2020. The ChestX-ray8 dataset [36] contains 112,120 frontal-view X-ray images

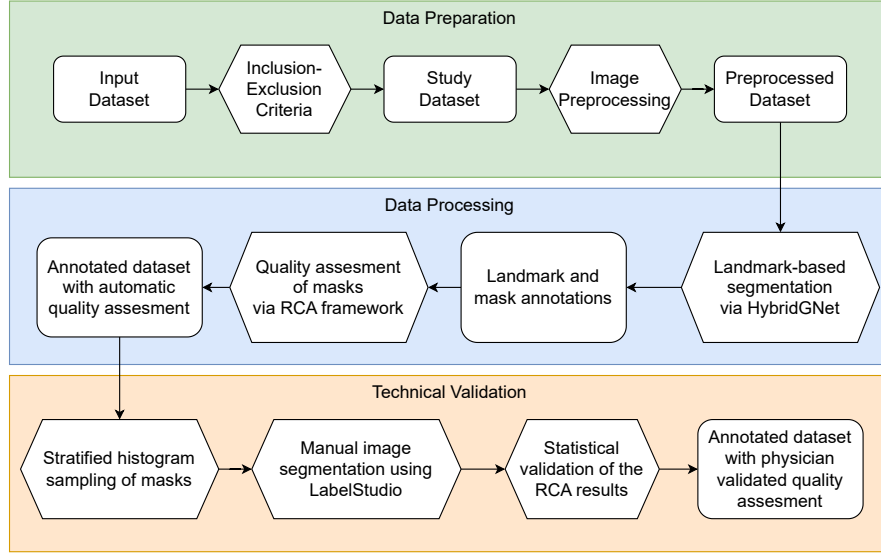


Figure 1: Data processing flowchart depicting the main steps involved in the building of the CheXmask dataset.

from 30,805 unique patients, annotated with text-mined fourteen disease image labels. The CheXpert dataset [17] comprises 224,316 chest radiographs collected from Stanford Hospital between October 2002 and July 2017. The MIMIC-CXR-JPG dataset [20] includes 377,110 CXR images from 227,827 imaging studies involving 65,379 patients; these were collected at the Beth Israel Deaconess Medical Center Emergency Department in the United States from 2011 to 2016. The Padchest dataset [6] is composed of 160,861 images from over 67,000 patients, collected from Hospital San Juan in Spain from 2009 to 2017. Lastly, the VinDr-CXR dataset [28] consists of 18,000 manually annotated images gathered from two primary hospitals in Vietnam.

Inclusion-exclusion criteria

Our study incorporated only frontal images, captured in posteroanterior (PA) or anteroposterior (AP) views (no lateral views were included). Specific selection criteria varied among the datasets due to their differing metadata, as summarized in Figure 2.

The CANDID-PTX, ChestX-ray8, and VinDr-CXR datasets contain only frontal images, so all images were included. We incorporated images from the Padchest dataset based on the “Projection” metadata field, including only those labeled as PA or AP, reducing the initial count of 160,861 images to 96,287. For the CheXpert dataset, images labeled as “Frontal” in the “Frontal/Lateral” metadata column were included, resulting in 187,825 out of the total 224,316 images. Lastly, images from the MIMIC-CXR-JPG dataset were included if the “ViewPosition” tag was listed as “PA” or “AP”, which resulted in 243,334 out of the initial 377,110 images.

Image preprocessing

For compatibility with HybridGNet input format, we preprocessed images to attain a uniform size suitable for generating pseudo-landmark contours. The standard size of the training data was 1024x1024 pixels, obtained by padding the images to square dimensions then resizing them to the required size, when needed.

Images from ChestX-ray8 dataset already satisfied the required dimension and format, thus requiring no pre-processing. The CANDID-PTX dataset met the required image dimension in DICOM

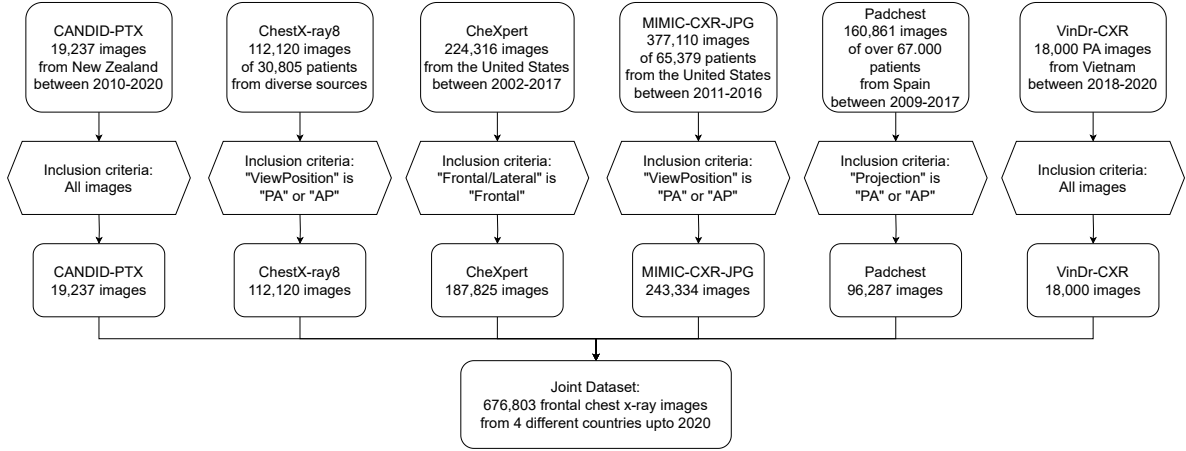


Figure 2: Summary diagram of inclusion-exclusion criteria.

format, so we converted images to PNG. The datasets of CheXpert, MIMIC-CXR-JPG, Padchest, and VinDr-CXR underwent a standard pre-processing pipeline, which included padding the images to a square shape and then resizing them to 1024x1024 pixels. Additionally, VinDr-CXR images needed extraction from DICOM files before the pre-processing. All these pre-processing procedures are reversible, which allows for the preservation of the original image shapes for releasing the generated annotations.

Data processing

Anatomically plausible segmentation via HybridGNet

The DL model HybridGNet segments organ contours by detecting the coordinates of anatomical landmarks. Thus, it is trained to minimize the distance between the predicted landmark positions and their GT location. The model incorporates an encoder-decoder architecture that combines standard convolutions for image encoding and graph generative models to achieve anatomically plausible representations. Pixel-level masks are obtained by filling in the contours defined by the landmarks predicted by HybridGNet. A detailed description of the HybridGNet model can be found in the work of Gaggion et al. [12, 14]. A modified training procedure was presented in a later work [13], to avoid domain memorization issues which emerge when dealing with heterogeneous labels in multi-centric scenarios (i.e. images from some medical centers contain only lung annotations, while others include both lung and heart masks).

In this work, we used the modified multi-centric training procedure [13] to retrain the HybridGNet model with the complete dataset used in prior studies (i.e., both the train and test splits of previous works). This dataset, which will be referred to as Chest-Xray-Landmark dataset, was released previously and is available on GitHub^a. The Chest-Xray-Landmark dataset is curated from a variety of sources, providing annotations for a total of 911 images. These images are sourced from different repositories, including the JSRT dataset [34] (available from its dedicated website [3]), a small subset of the Padchest dataset [6] (accessible on its designated website [5]), as well as the Montgomery [7] and Shenzhen [18] datasets (both obtainable from the NIH National Library of Medicine [4]). Please note that the Chest-Xray-Landmark dataset does not provide the images themselves; users are required to source the original images from their respective origins. All 911 images within this dataset contain lung annotations and only 383 images also include heart labels. The training process incorporated various data augmentation techniques, such as random rotations, scaling, and color shifting, to enhance

^a<https://github.com/ngaggion/Chest-xray-landmark-dataset>

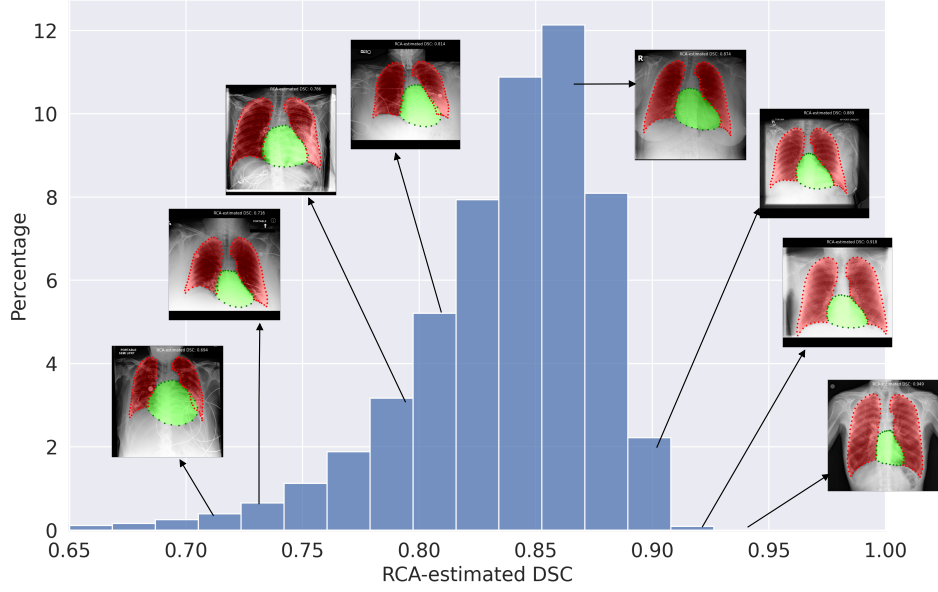


Figure 3: Histogram showcasing the distribution of the RCA-estimated DSC for the complete CheX-mask database. Example images with their landmark-based segmentations were drawn with lines to their corresponding histogram bin.

robustness and generalizability.

A fully working demo of the model used to segment the complete dataset is available on Hugging Face Spaces at <https://huggingface.co/spaces/ngaggion/Chest-x-ray-HybridGNet-Segmentation>.

Reverse classification accuracy to measure segmentation quality

We employed RCA to estimate the Dice Similarity Coefficient (DSC) [8] in the absence of ground truth. DSC is a widely used metric for assessing the similarity between two sets of binary labels. By using RCA, it is possible to produce accurate estimations of DSC, thereby making it easier to perform quality control in the absence of GT. The DSC quantifies the overlap between the predicted segmentation and the GT segmentation, ranging from 0 (no overlap) to 1 (perfect overlap). It is defined as $DSC = \frac{2TP}{2TP + FP + FN}$, where TP , FP , and FN denote the number of true positive, false positive, and false negative pixels, respectively.

To obtain the estimated DSC for a segmentation mask, the RCA framework consists in training a new segmentation method using solely this image with the predicted mask as GT. This new trained model (i.e., the reverse classifier) is then applied to segment a reference image set, with known GT masks. The hypothesis is that the segmentation accuracy of the reverse model on the reference set correlates to the segmentation quality of the original image: if the original segmentation of the evaluated image was good, the segmentation accuracy of the reverse model on this reference set should also be good.

In the original work, the authors experiment with three different methods for implementing the RCA classifier: Atlas Forests, Deep Learning, and Atlas-based Label Propagation. In this case, we used the last one, which employs a non-rigid registration technique for single-atlas label propagation, where the single atlas corresponds to the evaluated image together with its predicted segmentation. A summary value of the distribution of DSC across the reference set images is then used as performance metric. We will use for these values the *mean* and *max* RCA-estimated DSC. The RCA-estimated DSC is expected to correlate well with the real DSC that one would obtain if GT data was available.

Figure 3 showcases the distribution of RCA-estimated DSC for the complete CheXmask dataset with examples sampled from each bin.

The use of DL in atlas registration has shown improved performance over classic multi-atlas approaches, while also offering a significant speed-up in computation. In order to speed up the evaluation process, we implemented a DL-based atlas registration procedure [26] as label propagation method (that we refer as Deep RCA), which is several orders of magnitude faster than traditional registration methods. Specifically, we divide the registration procedure into two stages: first performing rigid registration to globally align the corresponding images, and then proceeding with deformable registration to correct local misalignments.

Data Records

The CheXmask dataset is structured as one file of comma-separated values (CSV) for each CXR dataset. The CSV content is explained in Table 1. We do not include images or metadata from the original datasets. We include an image ID in the first column of the CSV files, which has the same column name as the ID column of the respective dataset, and allows to match the rows in this CSV with the original dataset’s instances. We also include the pre-processed version of the masks, allowing to use all datasets in the same image resolutions (please also note that CANDID-PTX and ChestX-ray8 masks are already in the desired resolution). This database [11] is available at PhysioNet [16], in the following repository: <https://physionet.org/content/chexmask-cxr-segmentation-data/>.

Table 1: Description of the columns in CheXmask CSV files.

Column	Description
Image ID	Contains references to the original images as per the original metadata, thus the column name changes across dataset.
Dice RCA (Max)	Provides the maximum Dice Similarity Coefficient for the Reverse Classification Accuracy (RCA), indicating the quality of the segmentation.
Dice RCA (Mean)	Provides the mean Dice Similarity Coefficient for the Reverse Classification Accuracy (RCA), indicating the quality of the segmentation.
Landmarks	Includes a set of points representing the contour of the organs, as obtained by the HybridGNet.
Left Lung	Contains segmentation masks of the left lung in run-length encoding (RLE).
Right Lung	Contains segmentation masks of the right lung in RLE.
Heart	Contains segmentation masks of the heart in RLE.
Height	Height of the segmentation mask, necessary to decode RLE.
Width	Width of the segmentation mask, necessary to decode RLE.

Technical Validation

The technical validation of the CheXmask dataset was conducted through a three-tiered approach: (1) evaluating the quality of HybridGNet segmentations by measuring the DSC on a gold-standard subset revised by an expert physician; (2) validating the relevance of the RCA-estimated DSC as performance metric; (3) evaluating the quality of the masks for the six source datasets separately, by exploring their suitability as a training GT set for the segmentation task.

1) Validation via physician annotations as gold-standard

We built a gold-standard set of masks (for images sampled from the six original datasets) labeled by an experienced physician to evaluate the segmentation quality of CheXmask lungs and heart masks. The

physician, counting with more than five years of experience, performed a manual revision of the right lung, left lung and heart landmark-based segmentations. We used an open-source labeling platform called LabelStudio to do so. The landmarks predicted by HybridGNet for a subset of images were uploaded as predictions to LabelStudio, where the annotator corrected these predicted landmarks by moving them to their correct specific contour positions. This guarantees that the GT masks are based on the same collection of nodes than the evaluated masks, and reduces the time needed for annotation. Figure 4 shows the annotation interface. We include the source code to launch this interface in our repository, as an easy tool to perform further validations. We measured the DSC of HybridGNet’s masks comparing with the manually-corrected landmark-annotations, and report statistics on the DSC distribution for the six public datasets.

The sample size of the subset used as gold-standard evaluation was chosen based on the work by El Jurdi and Colliot [21], which suggests sampling 100 images for a 4%-wide confidence interval in segmentation tasks. We sampled 30 images per dataset, obtaining a final gold-standard set of 180 images. We combined random sampling and histogram-based sampling to ensure the presence of high and low quality masks. For each dataset, ten images were sampled based on a 10-bins histogram of the RCA-estimated DSC of all the dataset’s images, taking one random image per bin. Another twenty images were then randomly selected from the whole dataset.

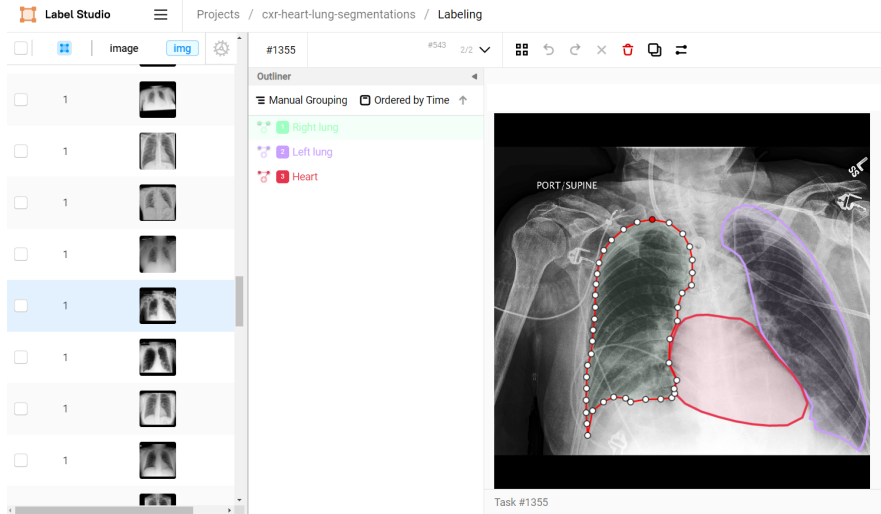


Figure 4: Label Studio setup for manual landmark-based segmentation

Table 2 summarizes the results of physician validation for lung and heart segmentation across the six datasets on their 30-samples subsets. We report DSC, Hausdorff Distance (HD), and Hausdorff 95% Distance (HD95) as evaluation metrics. HD and HD95 are distance error metrics, providing insights into the difference between the contours of the GT masks and the predicted masks.

In terms of lung segmentation, the datasets generally exhibit high DSC values, ranging from 0.947 to 0.981. This indicates a strong overlap between the GT masks and the predicted lung masks. Among the datasets, CANDID-PTX exhibits the highest agreement, while CheXpert has the lowest mean DSC. CheXpert shows relatively higher HD and HD95 values compared to other datasets, suggesting a larger discrepancy between the contours.

For heart segmentation, the DSC values range from 0.950 to 0.980, indicating a strong overlap between the GT and predicted heart masks across all datasets. The HD and HD95 values for heart segmentation are generally lower compared to lung segmentation, indicating a smaller distance between the contours.

Overall, the results indicate that the segmentation models perform well in capturing the lung and

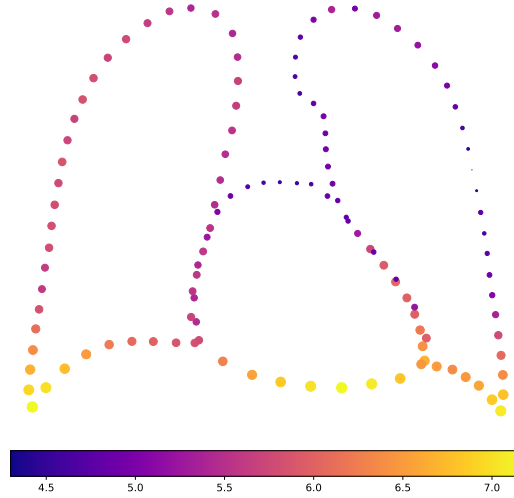


Figure 5: Illustration of the Mean Squared Error (MSE) per landmark across the entire gold-standard set, depicted in logarithmic scale for improved visual clarity. The color intensity of each landmark represents the magnitude of the MSE. Moreover, the size of each landmark is proportional to its respective MSE, thus larger landmarks indicate a greater prediction error.

heart structures. CANDID-PTX consistently demonstrates the highest agreement for both lungs and heart, while CheXpert exhibits relatively lower agreement. These findings highlight the variations in segmentation performance across different datasets and provide valuable insights for further improving the accuracy of the segmentation algorithms.

Per-landmark error analysis

In addition to the segmentation metrics, we also computed the mean squared error (MSE) per landmark across the entire set. This measure provided deeper insights into the positional accuracy of the landmarks predicted by HybridGNet. Interestingly, the results suggested that certain landmarks were more prone to prediction errors than others. Specifically, we observed higher MSE values for landmarks corresponding to the lower part of the lungs and the heart, indicating that these regions are more difficult for the network.

Figure 5 illustrates the MSE per landmark across the whole dataset. Each landmark is color-coded based on its respective MSE value, using a logarithmic scale for better visualization. The size of each landmark also corresponds to the magnitude of the MSE, with larger landmarks indicating higher errors. As can be seen in this figure, the bottom part of the lungs and the heart landmarks tend to be larger, denoting higher prediction errors. This aligns with the observations made by the expert physician responsible for the manual annotations, who noted that the boundaries of the heart and the lower section of the lungs tend to exhibit more blurriness compared to the upper part.

2) Validation via RCA-estimated DSC

As previously mentioned, computing the RCA-estimated DSC requires training a rigid and deformable registration model based on deep learning. To this end, we employed 80% of the ChestXray-Landmarks database for training. This model was used to propagate the predicted labels from the image atlas to

Table 2: Results of physician validation with standard deviations ($n = 30$ per dataset)

Dataset	Lungs			Heart		
	DSC	HD	HD95	DSC	HD	HD95
CANDID-PTX	0.981 ± 0.038	23.7 ± 43.0	11.9 ± 23.6	0.950 ± 0.111	23.6 ± 51.9	21.0 ± 46.6
CheXpert	0.947 ± 0.087	39.3 ± 57.7	27.0 ± 46.3	0.961 ± 0.083	18.0 ± 37.2	14.8 ± 30.3
ChestX-Ray8	0.961 ± 0.097	32.3 ± 63.4	18.8 ± 44.1	0.958 ± 0.133	18.6 ± 58.6	17.2 ± 55.1
MIMIC-CXR-JPG	0.965 ± 0.077	29.7 ± 51.1	15.1 ± 32.2	0.954 ± 0.101	22.9 ± 51.9	21.2 ± 47.8
Padchest	0.966 ± 0.089	23.0 ± 42.7	12.7 ± 27.5	0.980 ± 0.035	12.8 ± 24.3	10.4 ± 18.8
VinDr-CXR	0.978 ± 0.048	18.8 ± 40.1	8.9 ± 18.5	0.953 ± 0.119	21.9 ± 55.4	20.6 ± 52.6

the reference images (composed of the 5 most similar images to the atlas from the training set), and finally compute the RCA-estimated DSC (mean and max across the 5 values).

To validate that RCA-estimated DSC accurately approximates the real DSC, we proceeded to use simulated segmentation masks of various qualities for which we can compute the real coefficient. To this end, we generated a set of candidate segmentations of various quality (i.e. with real DSC ranging from low to high values), by training 12 UNet models on the same dataset but halting at various epochs (models trained for a few epochs produced lower quality segmentations while those trained until convergence produce better segmentations). Then, we assessed the correlation of RCA-estimated DSC on the remaining 20% test-split of the ChestXray-Landmarks database. The Pearson Correlation scores for the Max and Mean RCA-estimated DSC across the reference set were 0.93 and 0.94 respectively, demonstrating a strong correlation (Figure 6). We found that the RCA-estimated value tends to slightly sub-estimate the true DSC, suggesting the performance measured with RCA follows a conservative approach (i.e., more pessimistic than the actual true performance). We selected the *mean* RCA-estimated DSC as RCA estimation metric for the following experiments since it achieved better correlation.

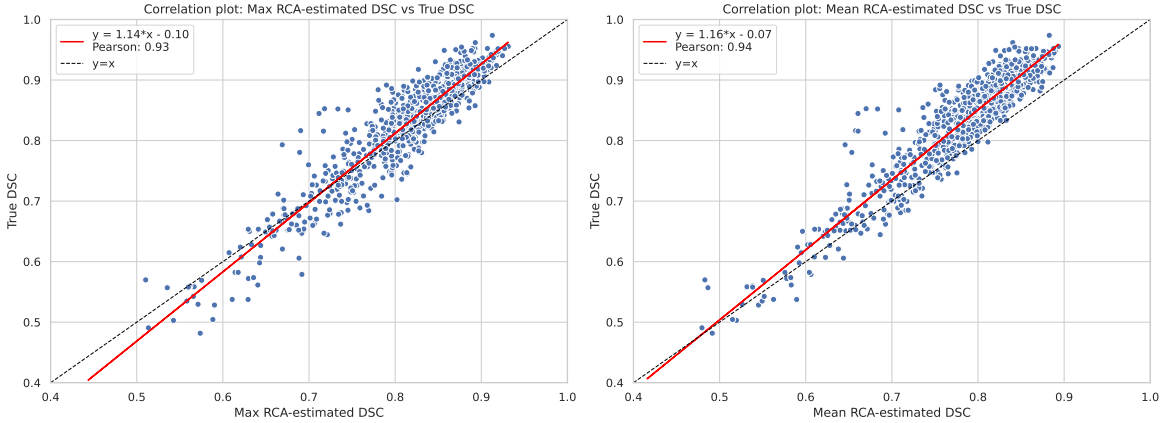


Figure 6: Correlation between the RCA values and the true DSC on the 20% test-subset of the ChestXray-Landmarks database

Evaluation of segmentation quality through RCA

As part of the technical validation on the quality of lungs and heart masks, we report a statistical description of the distribution of mean RCA-estimated DSC for all images in each dataset (Table 3). The full histograms are presented in Figure 7. For all datasets, the mean RCA-estimated DSC is higher than 0.88. As explained above, the true DSC distribution probably has a larger mean, since the RCA-estimated DSC was found to under-estimate the performance.

Table 3: Mean image level RCA-estimated DSC: statistical analysis

Dataset name	Sample size (n)	Mean	Std	Min	1%	5%	25%	50%	75%	Max
CANDID-PTX	19235	0.853	0.031	0.573	0.756	0.795	0.837	0.858	0.875	0.920
Chestx-ray8	112120	0.841	0.044	0.148	0.691	0.764	0.823	0.850	0.869	0.925
CheXpert	187825	0.830	0.038	0.400	0.718	0.761	0.808	0.835	0.857	0.918
MIMIC-CXR-JPG	243334	0.831	0.051	0.149	0.650	0.738	0.811	0.842	0.864	0.921
Padchest	96184	0.851	0.042	0.103	0.672	0.790	0.838	0.858	0.874	0.927
VinDr-CXR	18000	0.850	0.033	0.469	0.738	0.794	0.835	0.855	0.872	0.919

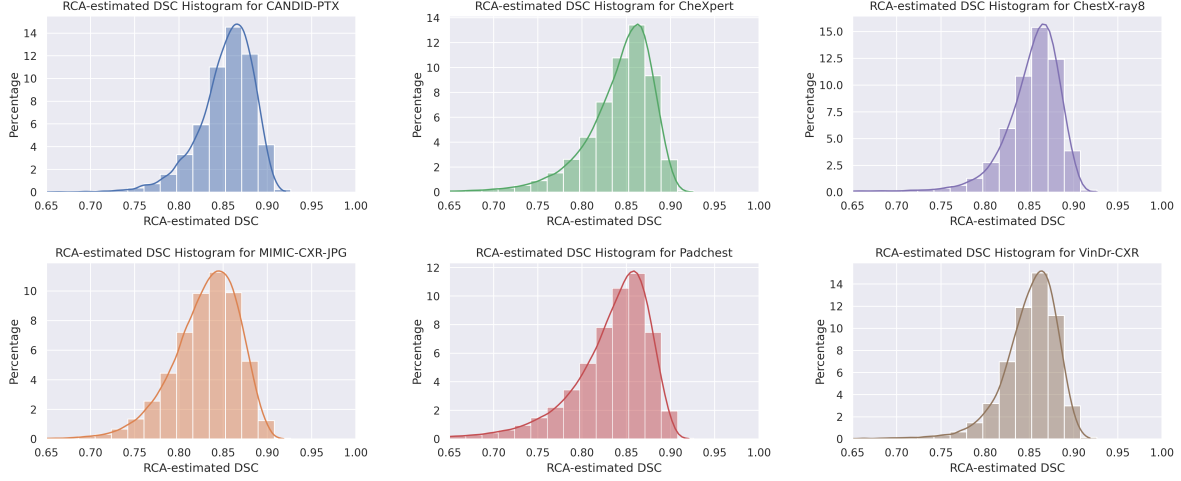


Figure 7: Histograms illustrating the distribution of segmentation quality across the six datasets. Each histogram represents the distribution of RCA (Reverse Classification Accuracy) estimations of DSC for a specific dataset, providing a visual representation of the segmentation performance. The histograms are truncated at 0.65 for visualization purposes. The full range of values is considered in the statistical analysis (Table 3).

Demographic bias analysis through RCA

To explore potential biases [24, 30] in the masks quality, we conducted a detailed analysis incorporating the metadata associated with the MIMIC-CXR-JPG, Padchest, ChestX-ray8, and CheXpert datasets. We took into account parameters such as disease findings, the sex and age of the patients, and the X-ray capture view (PA or AP).

Figure 8 presents histograms that clearly demonstrate a superior RCA-estimated DSC for PA images compared to AP images across all investigated datasets. This is consistent with the fact that PA images tend to come from hospitalized patients, who are more difficult to position in standard views and usually include artifacts or cables that may act as confounders or occlude body parts, making anatomical segmentation more challenging. In contrast, discrepancies based on sex were less pronounced, with a slightly diminished segmentation quality observed in females. This is also consistent with the fact x-ray imaging of the upper thorax women’s breasts tend to occlude the imaged organs, resulting in poorer image contrast for the relevant anatomy [15] and making the segmentation task more challenging. Furthermore, the influence of disease findings on segmentation quality varied amongst datasets, with significant disparities evident in the ChestX-ray8 and CheXpert datasets.

Upon evaluating the relationship between patient age and RCA-estimated DSC in the CheXpert dataset, no substantial patterns emerged. It is important to note, however, that all subjects are above the age of 18. On the contrary, when analyzing results for the ChestX-ray8 dataset which includes patients under 18, we observed lower RCA-estimated DSC for these cases. This is consistent with

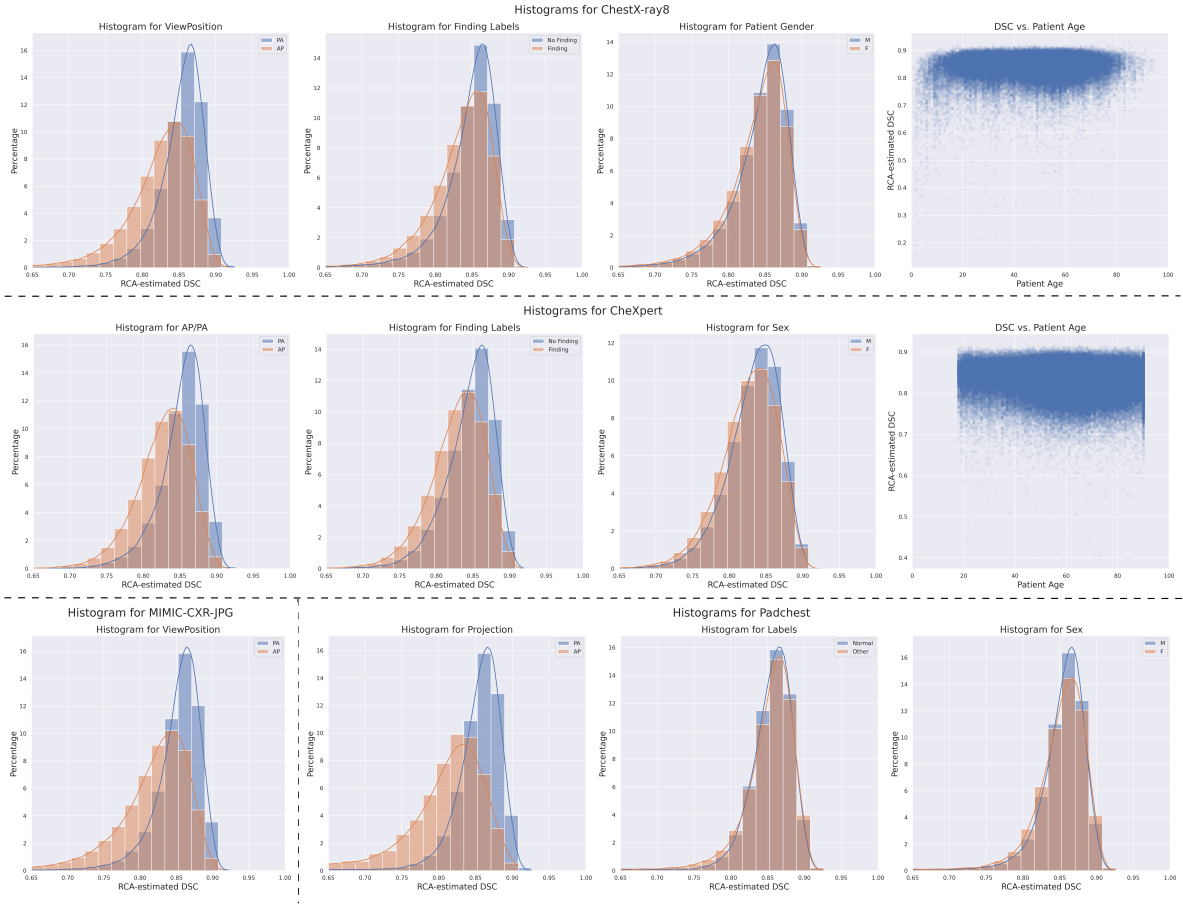


Figure 8: Histograms depicting RCA-estimated DSC distribution across diverse metadata categories, namely imaging protocol, sex, and disease findings, supplemented by scatter plots correlating RCA-estimated DSC and age. These visual aids elucidate the relationship between segmentation quality and metadata categories, particularly highlighting differences associated with X-ray capture view (PA versus AP), age and the presence of disease findings.

results reported in a previous analysis of age patterns for the HybridGNet model [14], where subjects below the age of 18 demonstrated subpar performance due to their absence in the training set. We hypothesize a similar trend may be affecting the sparse set of ChestX-ray8 X-rays taken for individuals under the age of 18.

These findings underscore the importance of desegregated analysis with respect to image acquisition protocol, disease presence and demographic attributes when understanding segmentation quality.

Out-of-distribution detection through RCA

In this section we focus on analyzing those segmentation masks which received lower RCA-estimated DSC values. Through empirical observation, we discovered that RCA-estimated DSC values lower than 0.7 tend to be indicative of out-of-distribution cases. To exemplify this, we randomly selected a subset of six images from each dataset with some of the lowest RCA values, which are showcased in Figure 9. By performing a manual and qualitative analysis of the lowest RCA-estimated DSC masks for each dataset, we found a wide variety of challenging scenarios, ranging from images with high noise levels or mislabeled lateral CXR, to an X-ray image of a mobile phone from the Padchest dataset. That is

Table 4: Dataset-level estimation of the quality of the segmentation via retraining the HybridGNet model from scratch.

Training Dataset	Lungs ($n = 911$)			Heart ($n = 383$)		
	DSC	HD	HD 95	DSC	HD	HD 95
CANDID-PTX	0.959 ± 0.020	36.4 ± 23.3	18.2 ± 10.5	0.919 ± 0.045	41.1 ± 22.1	36.8 ± 21.7
Chestx-ray8	0.960 ± 0.019	36.2 ± 22.7	17.7 ± 10.3	0.927 ± 0.031	39.9 ± 16.5	35.2 ± 16.3
CheXpert	0.946 ± 0.028	42.0 ± 23.5	22.9 ± 12.1	0.888 ± 0.053	58.2 ± 27.0	54.0 ± 26.4
Padchest	0.956 ± 0.028	37.7 ± 23.7	19.0 ± 12.3	0.924 ± 0.037	40.9 ± 19.6	36.6 ± 19.2
MIMIC-CXR	0.956 ± 0.020	37.5 ± 22.8	19.2 ± 10.5	0.921 ± 0.031	44.2 ± 19.9	39.6 ± 19.6
VinDr-CXR	0.959 ± 0.023	35.4 ± 23.3	17.7 ± 11.7	0.925 ± 0.035	38.4 ± 16.8	34.0 ± 16.2
20% test split	Lungs ($n = 181$)			Heart ($n = 76$)		
HybridGNet model trained on 80% train split [13]	0.967 ± 0.017	32.5 ± 20.6	15.6 ± 10.9	0.937 ± 0.029	34.6 ± 14.8	29.4 ± 13.4

why we recommend users of the CheXmask dataset to only use segmentations whose RCA-estimated DSC is higher than 0.7.

Table 3 can be useful to estimate the prevalence of such out-of-distribution images in each dataset. For instance, in the CANDID-PTX and VinDr-CXR datasets, the number of out-of-distribution images is relatively low. Notably, the VinDr-CXR dataset does not include any non-posteroanterior (PA) images, as all were manually annotated by physicians. However, the inclusion of zoomed-out images capturing both arms, which were absent in the HybridGNet training set, leads to lower-quality segmentations in those cases. In contrast, the larger datasets contain a higher number of cases mislabeled as PA or AP in their metadata. These mislabeled images could be easily identified and discarded by thresholding the RCA-estimated DSC, for example using the 1% or 5% quartile of a dataset as threshold. This approach can help to improve the overall quality and reliability of downstream applications developed using this datasets.

3) Validation via retraining segmentation models with CheXmask

We performed a final quality control at the dataset-level to evaluate the quality of HybridGNet segmentations per dataset (not individually as we did with RCA-estimated DSC). The hypothesis is that if the segmentations are good, they can be used as training set for a segmentation model that will obtain high performance. Thus, we trained six new HybridGNet models, by using the set of predicted masks of each of the six public datasets as training GT. We then evaluated model performance on the complete ChestXray-Landmarks database. All models were trained using the same configurations and hyperparameters as the original HybridGNet model [13]. As a comparison reference, we include the performance metrics reported on prior studies for the 20% test split of the ChestXray-Landmarks database which can be seen as an upper bound in performance, as this model was trained with the training split of the ChestXray-Landmarks database.

Table 4 shows the results for lungs and heart segmentation. A slight decrease is observed in terms of DSC, decreasing from 0.967 to 0.955 when using the segmented datasets as training set, in comparison to the reference values. This indicates that the overall quality of the segmentations is good enough to train models which perform well when compared with existing expert annotations.

Usage Notes

The CheXmask dataset is intended to serve as a resource for researchers working in the field of medical imaging, particularly those interested in computational anatomy, shape understanding, CXR analysis and image segmentation. It may be useful for a wide range of downstream applications, including deep learning model development and evaluation, training of generative models, clinical decision support systems, disease detection and diagnosis, and anomaly detection.

For the sake of completeness, we release segmentation masks for all images included in our study. However, based on the qualitative and quantitative analyses presented in this paper (see section Technical Validation), our recommendation for downstream tasks is to only use segmentations whose RCA-



Figure 9: Illustration of a random subset of images with some of the lowest RCA values from each dataset. These images were identified as having low-quality segmentations based on their corresponding RCA scores. The selection showcases various types of images, including those with noise, mislabeled lateral CXR, and even non-medical objects like a mobile phone in the Padchest dataset.

estimated DSC is higher than 0.7, to avoid including out-of-distribution images as well as low quality masks.

A critical point to note is that this dataset does not include the original CXR images due to proprietary and privacy considerations. Hence, researchers intending to utilize this dataset are required to source the original images directly from the respective datasets cited in the manuscript. It is imperative that all conditions of usage and restrictions established by the providers of these datasets are fully respected and complied with.

The segmentation masks within this dataset are encoded using the Run-Length Encoding (RLE) technique, a standardized method for representing binary masks in a compact manner. In order to decode these masks back to their original binary format, researchers can use Python along with widely-used scientific computing libraries such as NumPy and Pandas. We provide the necessary scripts for encoding and decoding these masks to facilitate their use.

For researchers aiming to employ machine learning or deep learning techniques for analyzing this dataset, we recommend the use of well-established libraries. In particular, the HybridGNet model, used in this work, was developed in PyTorch. The source code for this model is made available alongside the dataset, allowing researchers to reproduce our results, conduct comparative studies, or further refine the methodology.

The segmentation masks are available in two formats. The first retains the original resolution as provided by the respective imaging source. The second consists of preprocessed masks with an uniform resolution across different imaging sources. We also provide the corresponding image preprocessing scripts which researchers can apply to the original images, enabling the use of all datasets at the same resolution. These scripts have been designed to replicate the preprocessing steps applied to the original datasets, thereby facilitating the integration of all these major CXR datasets.

Limitations of Study

There are several limitations to consider in this study. Firstly, the validation of the subset of images was performed by a single physician, which may introduce potential biases or subjective interpretations. While efforts were made to ensure accuracy and consistency, further validation by multiple radiologists or experts could enhance the reliability and generalizability of the findings.

The selection process focused on including only CXR images in the posteroanterior (PA) or anteroposterior (AP) view, aiming to ensure consistency and homogeneity. However, due to potential inaccuracies or inconsistencies in the metadata (particularly in large-scale databases built from automatic analysis of electronic health records), there is a possibility that other types of images may have been included in the PA and AP views. Despite the utilization of the Reverse Classification Accuracy (RCA) framework, which provides a reliable metric to detect low-quality segmentations, it is important to acknowledge that some images that do not meet the specific PA or AP criteria may have inadvertently leaked into the dataset. While RCA helps identify such instances, there remains a potential for the inclusion of non-PA or non-AP images that could impact the segmentation quality assessment. Therefore, when working with these large datasets, it is crucial to exercise caution and acknowledge the possibility of mislabeled or misclassified images. Further investigations and refinements in the dataset selection process, including more robust metadata validation techniques, would be beneficial to enhance the reliability and accuracy of the segmentation results.

Finally, it is important to highlight that, since segmentation masks released in CheXmask are generated with HybridGNet, they tend to follow patterns of anatomical plausibility, even in presence of complex conditions like tuberculosis, which generate different type of organ occlusions. This is of particular importance for lung segmentation, where there are usually two ways in which masks can be annotated, either following the 'air' or 'anatomy' strategy [22, 23]. While the 'air' strategy focus on segmenting the air cavity of the lung, excluding areas of increased opacity due to infection, the 'anatomy' strategy provide a comprehensive view which follows the expected anatomy, including these opaque areas. Segmentation masks generated by HybridGNet (i.e. those in CheXmask) are more

similar to those generated via the 'anatomy' strategy. For a complete discussion about this, see Section V.5) and Figure 6 in Gaggion et. al [14].

Code availability

The code associated with this study is available in our Github repository^b. The repository encompasses Python 3 code for various components, including data preparation, data post-processing, technical validation, and the deep learning models. The data preparation section includes scripts for preprocessing the images. The data post-processing section provides scripts for converting the segmentation masks from run-length encoding to a binary mask format, examples of how to read the model and the necessary code to revert the pre-processing steps for each dataset. The technical validation section includes the code for the individual RCA analysis and the processing of the physician results. Additionally, the repository includes the code for the deep learning models used for image segmentation, including the HybridGNet model architecture, weights, training and inference scripts. The software prerequisites for running the code are outlined in the repository's README file.

Acknowledgements

This work was supported by Argentina's National Scientific and Technical Research Council (CONICET), which covered the salaries of E.F., D.M. and N.G. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research and ANPCyT (PICT-PRH-2019-00009).

Author contributions statement

E.F., N.G. and C.M. conceptualized the idea of this paper. N.G. performed the experiments. C.M. designed the LabelStudio interface and analyzed the results. L.M. implemented the deformable registration module for the RCA estimation. M.A. manually segmented the gold-standard set. D.H.M. and E.F. analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare not conflict of interest.

References

- [1] Chest-xray8 dataset. <https://nihcc.app.box.com/v/ChestXray-NIHCC>, accessed: August 27, 2023
- [2] Chexpert dataset. <https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2>, accessed: August 27, 2023
- [3] Japanese society of radiological technology (jsrt) dataset. <http://db.jsrt.or.jp/eng.php>, accessed: August 27, 2023
- [4] Nih tuberculosis chest x-ray datasets. <https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/index.html>

^b<https://github.com/ngaggion/CheXmask-Database>

- [5] Padchest dataset. <https://bimcv.cipf.es/bimcv-projects/padchest/>, accessed: August 27, 2023
- [6] Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797 (2020)
- [7] Candemir, S., others.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging* **33**(2), 577–590 (2014). <https://doi.org/10.1109/TMI.2013.2290491>
- [8] Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology* **26**(3), 297–302 (1945)
- [9] Feng, S., Azzollini, D., Kim, J.S., Jin, C.K., Gordon, S.P., Yeoh, J., Kim, E., Han, M., Lee, A., Patel, A., et al.: Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence* **3**(6), e210136 (2021)
- [10] Feng, S., Azzollini, D., Kim, J.S., Jin, C.K., Kim, E., Gordon, S., et al.: Candid-ptx. <https://auckland.figshare.com/articles/dataset/CANDID-PTX/14173982> (2021). <https://doi.org/10.17608/k6.auckland.14173982>
- [11] Gaggion, N., Mosquera, C., Aineseder, M., Mansilla, L., Milone, D., Ferrante, E.: CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.1). *PhysioNet* (2023), <https://doi.org/10.13026/dx54-8351>
- [12] Gaggion, N., Mansilla, L., Milone, D.H., Ferrante, E.: Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In: *MICCAI*. Springer International Publishing (2021)
- [13] Gaggion, N., Vakalopoulou, M., Milone, D.H., Ferrante, E.: Multi-center anatomical segmentation with heterogeneous labels via landmark-based models. In: *20th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE (2023)
- [14] Gaggion, N., Mansilla, L., Mosquera, C., Milone, D.H., Ferrante, E.: Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. *IEEE Transactions on Medical Imaging* (2022). <https://doi.org/10.1109/TMI.2022.3224660>
- [15] Ganz, M., Holm, S.H., Feragen, A.: Assessing bias in medical ai. In: *Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML)* (2021)
- [16] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C., Stanley, H.: Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000), <https://www.physionet.org/content/circ101/1.1/>
- [17] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 33, pp. 590–597 (2019)
- [18] Jaeger, S.e.a.: Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging* **33**(2), 233–245 (2014). <https://doi.org/10.1109/TMI.2013.2284099>
- [19] Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg - chest radiographs with structured labels (version 2.0.0). <https://physionet.org/content/mimic-cxr-jpg/2.0.0/> (2019). <https://doi.org/10.13026/8360-t248>

- [20] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
- [21] Jurdi, R.E., Colliot, O.: How precise are performance estimates for typical medical image segmentation tasks? In: IEEE International Symposium on Biomedical Imaging (ISBI 2023). IEEE, Cartagena de Indias, Colombia (April 2023), <https://hal.archives-ouvertes.fr/hal-04104891>
- [22] Karargyris, A., Siegelman, J., Tzortzis, D., Jaeger, S., Candemir, S., Xue, Z., Santosh, K.C., Vajda, S., Antani, S., Folio, L., Thoma, G.R.: Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays. *International Journal of Computer Assisted Radiology and Surgery* **11**(1), 99–106 (2016). <https://doi.org/10.1007/s11548-015-1242-x>
- [23] Larrazabal, A.J., Martínez, C., Glocker, B., Ferrante, E.: Post-dae: anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Transactions on Medical Imaging* **39**(12), 3813–3820 (2020)
- [24] Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences* **117**(23), 12592–12594 (2020)
- [25] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)
- [26] Mansilla, L., Milone, D.H., Ferrante, E.: Learning deformable registration of medical images with anatomical constraints. *Neural Networks* **124**, 269–279 (2020)
- [27] Nguyen, H.Q., Pham, H.H., tuan linh, l., Dao, M., khanh, l.: Vindr-cxr: An open dataset of chest x-rays with radiologist annotations (version 1.0.0). <https://physionet.org/content/vindr-cxr/1.0.0/> (2021). <https://doi.org/10.13026/3akn-b287>
- [28] Nguyen, H.Q., Lam, K., Le, L.T., Pham, H.H., Tran, D.Q., Nguyen, D.B., Le, D.D., Pham, C.M., Tong, H.T., Dinh, D.H., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data* **9**(1), 429 (2022)
- [29] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
- [30] Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. *nature communications* **13**(1), 4581 (2022)
- [31] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. pp. 234–241. Springer (2015)
- [32] Roulet, N., Slezak, D.F., Ferrante, E.: Joint learning of brain lesion and anatomy segmentation from heterogeneous datasets. In: *International Conference on Medical Imaging with Deep Learning*. pp. 401–413. PMLR (2019)
- [33] Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. *Annual review of biomedical engineering* **19**, 221–248 (2017)

- [34] Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology* **174**(1) (2000)
- [35] Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging* **36**(8), 1597–1606 (2017)
- [36] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2097–2106 (2017)