

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

Nicolás Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone and Enzo Ferrante

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.

Index Terms—Graph convolutional neural networks, anatomically plausible segmentation, landmark based segmentation, graph generative models, localized skip connections

I. 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 anatomical

N. Gaggion, L. Mansilla, D.H. Milone and E. Ferrante are with the Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, Argentina. (e-mails: ngaggion@sinc.unl.edu.ar, lmansilla@sinc.unl.edu.ar, dmilone@sinc.unl.edu.ar, eferrante@sinc.unl.edu.ar).

C. Mosquera is with the Health Informatics Department at Hospital Italiano de Buenos Aires and with Universidad Tecnológica Nacional, Buenos Aires, Argentina (e-mail: candelaria.mosquera@hospitalitaliano.org.ar).

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 AN-PCyT (PICT 2016-0651, PICT 2018-03907).

ically 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 MIC-CAI 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.

II. 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 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 CurveGCN [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.

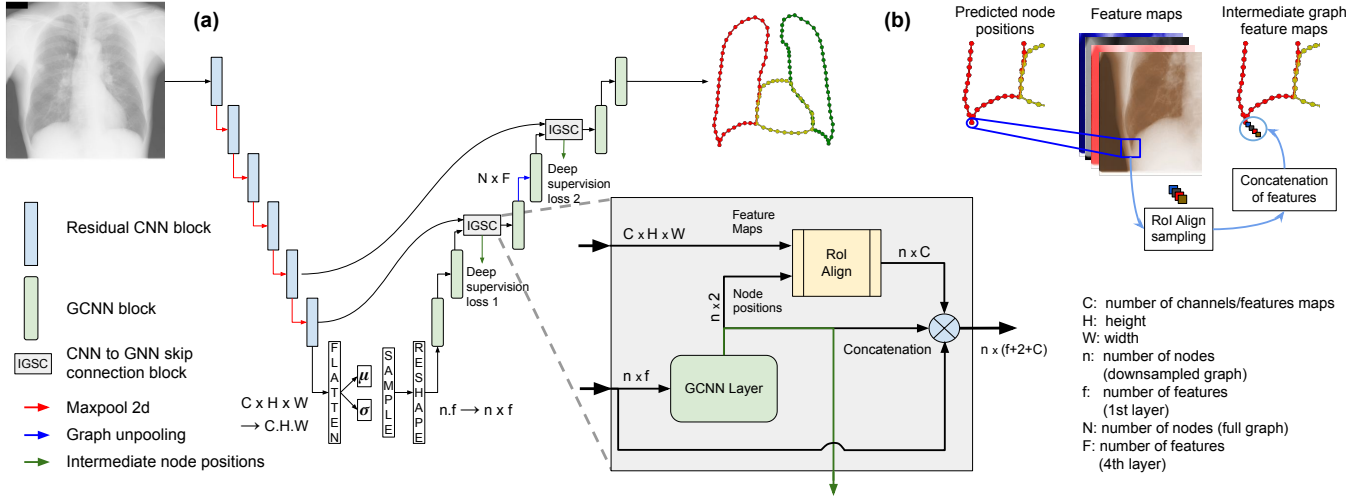


Fig. 1: HybridGNet architecture. (a) The proposed HybridGNet architecture combines standard convolutions for image feature encoding (cyan) with graph spectral convolutions (green) to decode plausible anatomical graph-based representations. After the input image is processed by the image encoder, we sample 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.

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.

III. ANATOMICAL SEGMENTATION VIA HYBRID GRAPH NEURAL NETWORKS

A. 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 $\mathbf{G} = \langle \mathbf{V}, \mathbf{A}, \mathbf{X} \rangle$ (see Section IV-A.1 for a detailed description of the graph construction for the heart and lungs structures used in this study). $\mathbf{V}$ 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 $\mathbf{V}$ 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 $\mathbf{g}_\phi * \mathbf{X} = \mathbf{g}_\phi \mathbf{L} \mathbf{X} = \mathbf{g}_\phi (\mathbf{U} \mathbf{A} \mathbf{U}^T) \mathbf{X} = \mathbf{U} \mathbf{g}_\phi \mathbf{A} \mathbf{U}^T \mathbf{X}$, 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 $\mathbf{g}_\phi = \sum_{k=0}^K \phi_k \mathbf{A}^k$. 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.

B. 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).

IV. EXPERIMENTAL SETUP

A. 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.

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]. 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 III-A 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.

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 providence, China.

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.

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

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 $f_e^{\mathcal{I}}$ to estimate the PCA coefficients, reconstructing the landmarks as a linear combination of the principal components.

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

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) *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 $f_e^{\mathcal{I}}$ and decoder $f_d^{\mathcal{I}}$ with standard skip-connections via concatenation, to guarantee comparable complexity.

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.

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 an ensemble model, averaging the predictions and incorporating standard post-processing techniques.

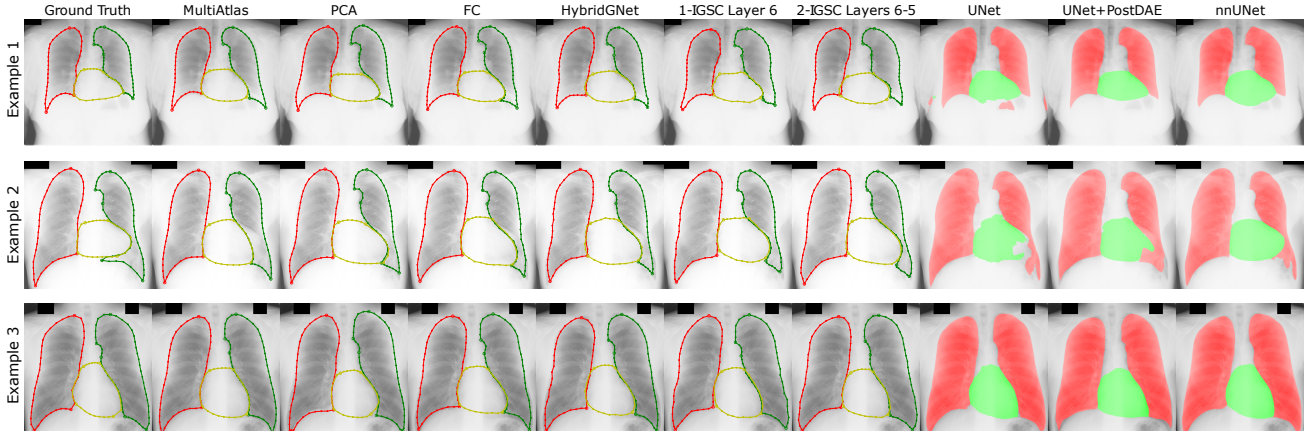
C. Implementation and training details

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 $f_e^{\mathcal{I}}$, 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

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 I: 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)	

**Fig. 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.

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 down-sampling 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.

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).

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.

V. 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.

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 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 I 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

TABLE II: 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

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 II 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.

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 IV-A.1 as an

TABLE III: 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 IV: 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)

example) or 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 III 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.

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 IV shows how HybridGNet model

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

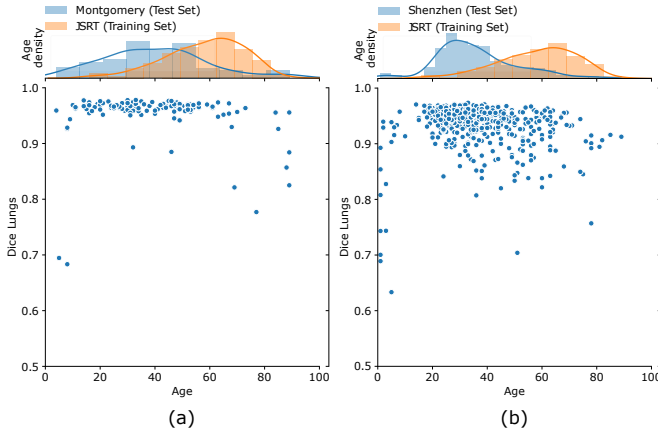


Fig. 3: Assessing the impact of domain shift by age distribution on lung segmentation. Scatter plot of the lung Dice coefficient vs. age of patient for every individual in both (a) Montgomery and (b) Shenzhen datasets. Histograms show the age distribution for test (blue) and training sets (orange).

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.

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.

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

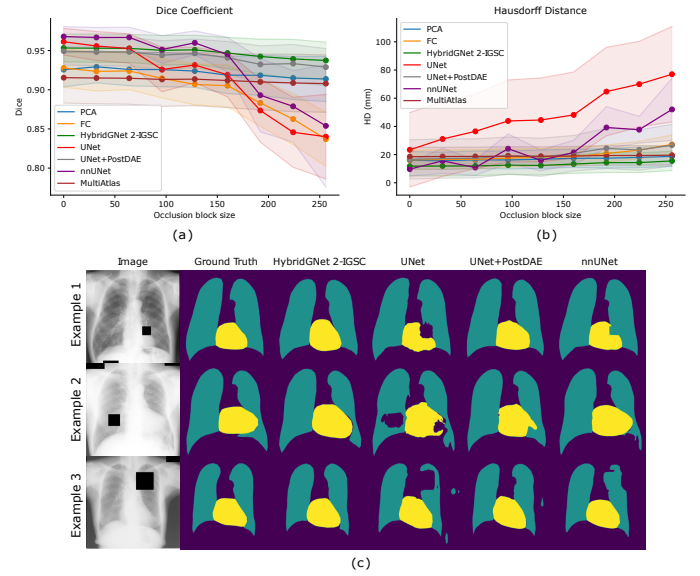


Fig. 4: Artificial occlusions study. (a) Dice and (b) HD distance for increasing block size in artificial occlusions. (c) Shows qualitative results.

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) *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

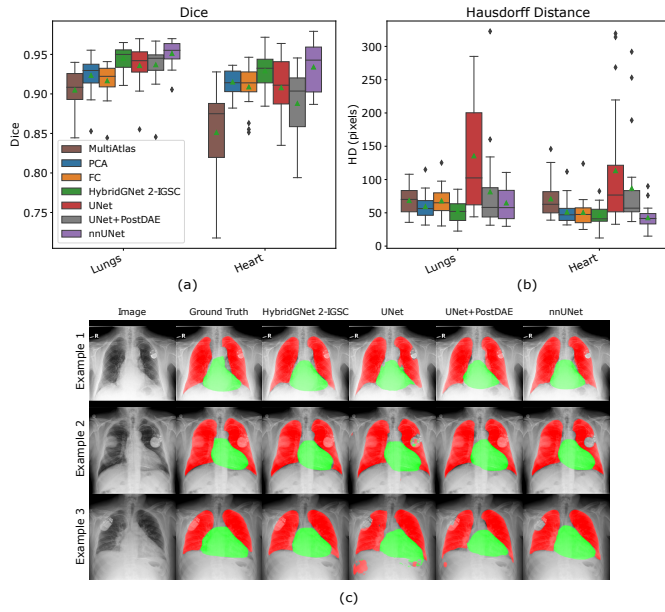


Fig. 5: Real occlusions study. (a) Dice and (b) HD distances for the pacemaker Padchest subset. Wilcoxon test showed no significant differences between HybridGNet and nnUNet, the best two performing models. However, qualitative results (c) show how HybridGNet results in more anatomically plausible segmentations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

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.

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 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 V-.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.

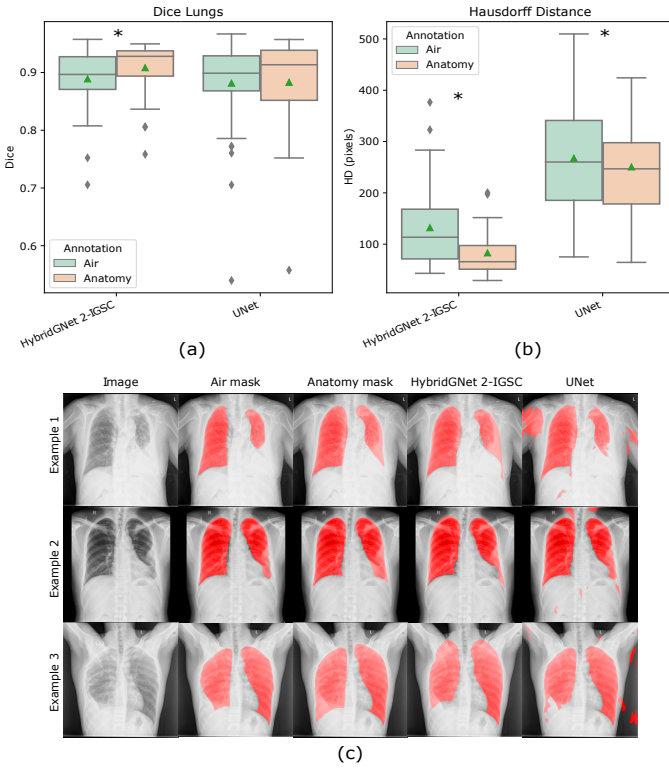


Fig. 6: Model behaviour on pathological anatomy. Top boxplots show Dice coefficient (a) and HD (b), taking the air mask as ground truth (green) and taking the anatomy mask as ground truth (orange) for the HybridGNet and UNet models. * indicates significant differences between means according to Wilcoxon's 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.

VI. 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,

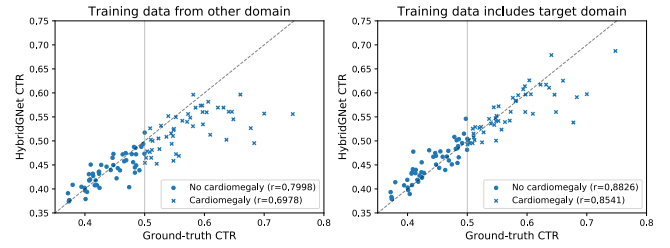


Fig. 7: CTR study. Ground-truth CTR vs HybridGNet CTR when training with JSRT dataset only (left) and the dataset augmented with cardiomegaly images from Padchest (right). The vertical line indicates the boundary between normal and abnormal CTR.

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.

VII. ACKNOWLEDGEMENTS

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. Koderia, 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 cardi thoracic ratio: an index of cardiac enlargement," *The American Journal of the Medical Sciences (1827-1924)*, vol. 157, no. 4, p. 513, 1919.