Improving anatomical plausibility and auditing fairness in deep segmentation networks

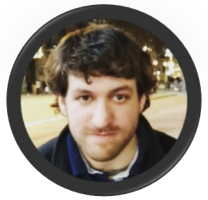
Enzo Ferrante

eferrante@sinc.unl.edu.ar

@enzoferrante

Research Institute for Signals, Systems and Computational Intelligence, sinc(i)
Argentina's National Research Council (CONICET), Universidad Nacional del Litoral (UNL)
Santa Fe, Argentina

DATAIA Visiting Professor at Université Paris-Saclay
Paris, France



Lucas Mansilla

Nicolás Gaggión

Rodrigo Echeveste

Diego Milone

Franco Matzkin

Agostina Larrazabal

Nicolás Nieto

Victoria Peterson

Candelaria Mosquera

Agustina Ricci

Rodrigo Bonazola

Josefina Catoni

Estanislao Claucich

Belen Bachli

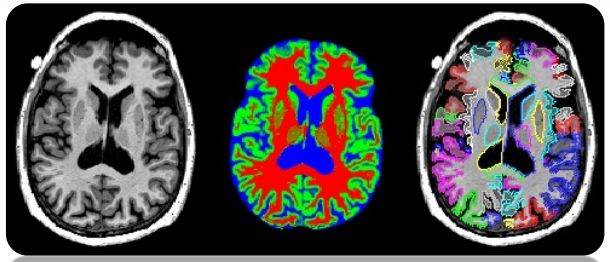
Machine Learning for Biological and Medical Image Computing

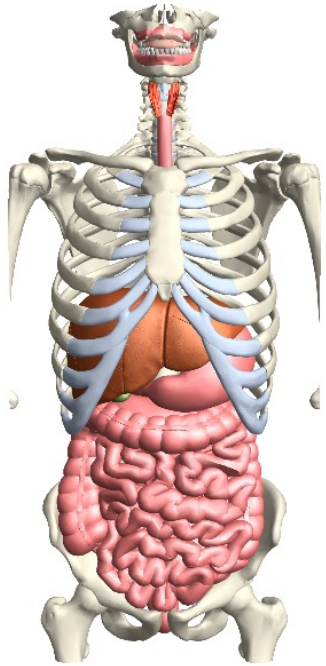
ML-BioMIC

UNL CONICET

s i n c (*i*)

Anatomical Plausibility

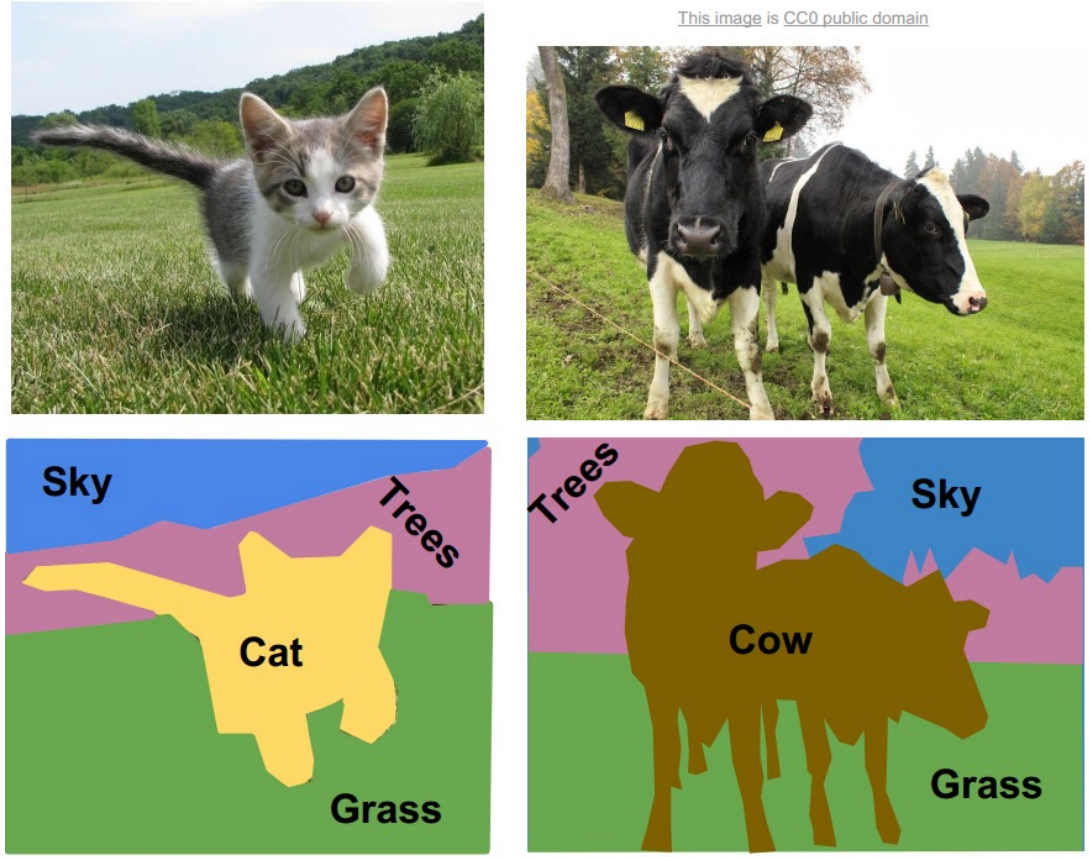




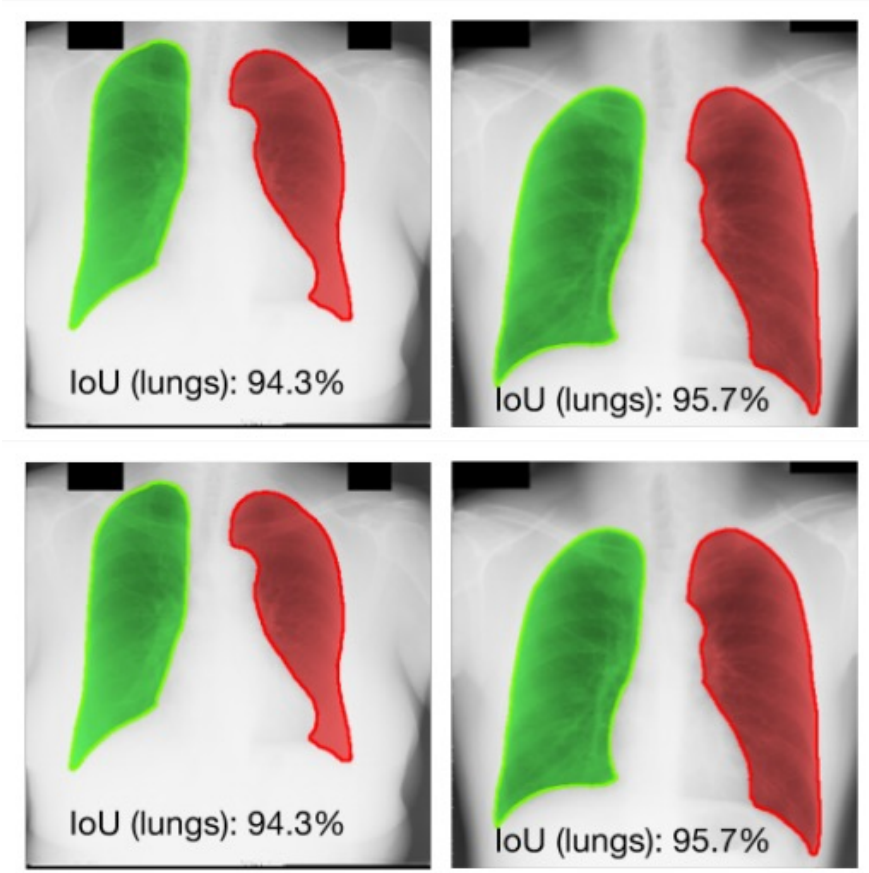
- Anatomical structures follow regular patterns
- Constrained space of solutions in terms of shape, topology and location
- We say that a segmentation mask is *anatomically plausible* if it lives in such constrained space.

Image source: https://www.wolfram.com/

Semantic segmentation 'in the wild'
vs
medical imaging segmentation

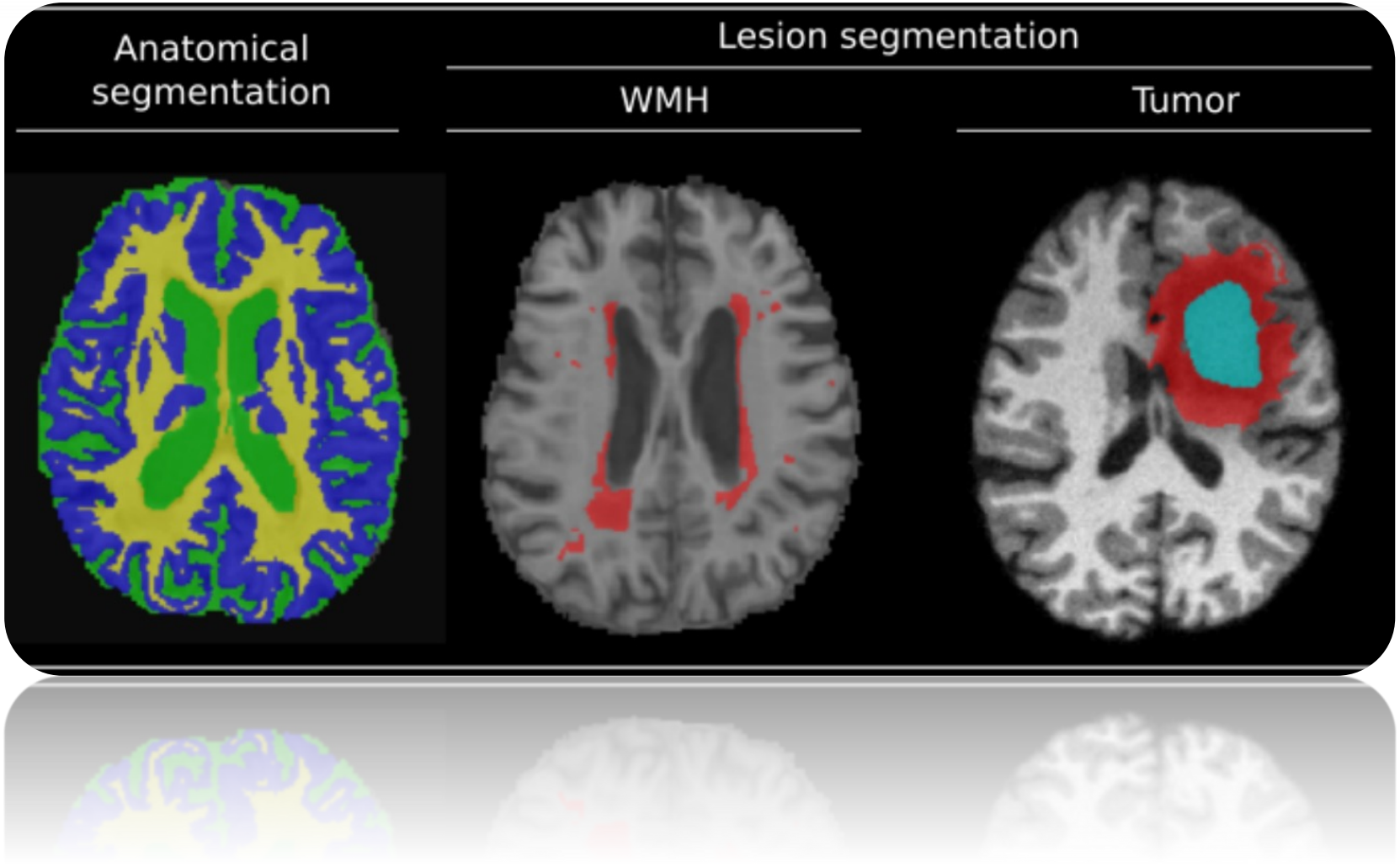


Tons of data, low regularity

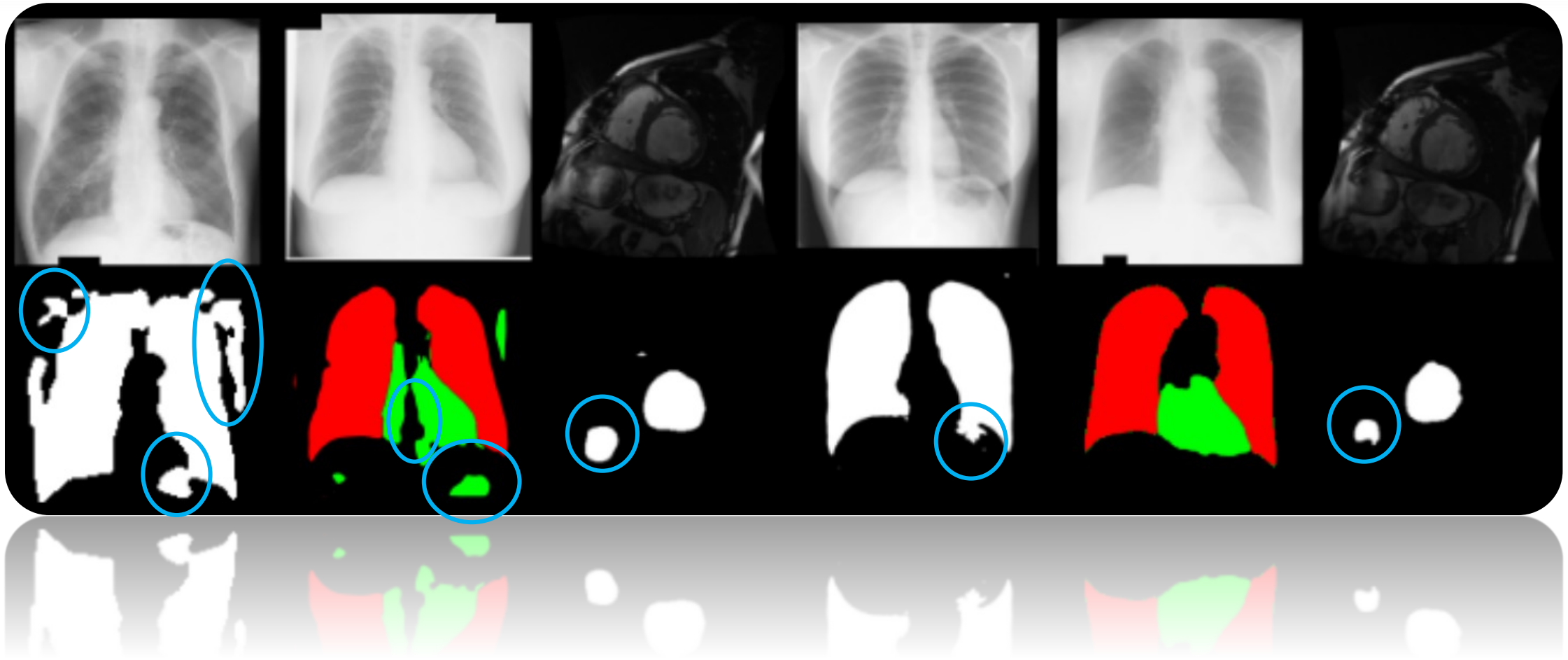


Less data, more regularity

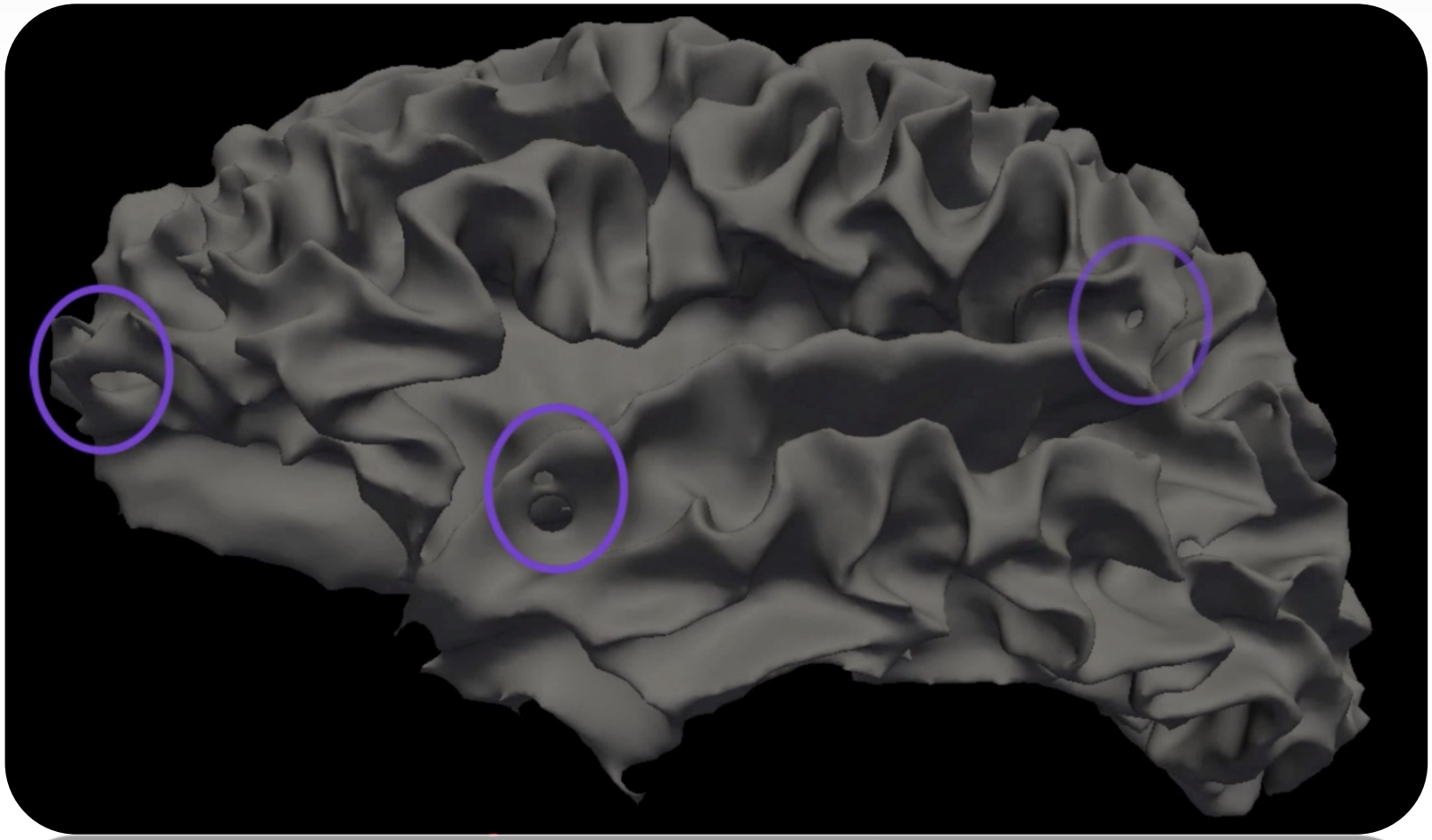
Anatomical Plausibility



Anatomical plausibility in image segmentation



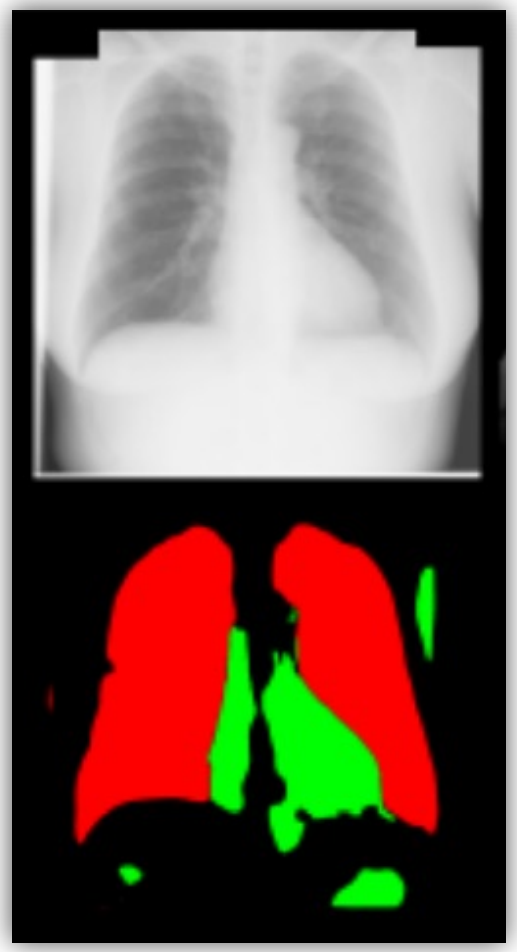
Anatomical plausibility in image segmentation



Example of topological defects in the **white matter surface** extracted by FreeSurfer

Source: https://andysbrainbook.readthedocs.io/en/latest/FreeSurfer/FS_ShortCourse/FS_12_FailureModes.html

Why does it happen?
CNN limitations



- CNN predictions have **local support**
- CNNs are **translation invariant**
- The **loss term** is usually defined at the **pixel level**
- This makes it **difficult** to introduce **global shape constraints**

Can we take advantage of this high data regularity to encourage anatomical plausibility via global shape constraints?

Anatomically Constrained Neural Networks

Imperial College London

Anatomically Constrained Neural Networks (ACNNs): Application to Cardiac Image Enhancement and Segmentation

Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A. Cook, Antonio de Marvao, Timothy Dawes, Declan P. O'Regan, Bernhard Kainz, Ben Glocker, and Daniel Rueckert

***Abstract*— Incorporation of prior knowledge about organ shape and location is key to improve performance of image analysis approaches. In particular, priors can be useful in cases where images are corrupted and contain artefacts due to limitations in image acquisition. The highly constrained nature of anatomical objects can be well captured with learning-based techniques. However, in most recent and promising techniques such as CNN-based segmentation it is not obvious how to incorporate such prior knowl-**

state-of-the-art methods operate as pixel-wise clas-

jectives do not incorporate the

underlying data and a prior on the solution space, where the latter is useful in cases where the images are corrupted or contain artefacts due to limitations in the image acquisition. For example, bias fields, shadowing, signal drop-out, respiratory motion, and low-resolution acquisitions are the few common limitations in ultrasound (US) and magnetic resonance (MR) imaging.

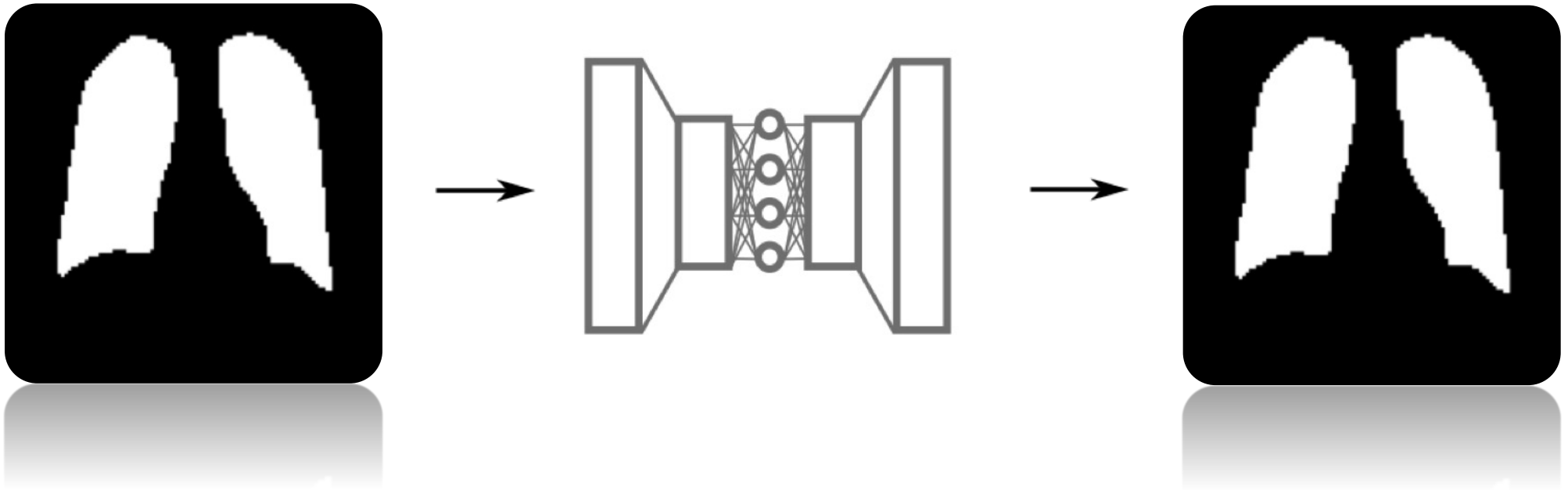
Incorporating prior knowledge into image segmentation algorithms has proven useful in order to obtain more accurate and plausible results as summarised in the recent survey [32]. Prior information can take many forms: boundaries and edge [13], [14]; topology specification;

Anatomically Constrained Neural Networks

Idea: Learn an embedding which contains global information about shape and topology

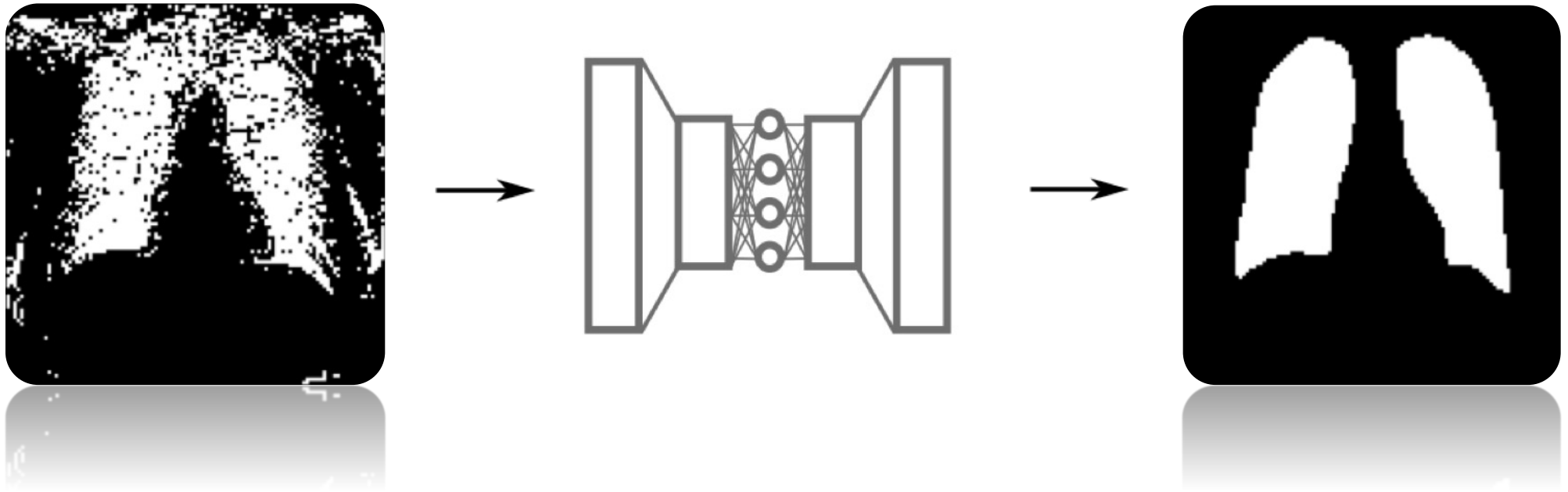
→ (d1, d2, d3, …….. dn)

Learning embeddings of anatomical masks



Trained using segmentation masks (not image information)

Learning embeddings of anatomical masks



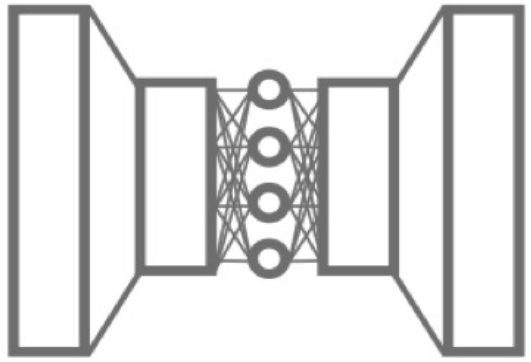
Trained using segmentation masks (not image information)

Denoising autoencoders

Encoder : f_θ g_θ **: Decoder**

$\hat{x}$

Noisy Image



Denoised Image

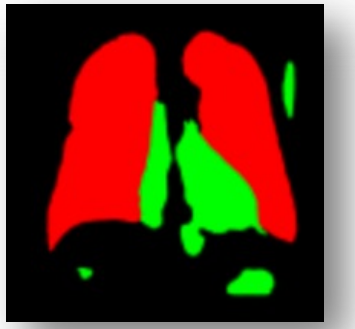
Embedding: $h = f_\theta(\hat{x})$

Noisy Image / Mask

E.g: MSE Loss Function $L(x;\theta) = (g_\theta(f_\theta(\hat{x})) - x)^2$

Decoder Encoder Original Image / Mask

Measuring similarity at the local and global level

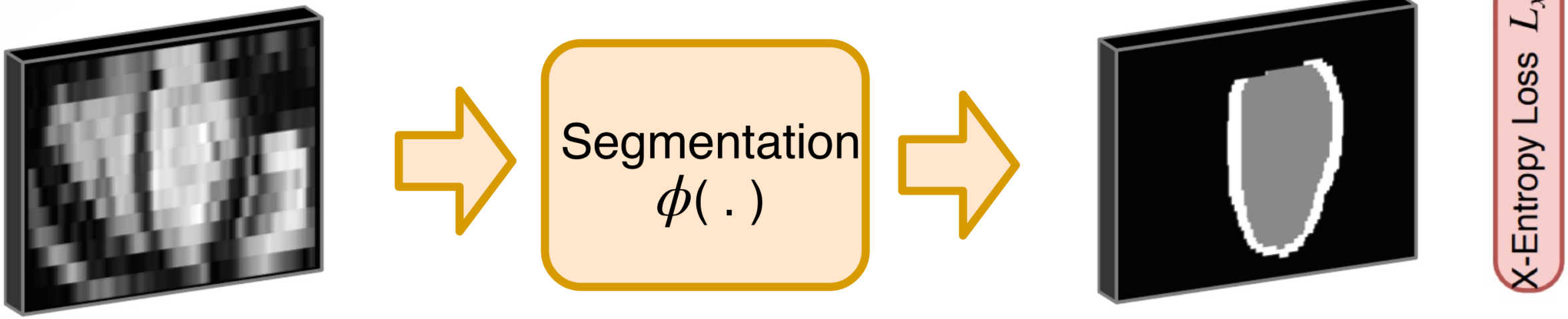


- **Local loss:** Cross entropy defined at the pixel level. ---> $\mathcal{L}_{ce}$

- **Global loss:** Euclidean distance between the embeddings

$$\mathcal{L}_{ae} = \| f(\phi(x); \boldsymbol{\theta}_f) - f(\hat{x}; \boldsymbol{\theta}_f) \|_2^2$$

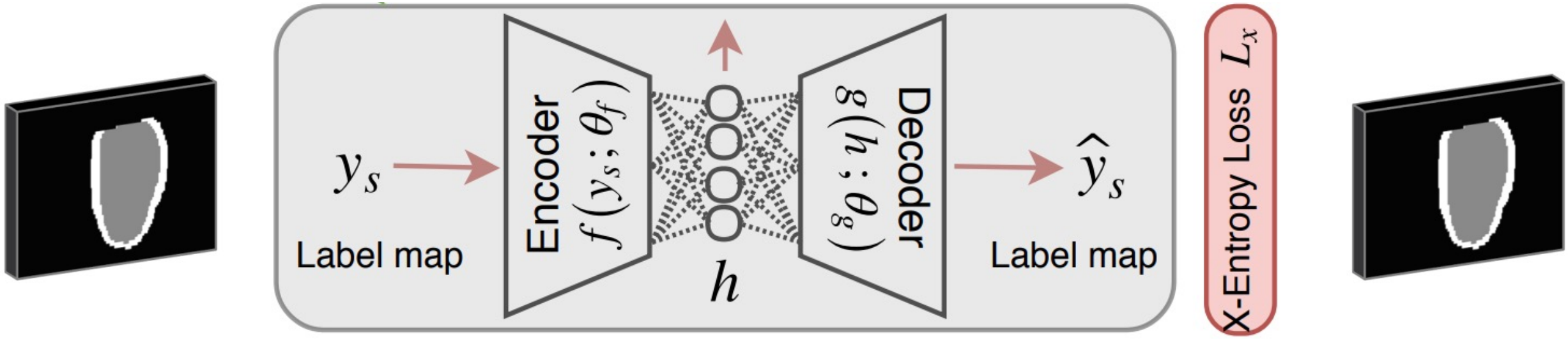
ACNN for image segmentation



Source: Oktay, Ferrante, Kamnitsas et al (IEEE TMI, 2018)

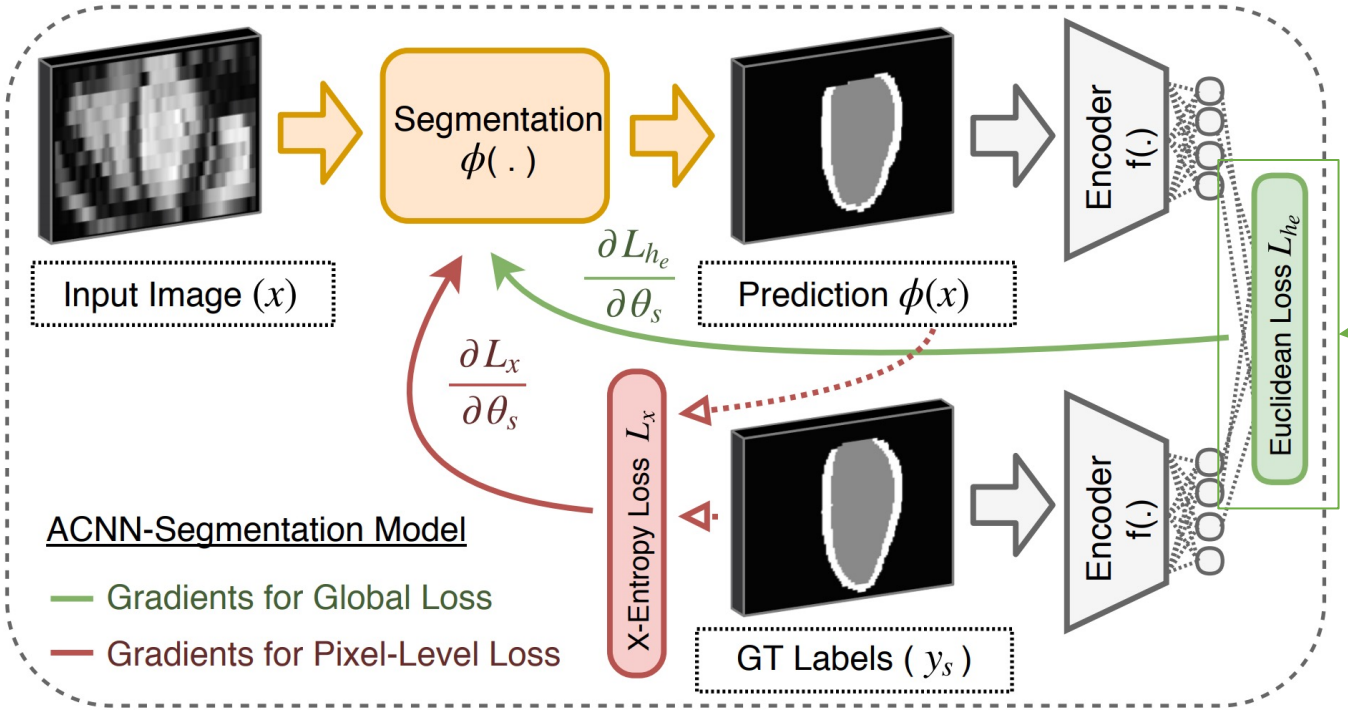
ACNN for image segmentation

Step 1: Train the autoencoder



Source: Oktay, Ferrante, Kamnitsas et al (IEEE TMI, 2018)

ACNN for image segmentation



Global loss

$$L_{h_e} = \| f(\phi(\boldsymbol{x}); \boldsymbol{\theta}_f) - f(\boldsymbol{y}; \boldsymbol{\theta}_f) \|_2^2$$

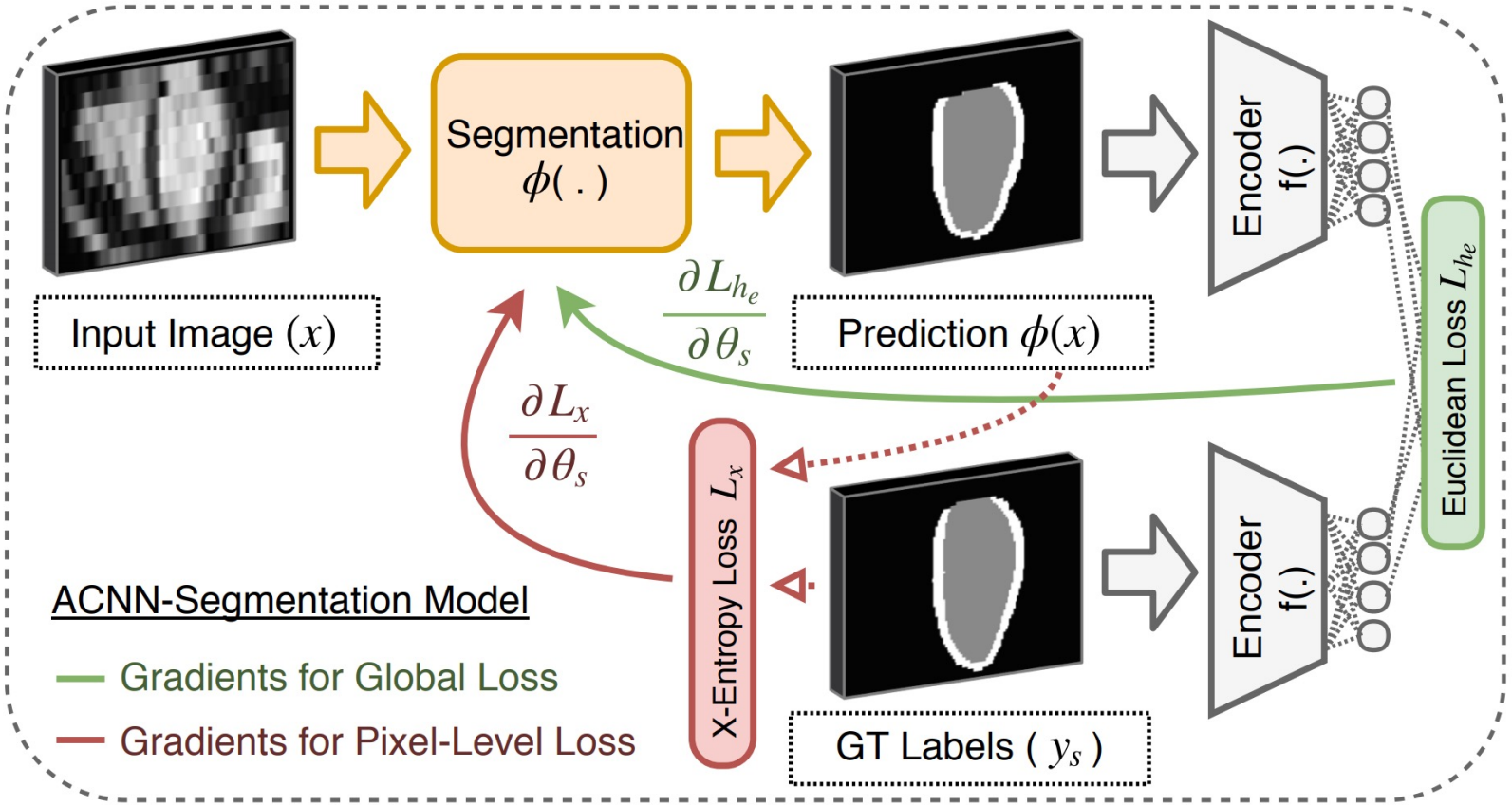
Cross entropy at the pixel level

$$\min_{\boldsymbol{\theta}_s} \left(L_x \left(\phi(\boldsymbol{x}; \boldsymbol{\theta}_s), \boldsymbol{y}\right) + \lambda_1 \cdot L_{h_e} + \frac{\lambda_2}{2} ||\boldsymbol{w}||_2^2 \right)$$

Source: Oktay, Ferrante, Kamnitsas et al (IEEE TMI, 2018)

ACNN for image segmentation

Step 2: Use the embeddings to formulate the loss function



Source: Oktay, Ferrante, Kamnitsas et al (IEEE TMI, 2018)

ACNN for image segmentation

MR Segmentation

US Segmentation

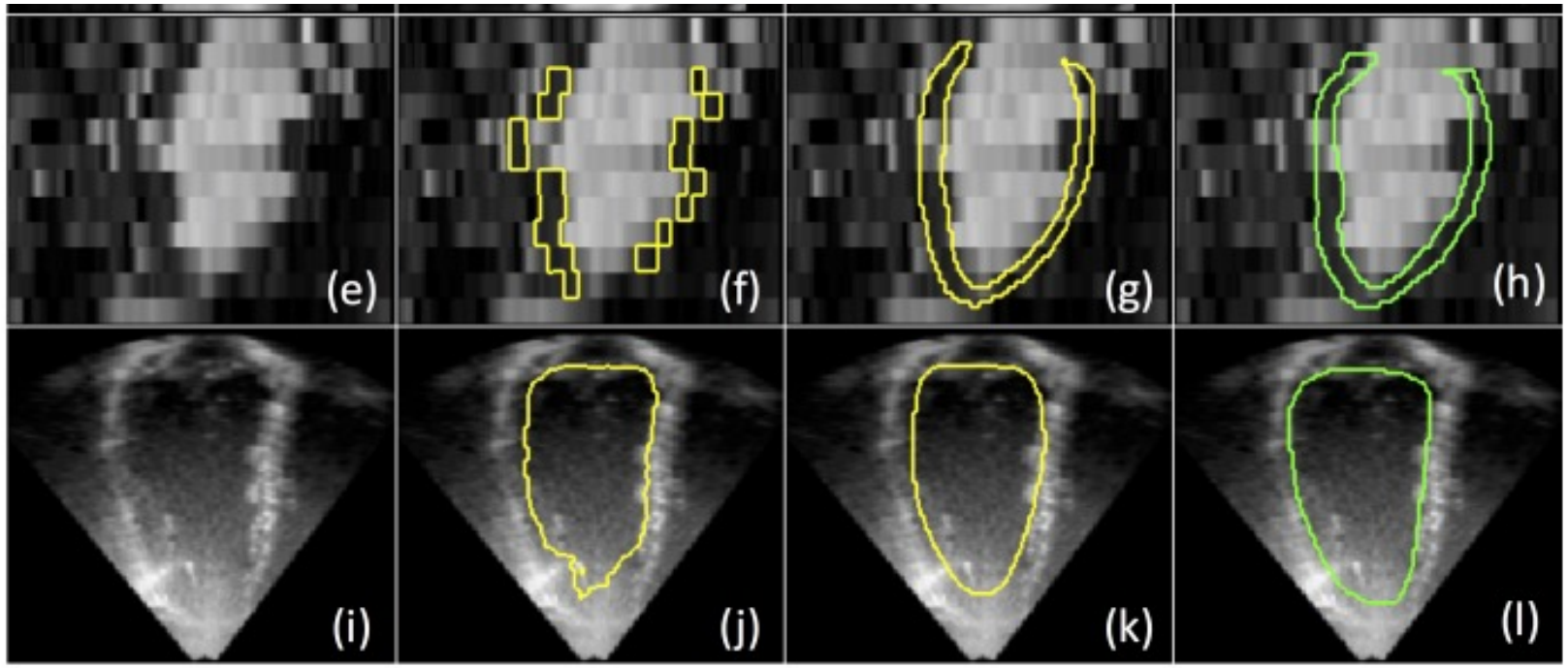


Image UNet ACNN GT

ACNN for image segmentation

	Endocardium			Myocardium			Capacity
	Mean Dist. (mm)	Hausdorff Dist. (mm)	Dice Score (%)	Mean Dist. (mm)	Hausdorff Dist. (mm)	Dice Score (%)	# Trainable Parameters
2D-FCN [44]	2.07±0.61	11.37±7.15	.908±.021	1.58±0.44	9.19±7.22	.727±.046	1.39×10^6
3D-Seg	1.77±0.84	10.28±8.25	.923±.019	1.48±0.51	10.15±10.58	.773±.038	1.60×10^6
3D-UNet [12]	1.66±0.74	9.94±9.22	.923±.019	1.45±0.47	9.81±11.77	.764±.045	1.64×10^6
AE-Seg [37]	1.75±0.58	8.42±3.64	.926±.019	1.51±0.29	8.52±2.72	.779±.033	1.68×10^6
3D-Seg-MAug	1.59±0.74	8.52±8.13	.928±.019	1.37±0.41	9.41±9.17	.785±.041	1.60×10^6
AE-Seg-M	1.59±0.48	**7.52±3.78**	.927±.017	1.32±0.26	**7.12±2.79**	.791±.036	1.91×10^6
ACNN-Seg	**1.37±0.42**	7.89±3.83	**.939±.017**	**1.14±0.22**	7.31±3.59	**.811±.027**	1.60×10^6
p-values	$p \ll 0.001$	$p \approx 0.890$	$p \ll 0.001$	$p \ll 0.001$	$p \approx 0.071$	$p \ll 0.001$	-

Can we use ACNN to encourage anatomical plausibility in Deformable Image Registration?

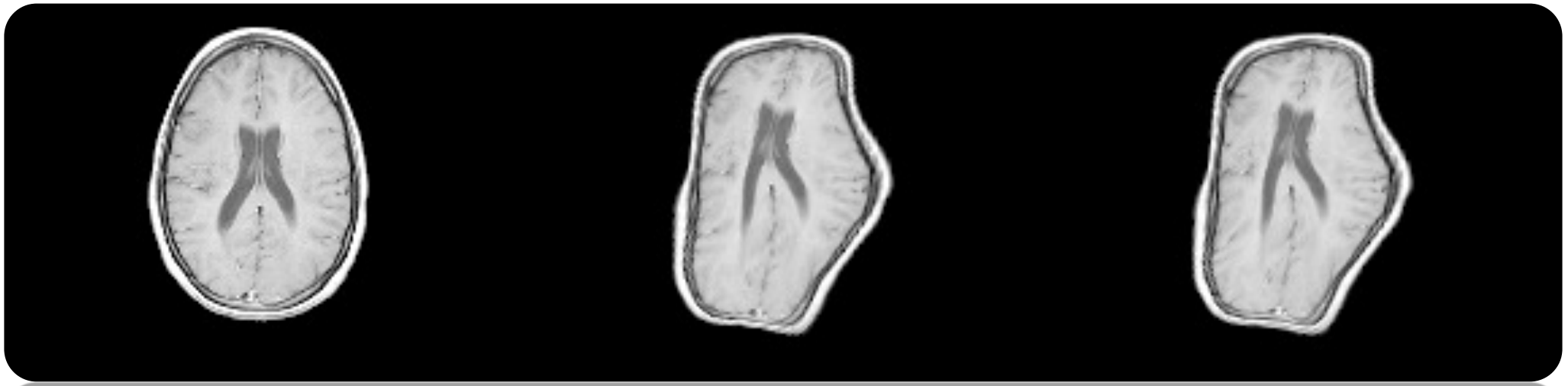


Image source: https://www.mathworks.com/matlabcentral/fileexchange/50702-fordanic-image-registration

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2020 Special Issue

Learning deformable registration of medical images with anatomical constraints

Lucas Mansilla, Diego H. Milone, Enzo Ferrante *

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

ARTICLE INFO

Article history:
Available online 30 January 2020

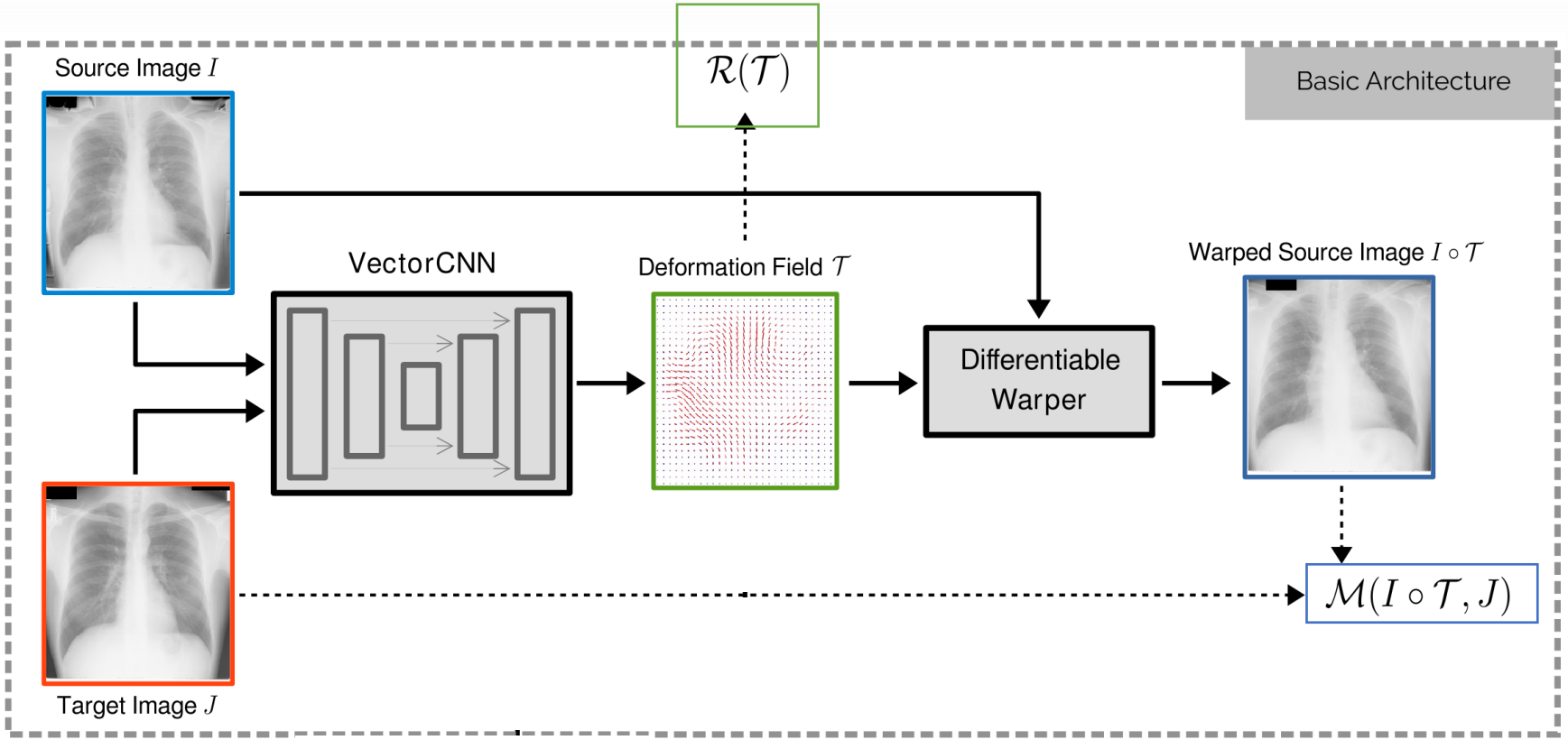
Keywords:
Medical image registration
Convolutional neural networks
X-ray image analysis

ABSTRACT

Deformable image registration is a fundamental problem in the field of medical image analysis. During the last years, we have witnessed the advent of deep learning-based image registration methods which achieve state-of-the-art performance, and drastically reduce the required computational time. However, little work has been done regarding how can we encourage our models to produce not only accurate, but also anatomically plausible results, which is still an open question in the field. In this work, we argue that incorporating anatomical priors in the form of global constraints into the learning process of these models, will further improve their performance and boost the realism of the warped images after registration. We learn global non-linear representations of image anatomy using segmentation masks, and employ them to constraint the registration process. The proposed AC-RegNet architecture is evaluated in the context of chest X-ray image registration using three different datasets, where the high anatomical variability makes the task extremely challenging. Our experiments show that the proposed anatomically constrained registration model produces more realistic and accurate

demonstrating the potential of this approach.
ltd. All rights reserved.

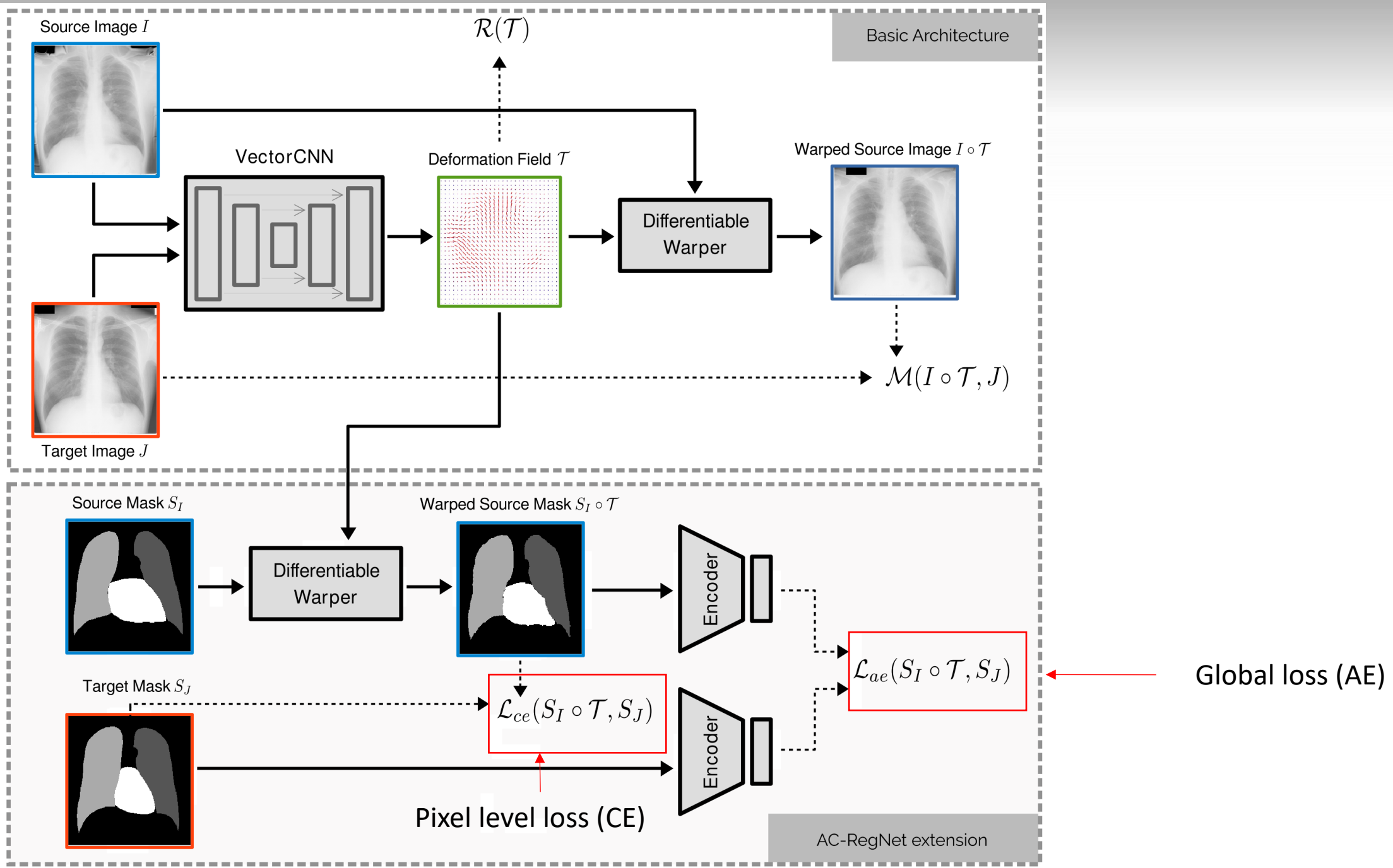
Learning deformable image registration with anatomical constraints



$$\mathcal{L}(I, J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T})$$

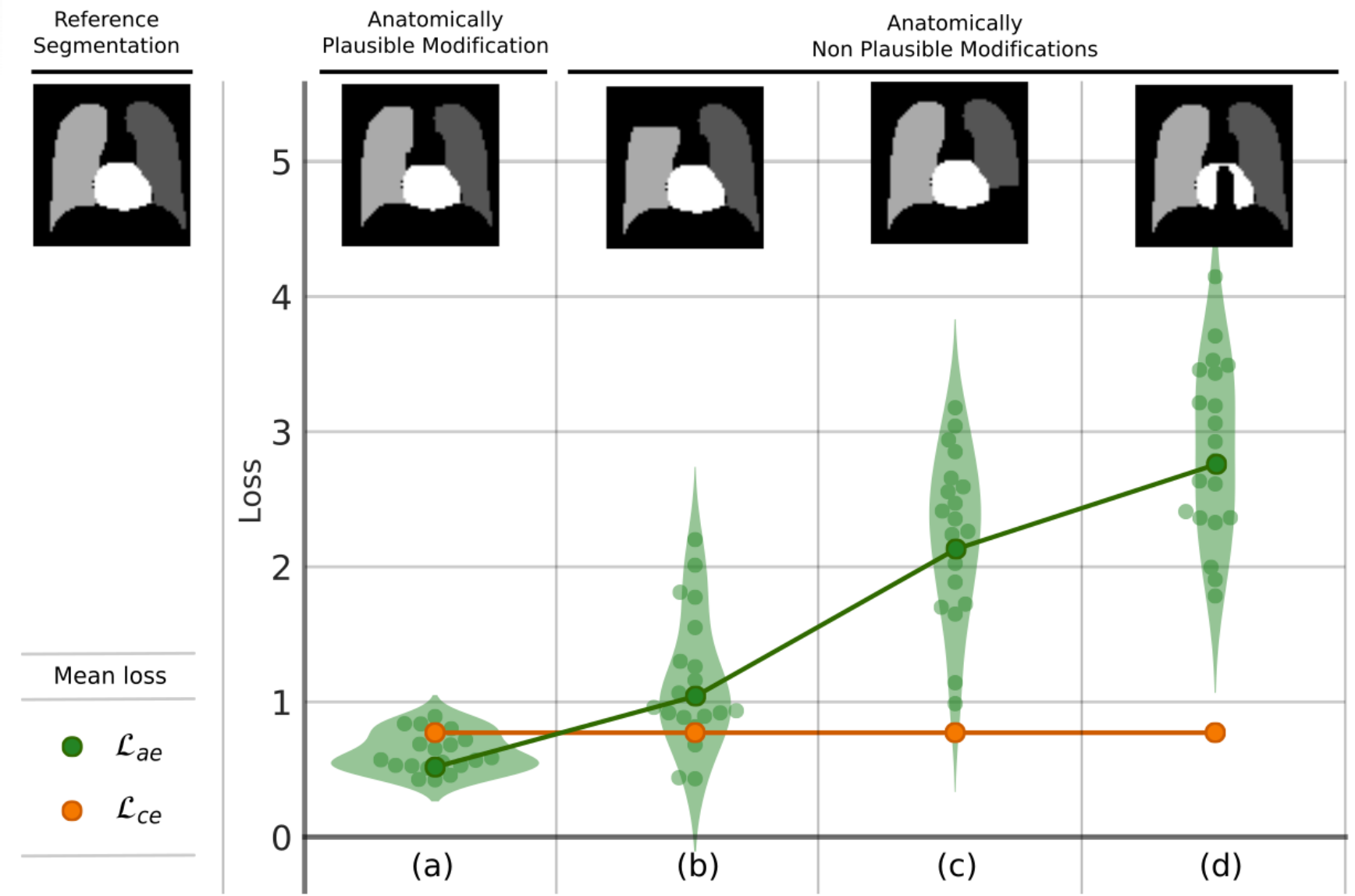
Similarity measure

Regularization term



Global loss (AE)

Measuring similarity at the local and global level



Mansilla, Milone & Ferrante
Neural Networks, 2020

Post-DAE: Autoencoders as a post processing step

Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders

Agostina J Larrazabal, César Martínez, Ben Glocker, Enzo Ferrante

***Abstract*— We introduce Post-DAE, a post-processing method based on denoising autoencoders (DAE) to improve the anatomical plausibility of arbitrary biomedical image segmentation algorithms. Some of the most popular segmentation methods (e.g. based on convolutional neural networks or random forest classifiers) incorporate additional post-processing steps to ensure that the resulting masks fulfill expected connectivity constraints. These methods operate under the hypothesis that contiguous pixels with similar aspect should belong to the same class. Even if valid in general, this assumption does not consider more complex priors like topological restrictions or convexity, which cannot be easily incorporated into these methods. Post-DAE leverages the latest developments in manifold learning via denoising autoencoders. First, we**

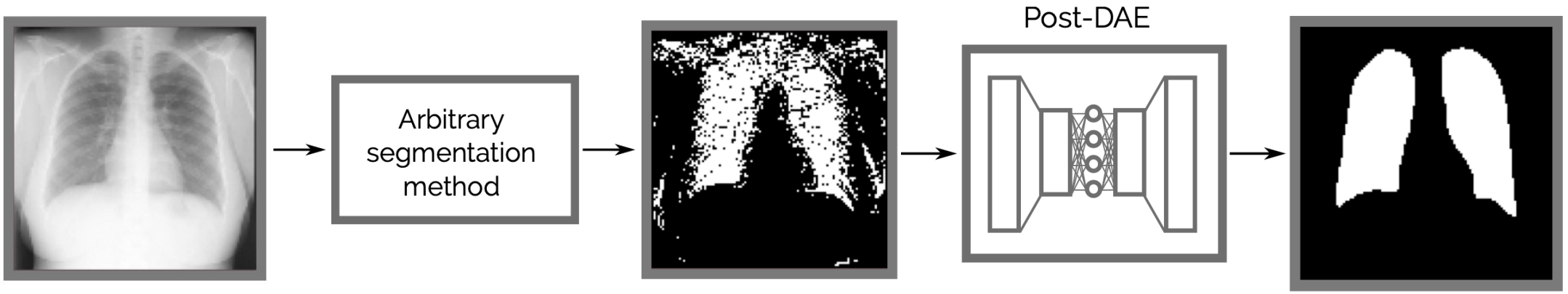
pipelines such as computer assisted diagnosis, morphometric analysis for population studies and radiotherapy planning. The correctness and anatomical plausibility of these results is thus of paramount importance, since it will directly influence the overall quality of subsequent analyses.

Convolutional neural networks (CNNs) proved to perform biomedical image segmentation in a highly accurate way [1]–[3]. CNNs constitute a particular type of neural network specially suited for regularly structured data, like 2D or 3D images, where hierarchical representations of the input are learned using stacked convolutional layers. At every layer, shared parameters (also referred as weights or kernel) are used to learn new representations of the input image. This sharing scheme reduces the number of parameters that should be learnt

In Collaboration with

Imperial College London

Autoencoders as a post-processing step



Denoising autoencoders

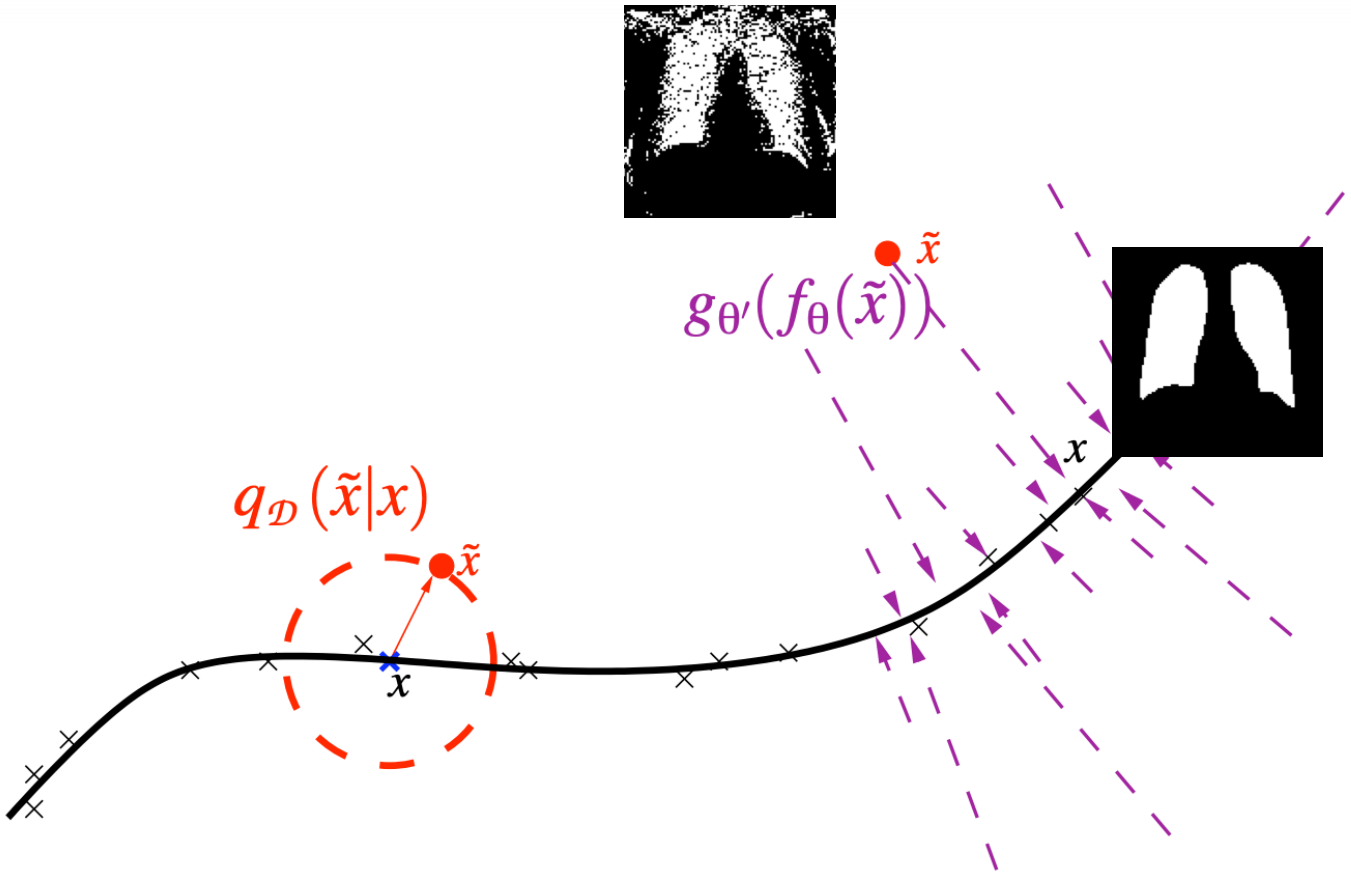
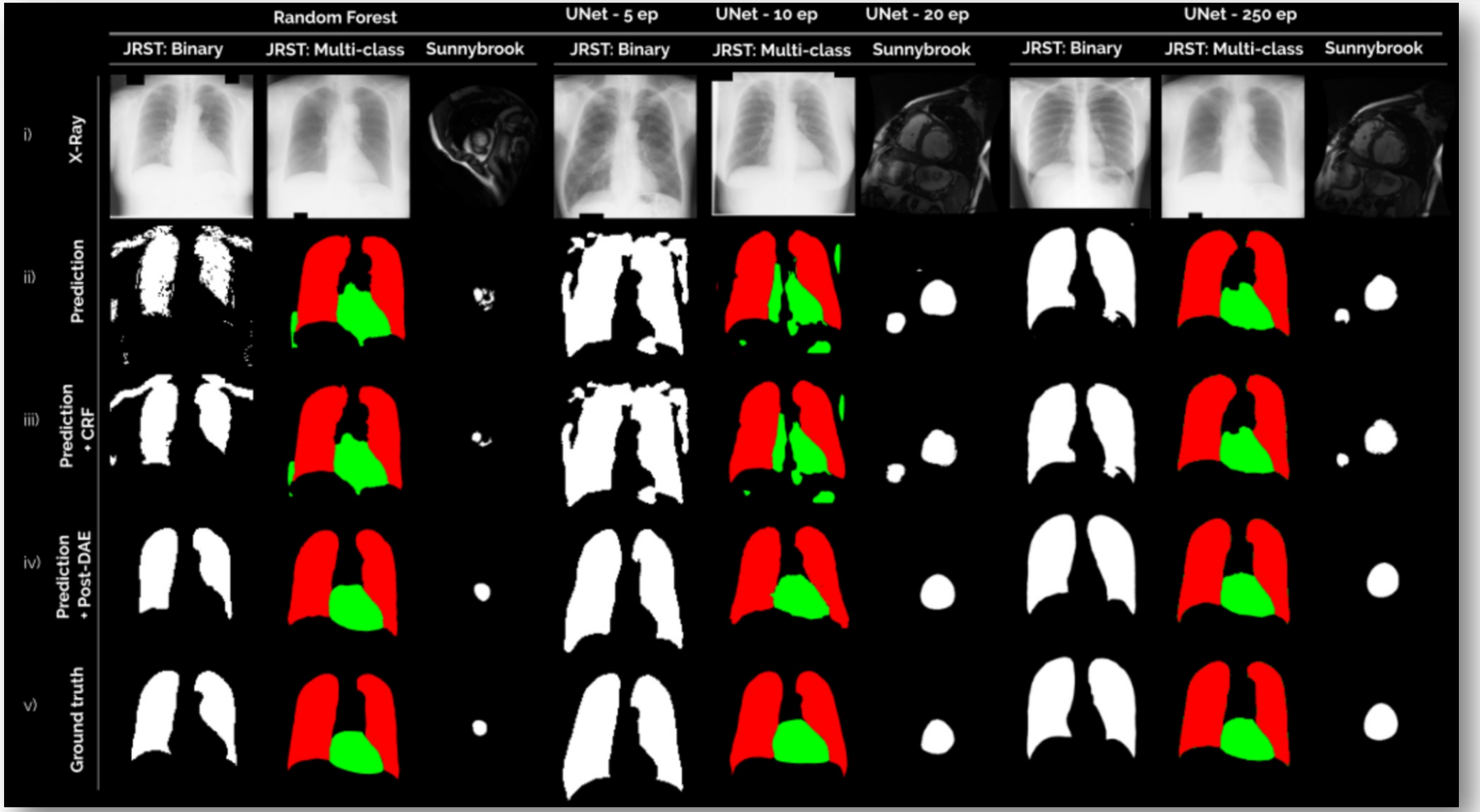
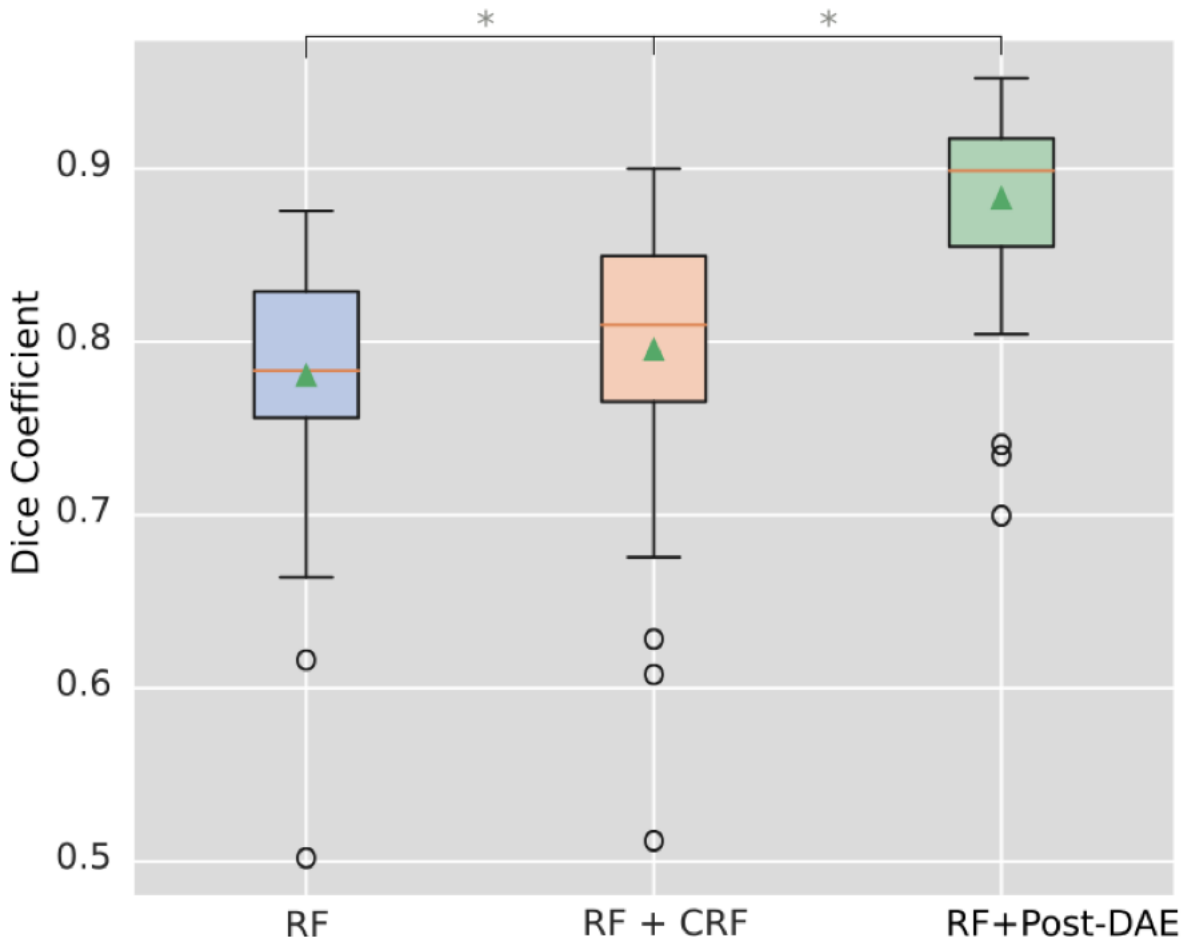


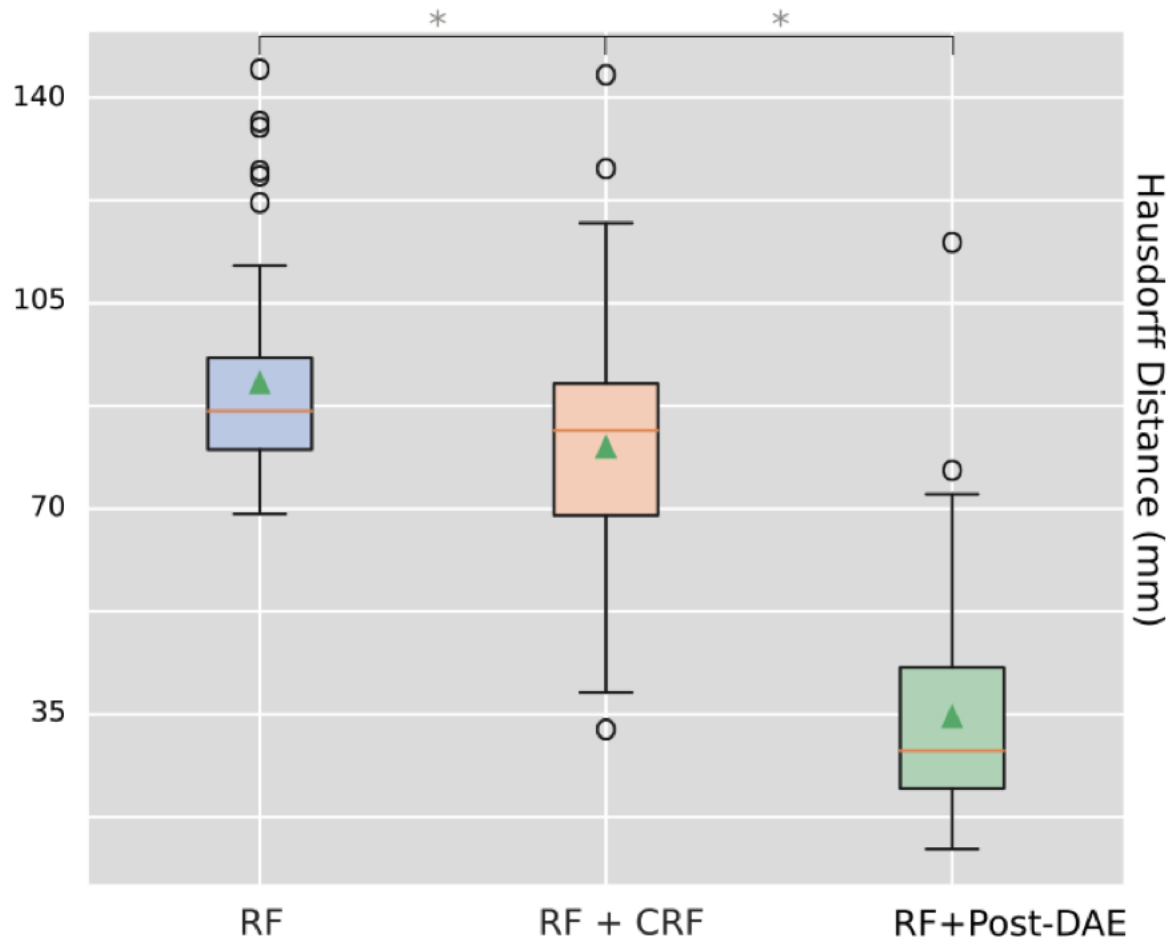
Image from: Vincent et al, 2010 (JMLR)

Experiments: quantitative results



Experiments: quantitative results





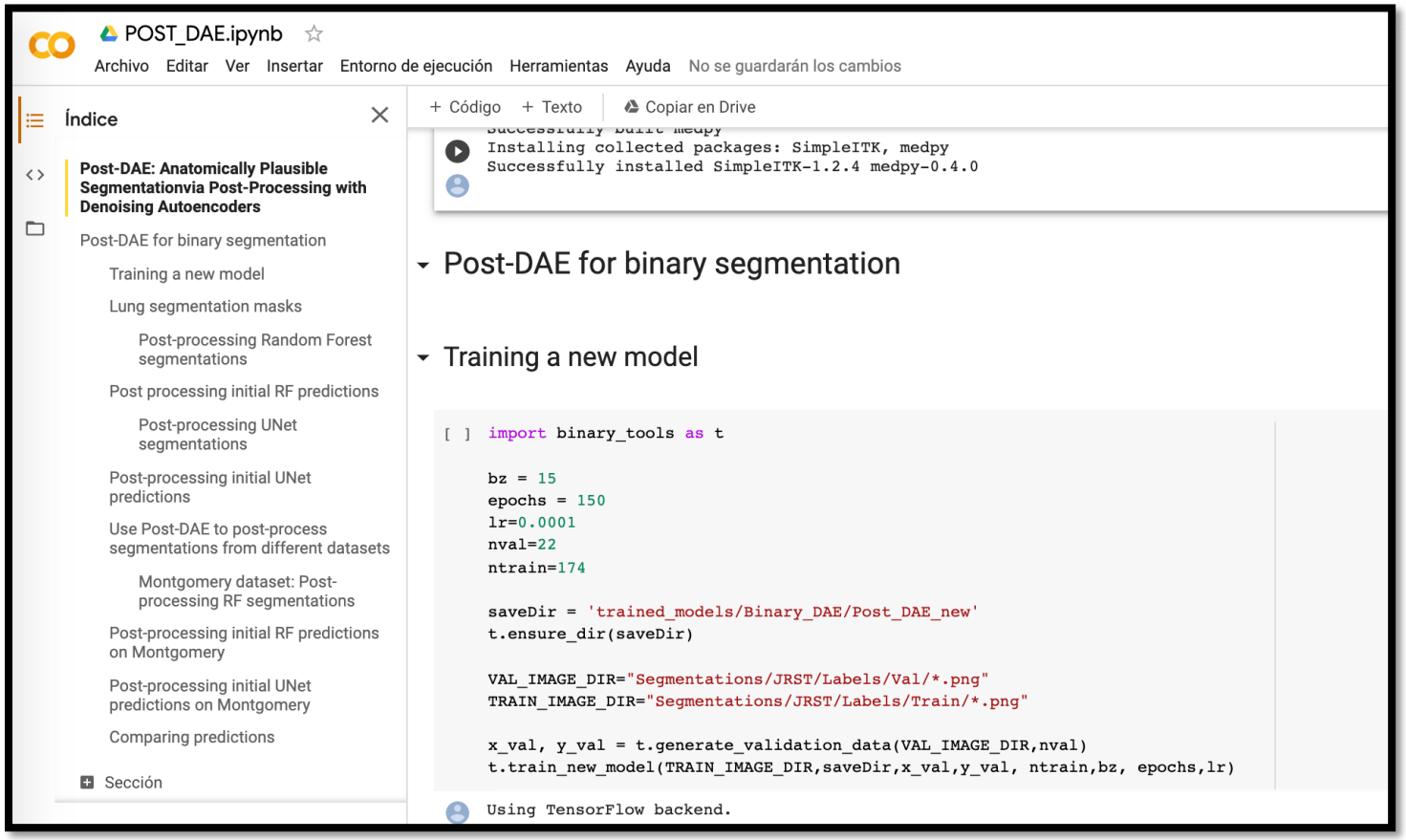
Anatomical Priors for Image Segmentation via Post-Processing with Denoising Autoencoders

MICCAI 2019

Visualization of segmentation masks before and after post-processing using Post-DAE

Source code

Agostina Larrazabal



https://tinyurl.com/postdae-segmentation

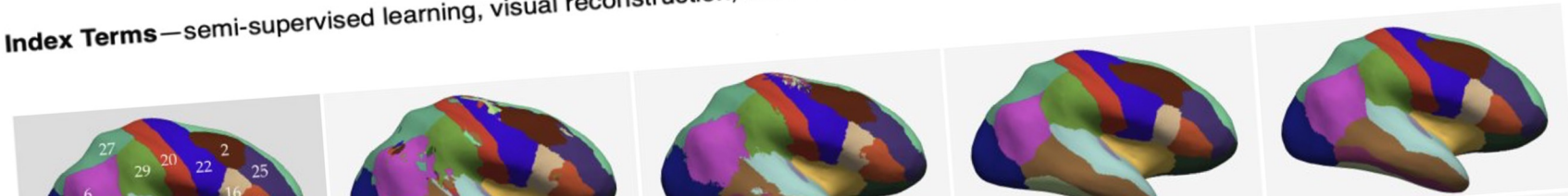
Denoising also helps during training

Supervision by Denoising

Sean I. Young, Adrian V. Dalca, Enzo Ferrante, Polina Golland, Christopher A. Metzler, Bruce Fischl, and Juan Eugenio Iglesias

Abstract—Learning-based image reconstruction models, such as those based on the U-Net, require a large set of labeled images if good generalization is to be guaranteed. In some imaging domains, however, labeled data with pixel- or voxel-level label accuracy are scarce due to the cost of acquiring them. This problem is exacerbated further in domains like medical imaging, where there is no single ground truth label, resulting in large amounts of repeat variability in the labels. Therefore, training reconstruction networks to generalize better by learning from both labeled and unlabeled examples (called semi-supervised learning) is problem of practical and theoretical interest. However, traditional semi-supervised learning methods for image reconstruction often necessitate handcrafting a differentiable regularizer specific to some given imaging problem, which can be extremely time-consuming. In this work, we propose "supervision by denoising" (SUD), a framework that enables us to supervise reconstruction models using their own denoised output as soft labels. SUD unifies stochastic averaging and spatial denoising techniques under a spatio-temporal denoising framework and alternates denoising and model weight update steps in an optimization framework for semi-supervision. As example applications, we apply SUD to two problems arising from biomedical imaging—anatomical brain reconstruction (3D) and cortical parcellation (2D)—to demonstrate a significant improvement in the image reconstructions over supervised-only and stochastic averaging baselines.

Index Terms—semi-supervised learning, visual reconstruction, denoising, fully convolutional networks, proximal methods.



In Collaboration with

MASSACHUSETTS GENERAL HOSPITAL

HARVARD MEDICAL SCHOOL

Massachusetts Institute of Technology

Young et al, 2023
IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE T-PAMI)

Can we extend these ideas to landmark-based segmentation?

Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

Nicolás Gaggion, Lucas Mansilla, Diego Milone, Enzo Ferrante

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

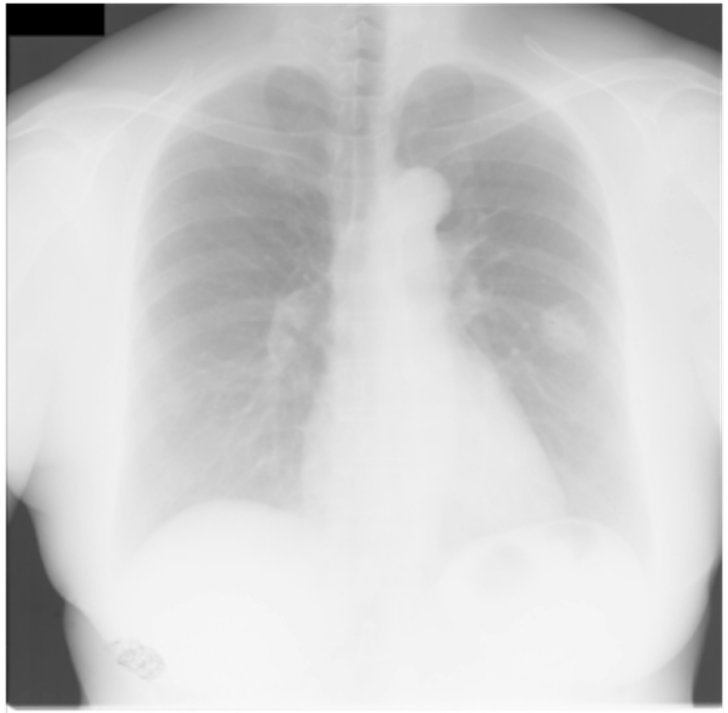
Abstract. In this work we address the problem of landmark-based segmentation for anatomical structures. We propose HybridGNet, an encoder-decoder neural architecture which combines standard convolutions for image feature encoding, with graph convolutional neural networks to decode plausible representations 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.

Keywords: Landmark-based segmentation · Graph convolutional neural networks · Spectral convolutions.

MICCAI 2021

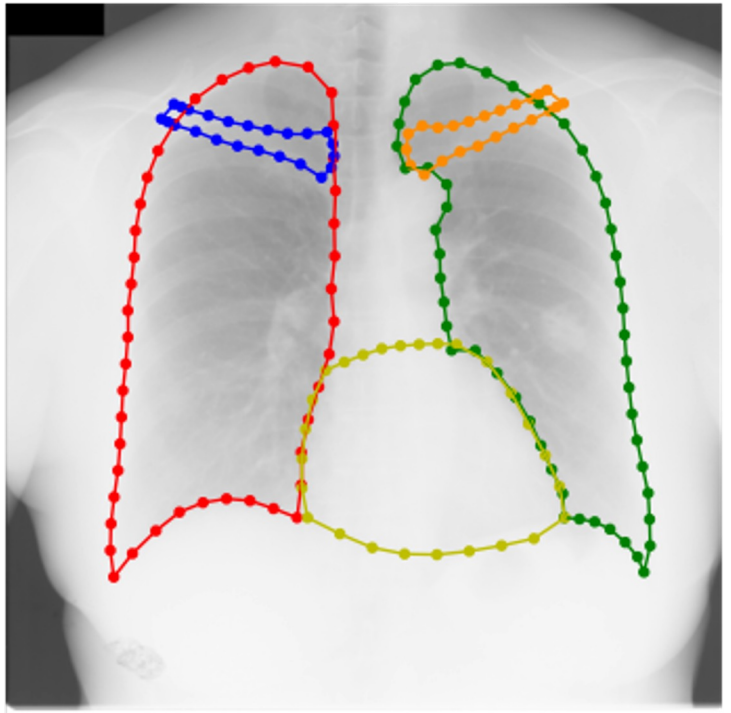
Landmark based anatomical segmentation

Image

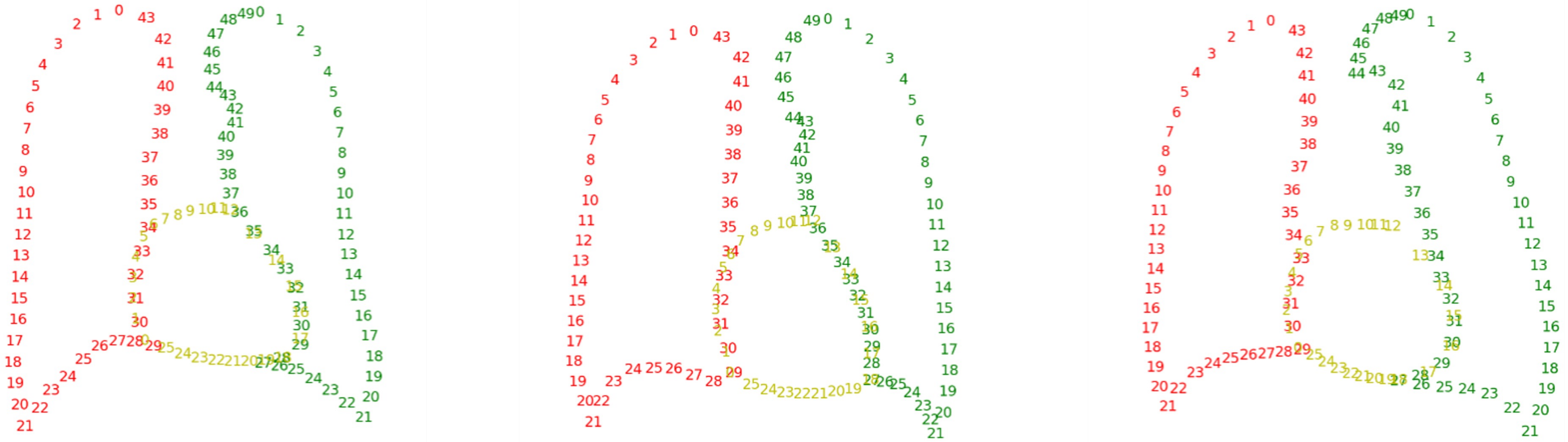


Segmentation

Landmarks

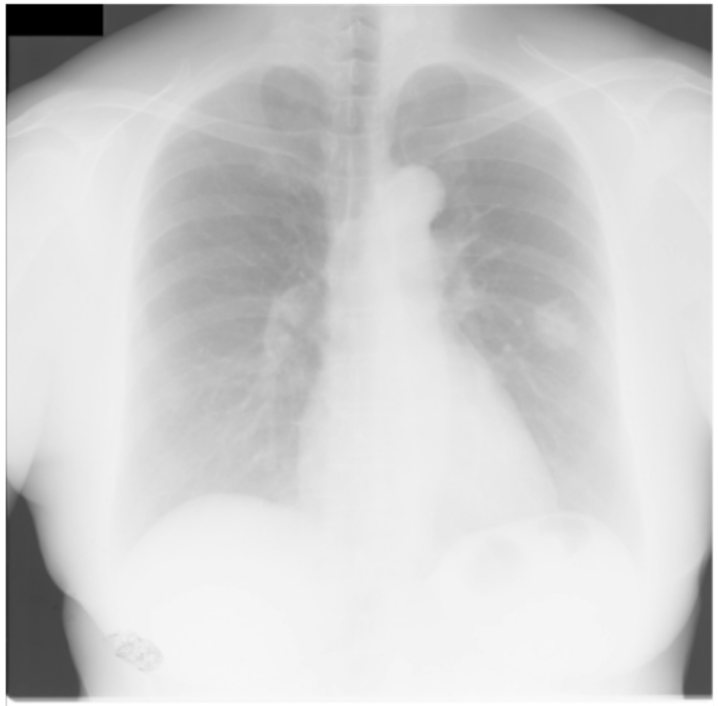


Landmark based anatomical segmentation

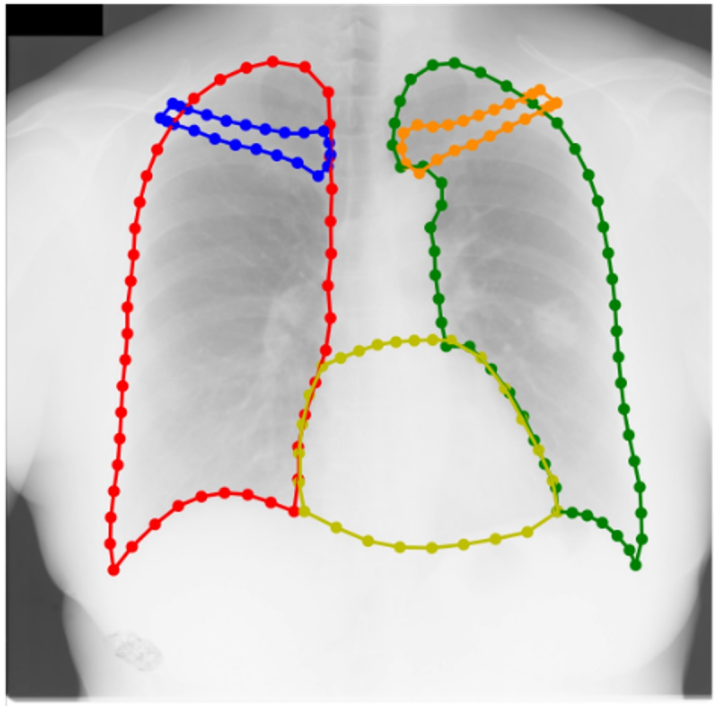


Graph construction

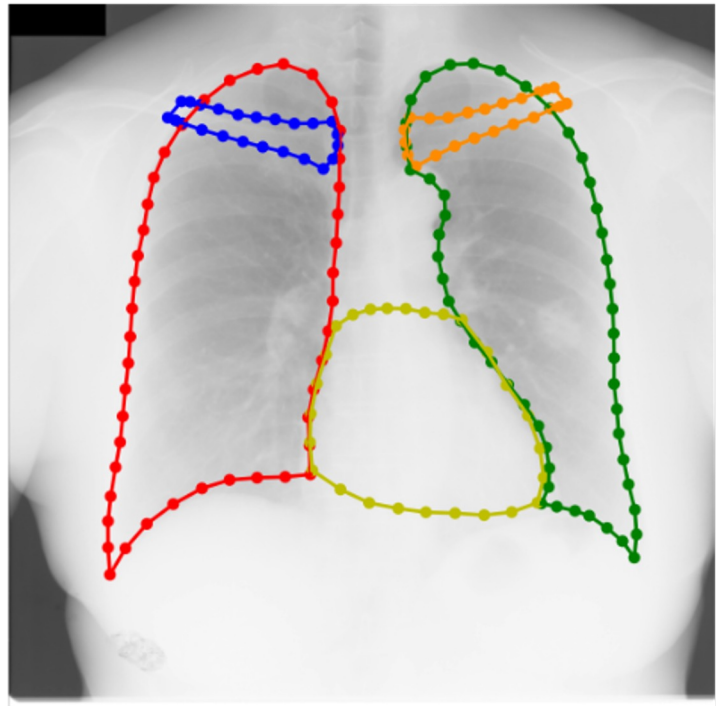
Image



GT



Landmark Segmentation



$$\mathcal{G} =< V, \mathbf{A}, \mathbf{X} >$$

V is the set of nodes

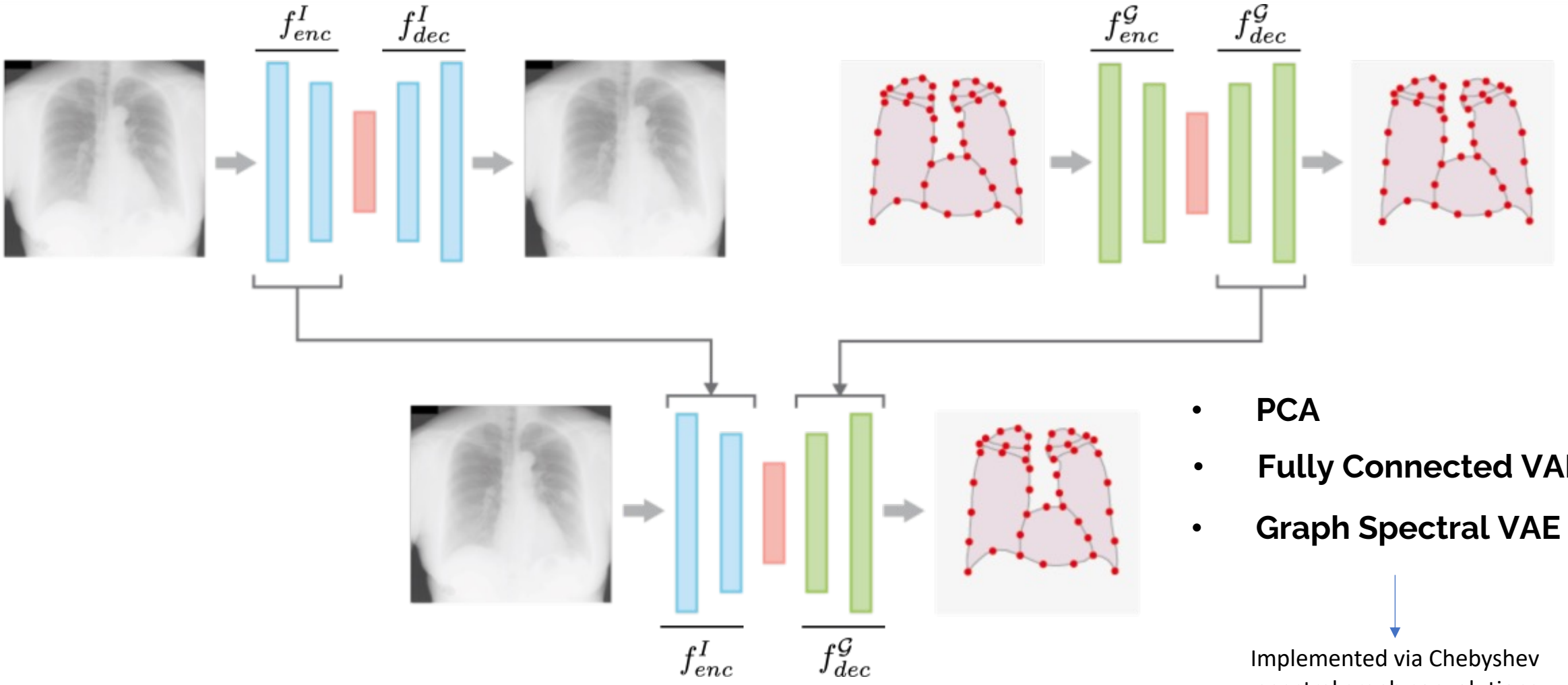
$$\mathbf{A} \in \{0,1\}^{|V|\times|V|}$$

Adjacency matrix (fixed)

$$\mathbf{X} \in \mathbb{R}^{|V|\times d}$$

Point coordinates as a node features

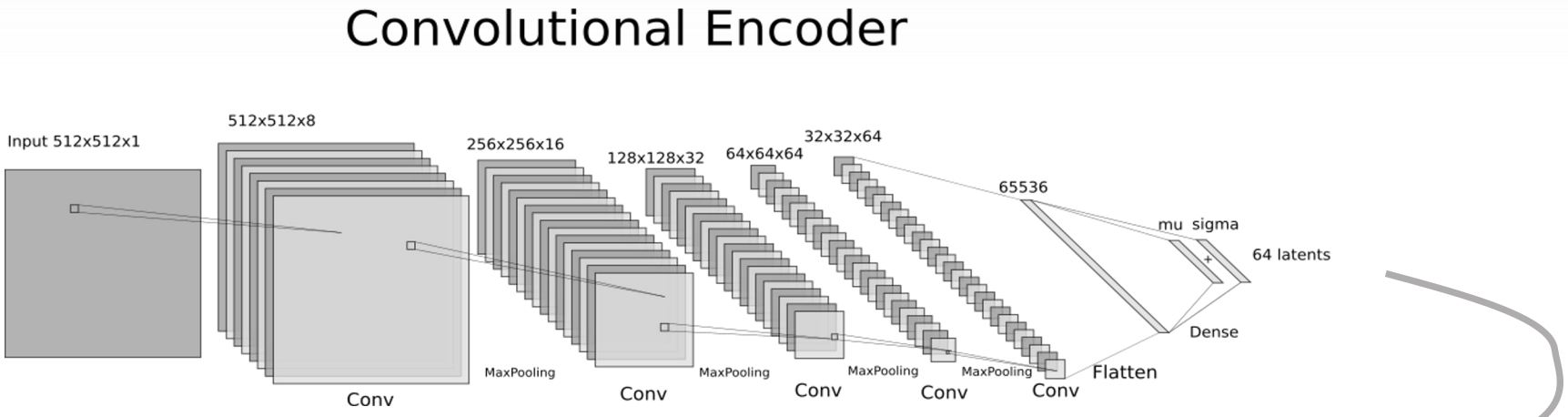
Hybrid GNet Architecture



- **PCA**
- **Fully Connected VAE**
- **Graph Spectral VAE**

Implemented via Chebyshev spectral graph convolutions (Defferrard et al, NeurIPS 2016)

Hybrid GNet Architecture



Trained end-to-end with an MSE loss on the nodes position

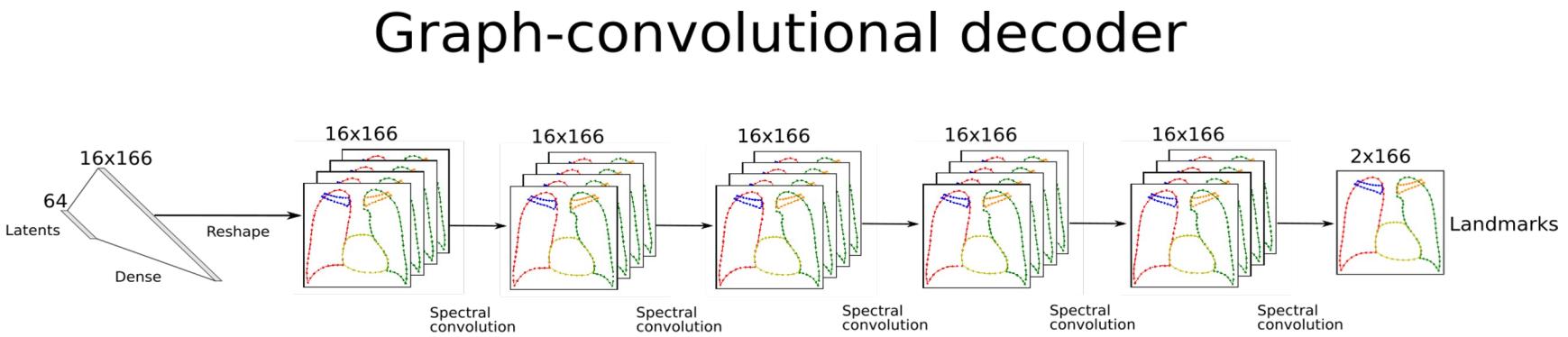


Image to graph skip connections

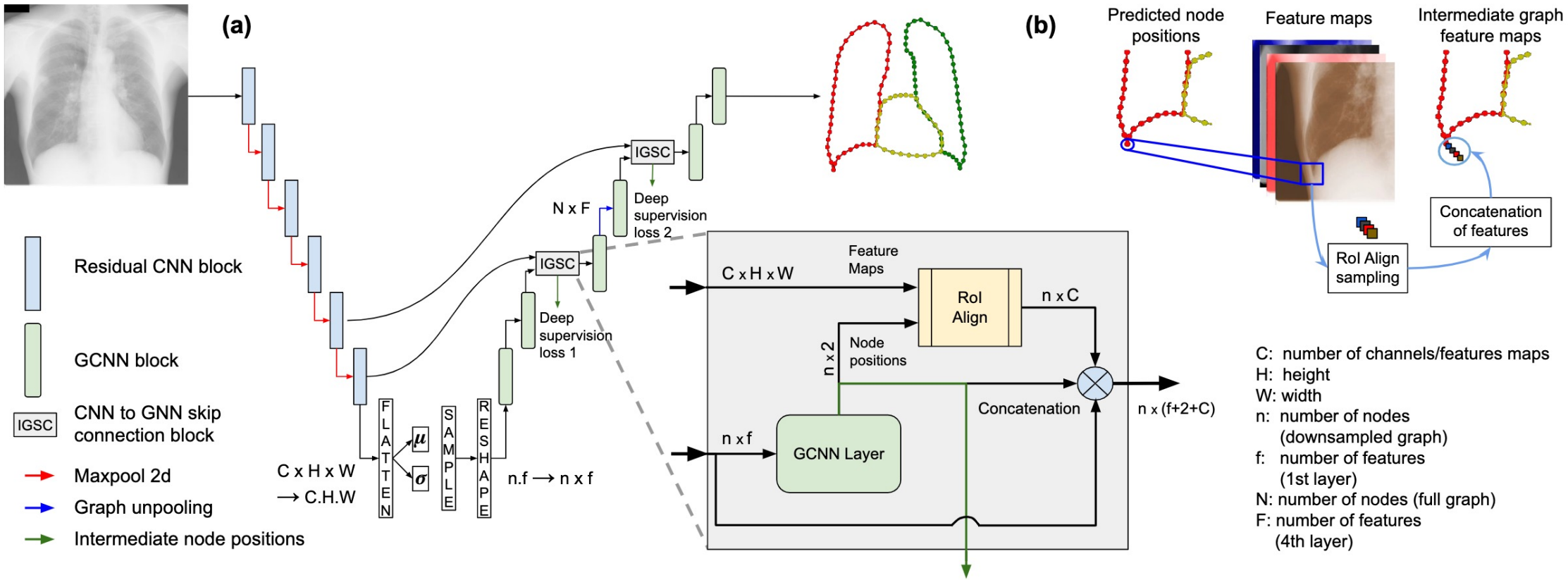


Image to graph skip connections

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 anatom-**

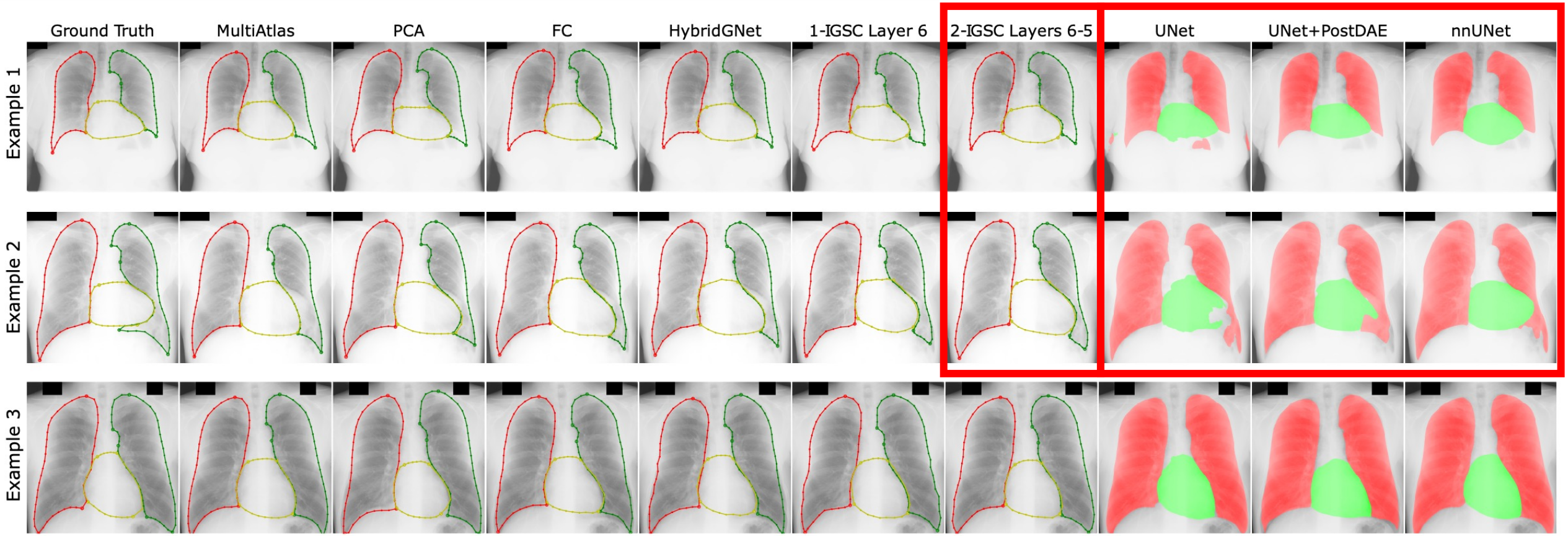
features from annotated datasets. Casting image segmentation as a pixel labeling problem is desirable in scenarios where topology and location do not tend to be preserved across individuals, like lesion segmentation. However, organs and anatomical structures usually present a characteristic topology 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

IEEE TMI 2022

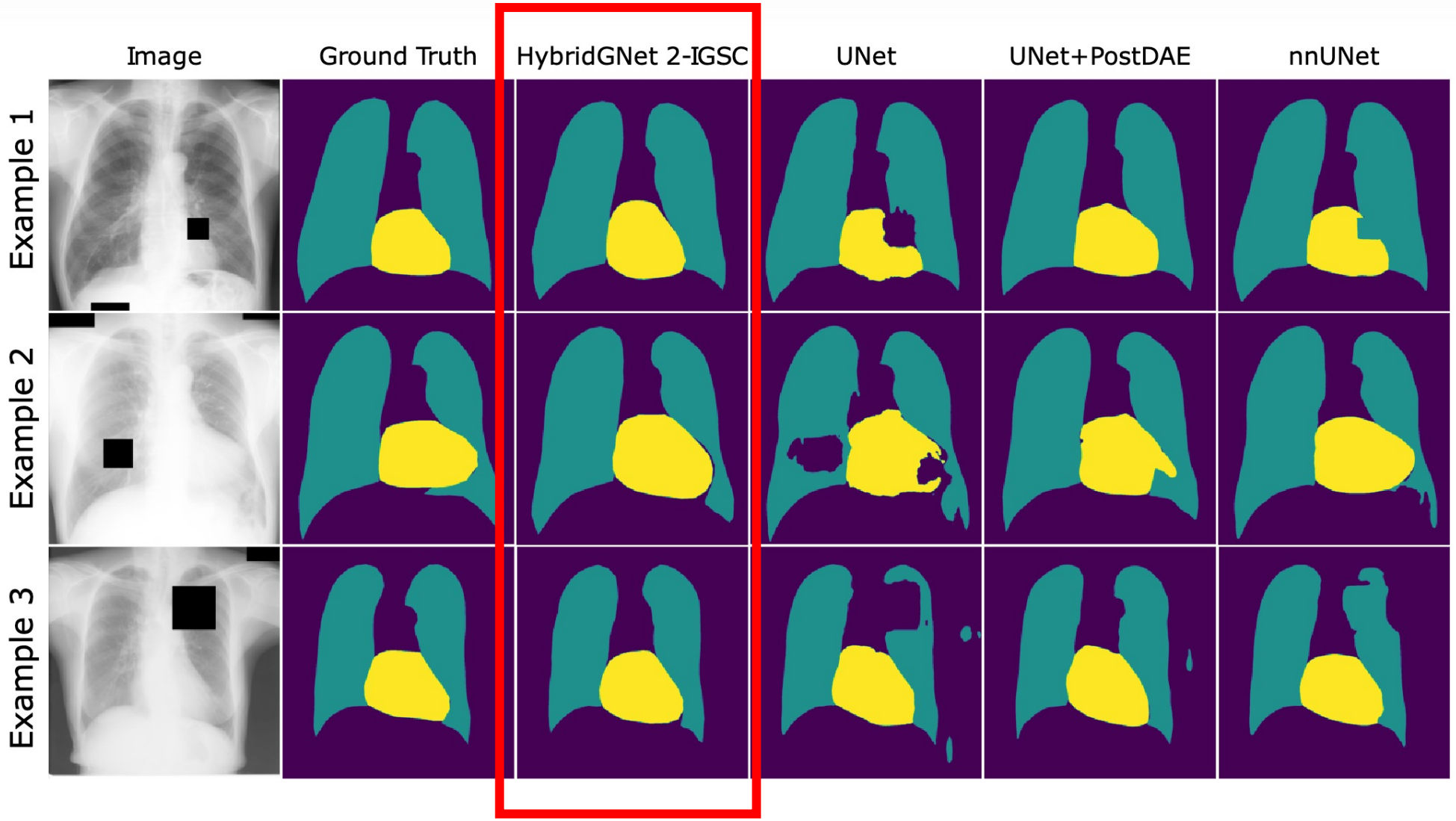
Experiment: Comparison with landmark-based baseline methods

Model		MSE	Dice Lungs	HD Lungs	Dice Heart	HD Heart
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)

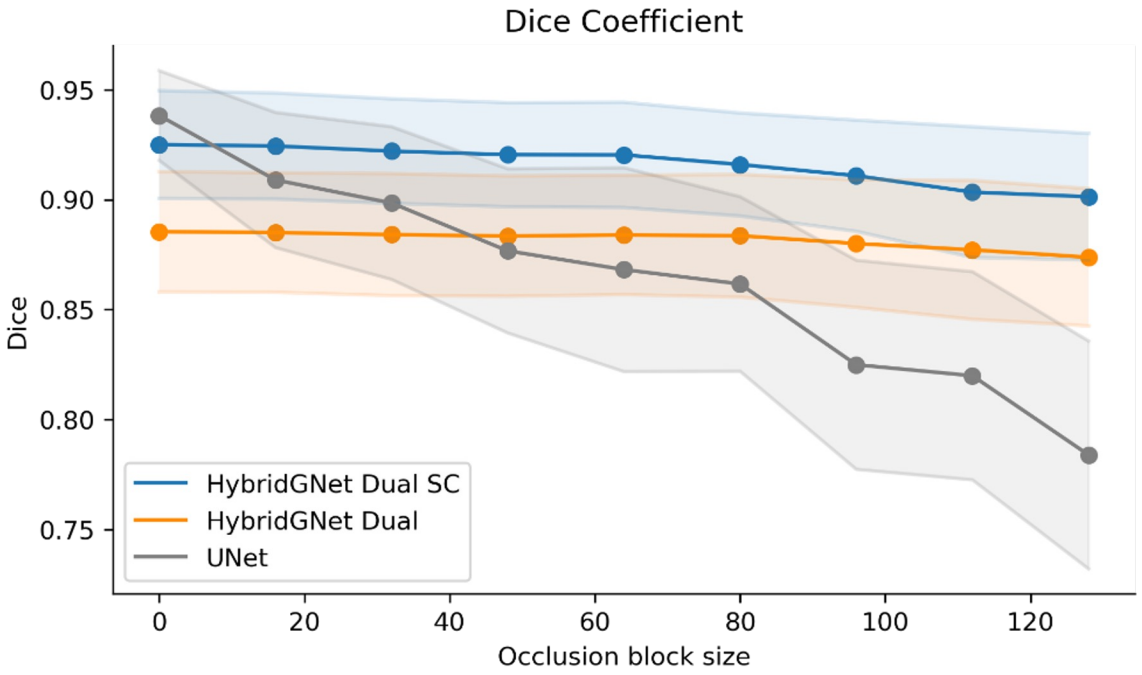
Experiment: Comparison with landmark-based baseline methods

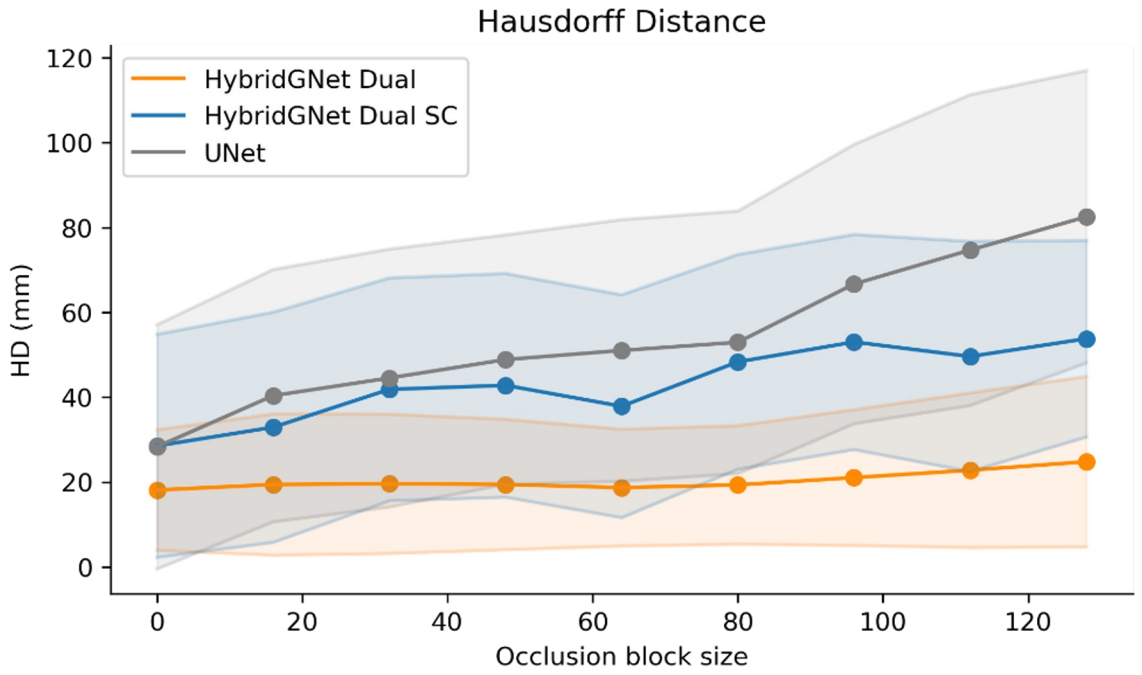


Experiment: Robustness to simulated image occlusion on dense segmentation



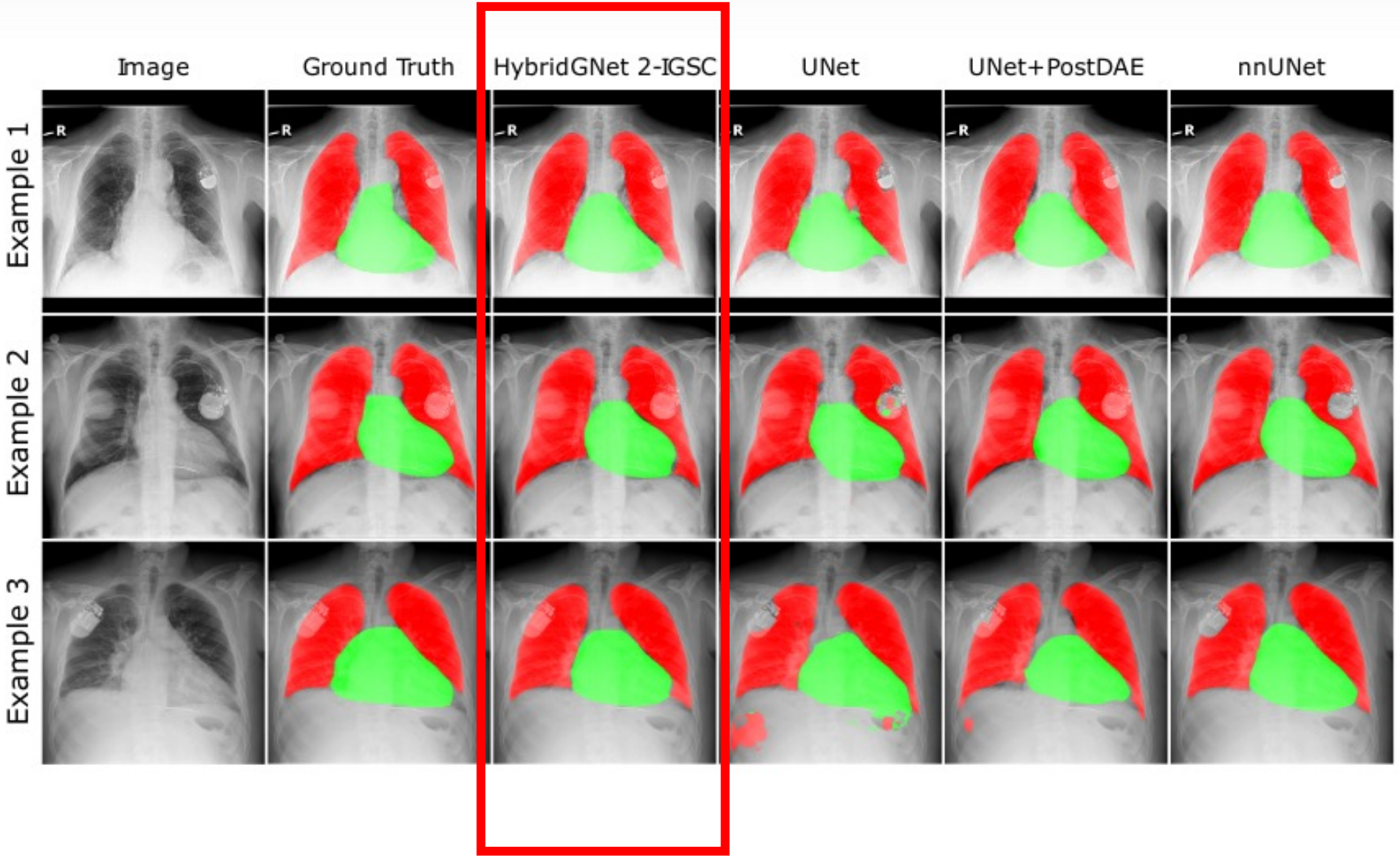
Experiment: Robustness to image occlusion on dense segmentation



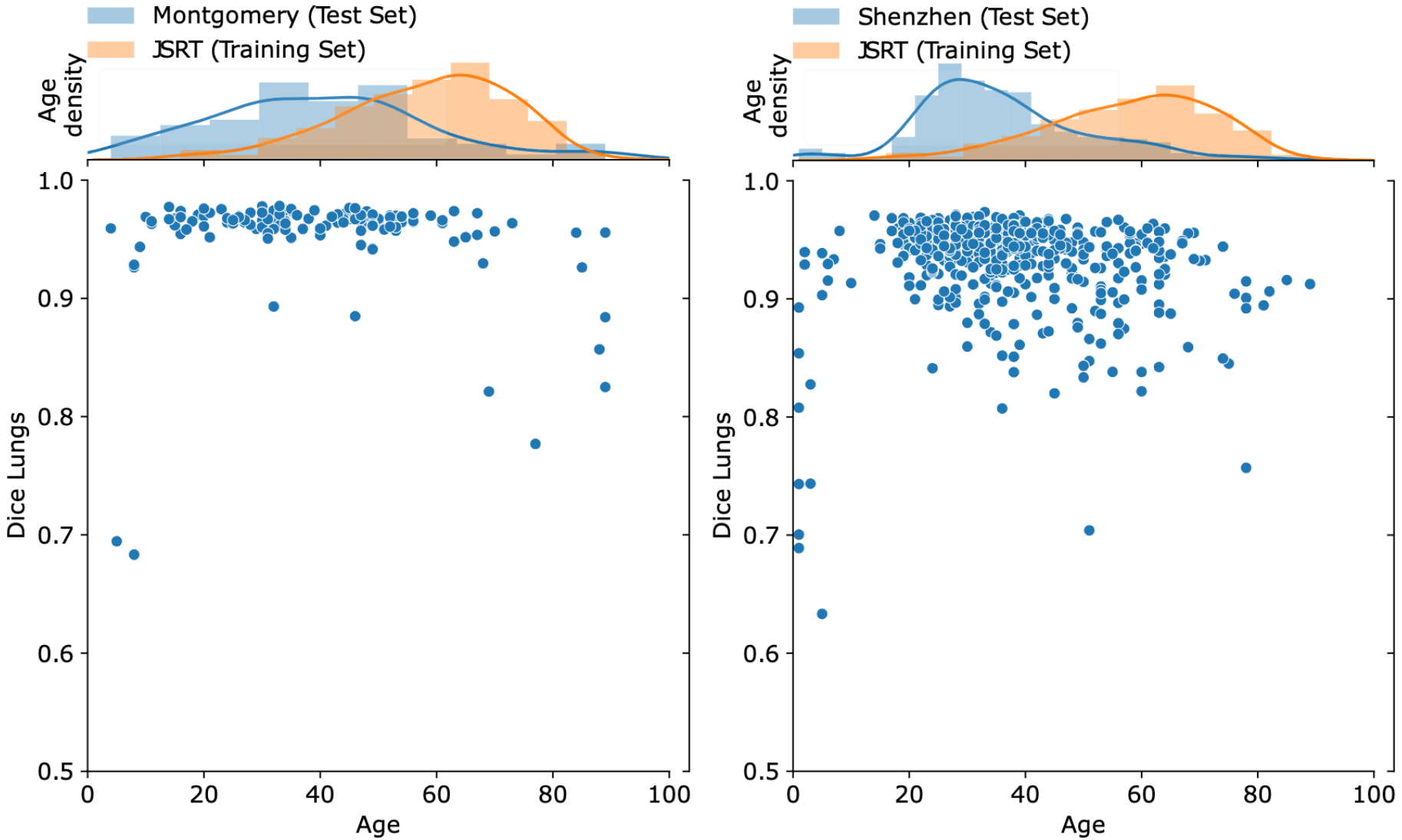


We compared the models evaluating the Dice coefficient and Hausdorff distance on the dense masks obtained from the UNet and the convolutional decoder of the dual models.

Experiment: Robustness to real image occlusion on dense segmentation

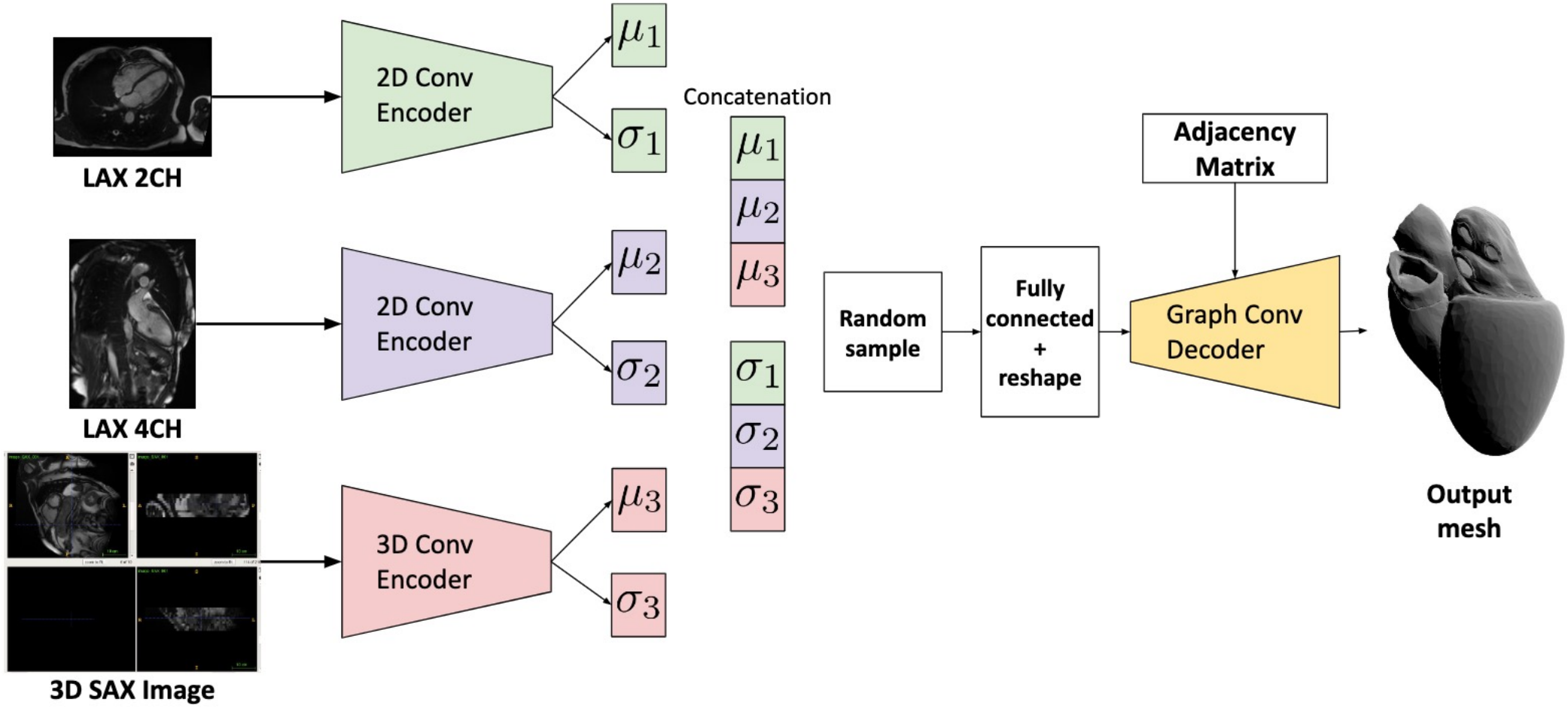


Experiment: Assessing the impact of domain shift by age distribution on lung segmentation



Extension to 3D

Extend HybridGNet to 3D meshes derived from volumetric images instead of 2D contours



Under evaluation, 2023

MULTI-CENTER ANATOMICAL SEGMENTATION WITH HETEROGENEOUS LABELS VIA LANDMARK-BASED MODELS

Nicolás Gaggion[⋆] *Maria Vakalopoulou*[†] *Diego H. Milone*[⋆] *Enzo Ferrante*[⋆]

[⋆] Research Institute for Signals, Systems and Comp. Intelligence, sinc(i), CONICET-UNL, Argentina
[†] MICS, 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

section 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

CentraleSupélec | université PARIS-SACLAY

IEEE ISBI 2023

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

Image by Vectorportal.com, CC BY

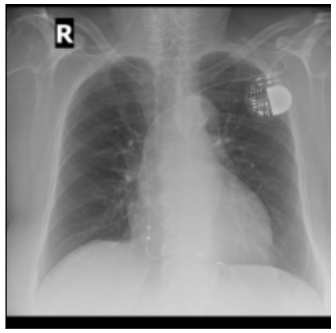
Hospital A

Hospital B

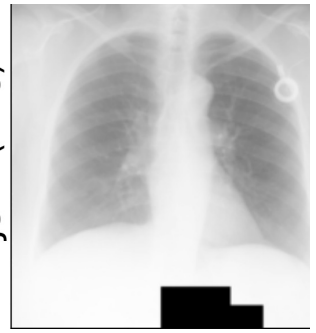
Hospital C

Montgomery (L)

Padchest (LH)

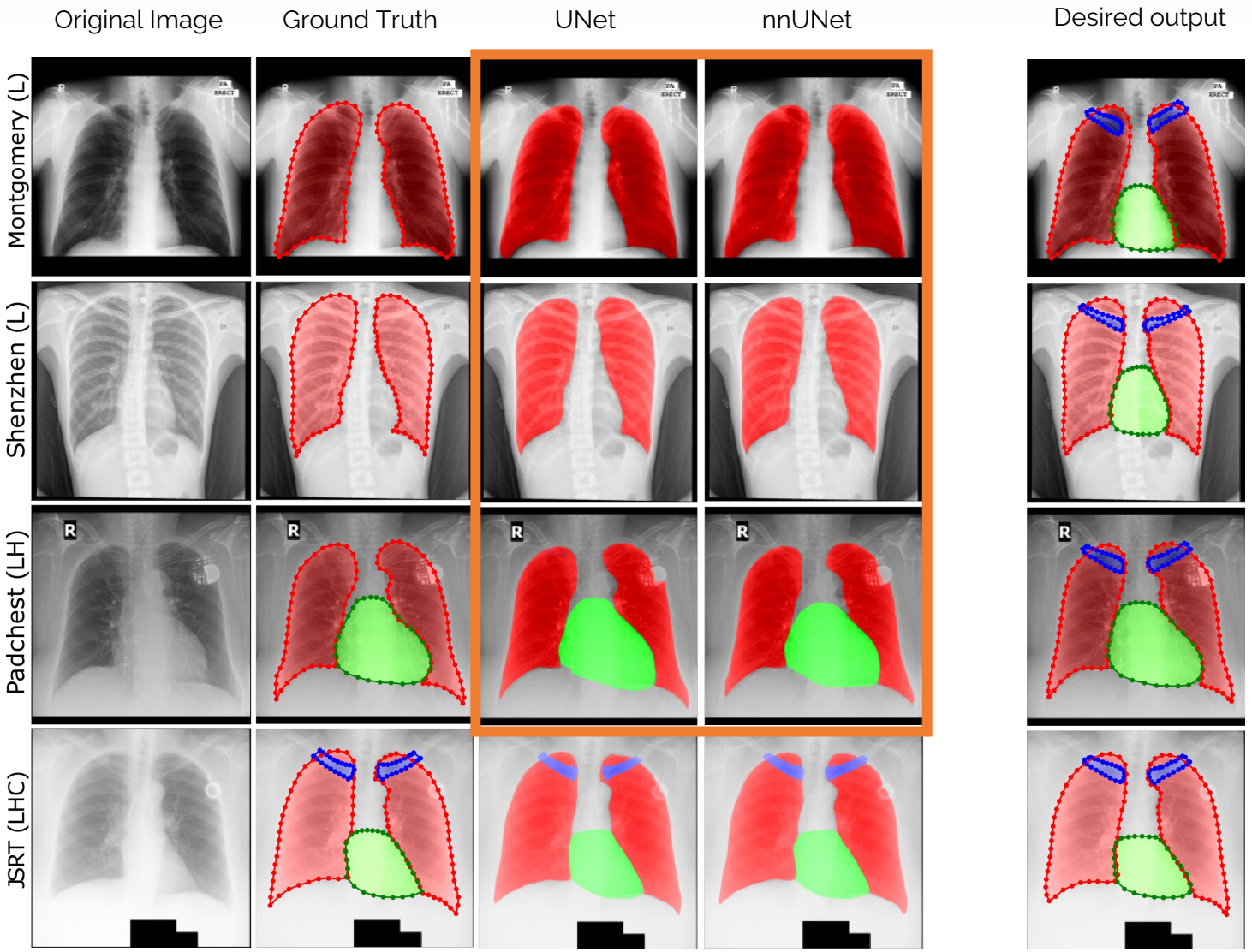


JSRT (LHC)



Problem: domain memorization

when training using multi-center data with heterogeneous labels

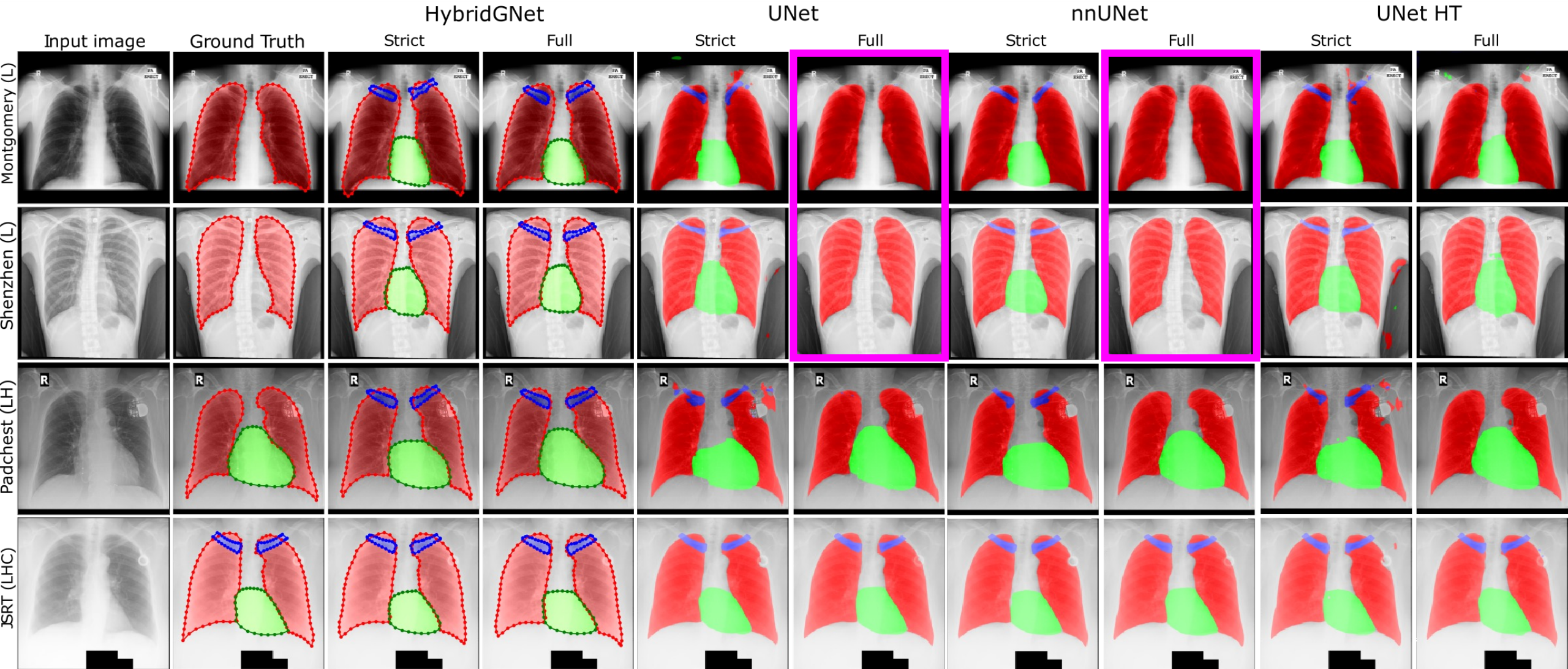


- Naïve pixel level approaches tend to fail in multicentric scenarios with heterogeneous labels due to domain memorization issues

- We propose to overcome this approach with landmark based models

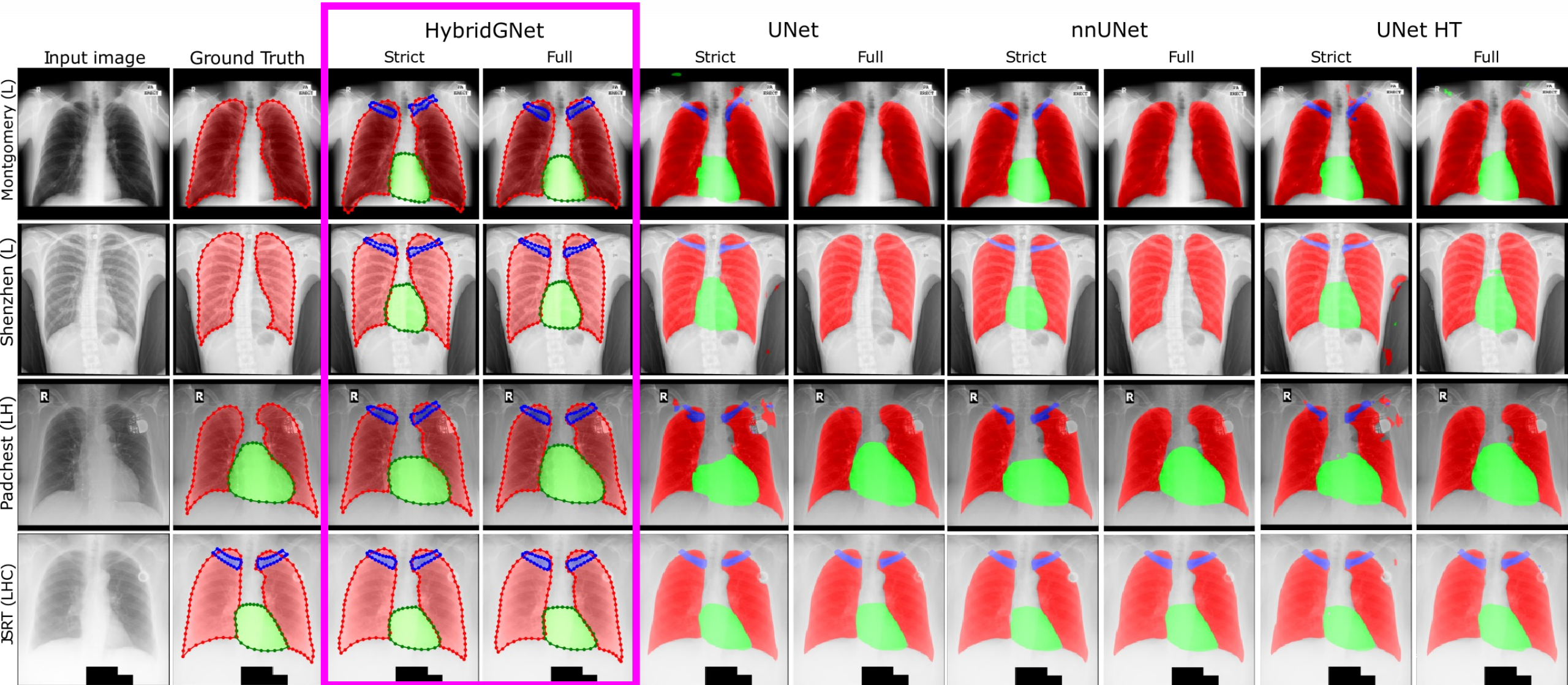
Qualitative results

UNet and nnUNet suffer from domain memorization in the full setting (e.g. the heart is not predicted in Montgomery and Shenzhen)



Qualitative results

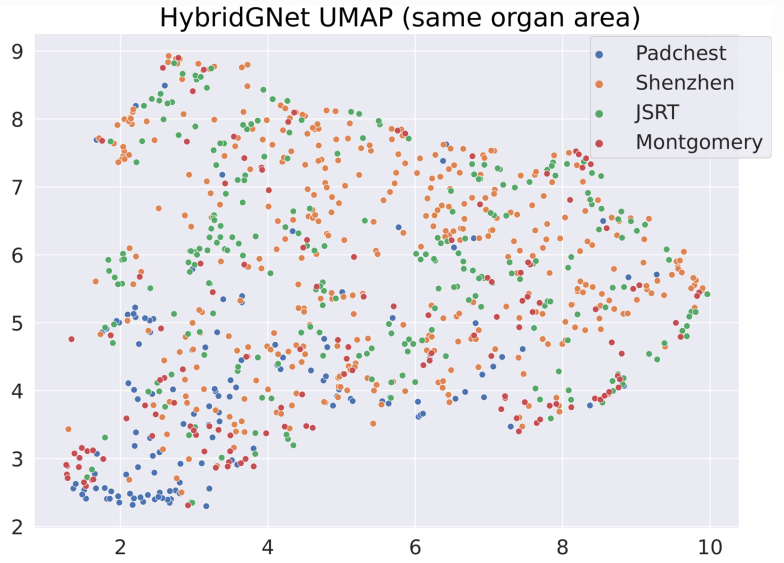
HybridGNet presents anatomically plausible results for all structures when trained in both settings

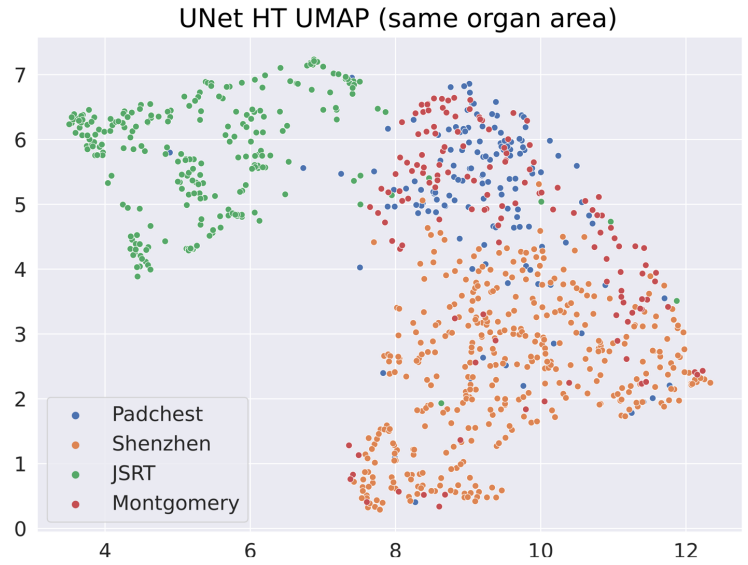


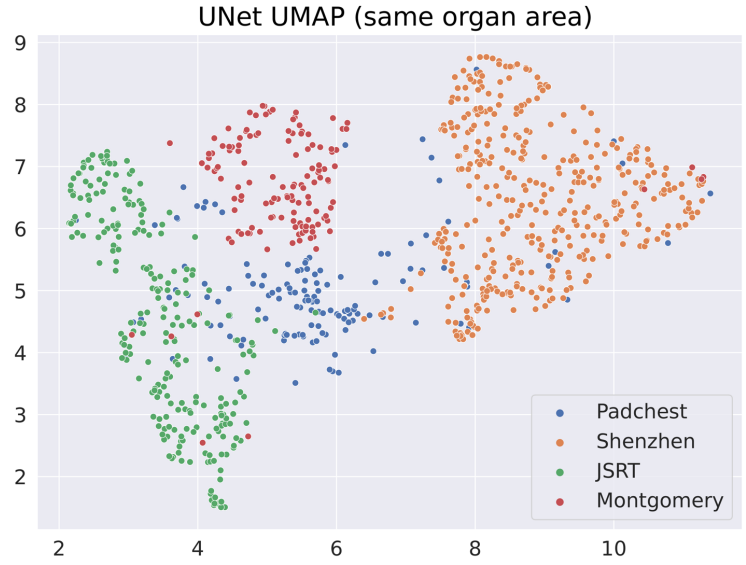
Why is this happening?

- We performed dimensionality reduction on the bottleneck latent space of HybridGNet and the UNets

- UNet and UNet HT tend to clusterize images per dataset, while HybridGNet doesn't, explaining the improved robustness to domain-label memorization.

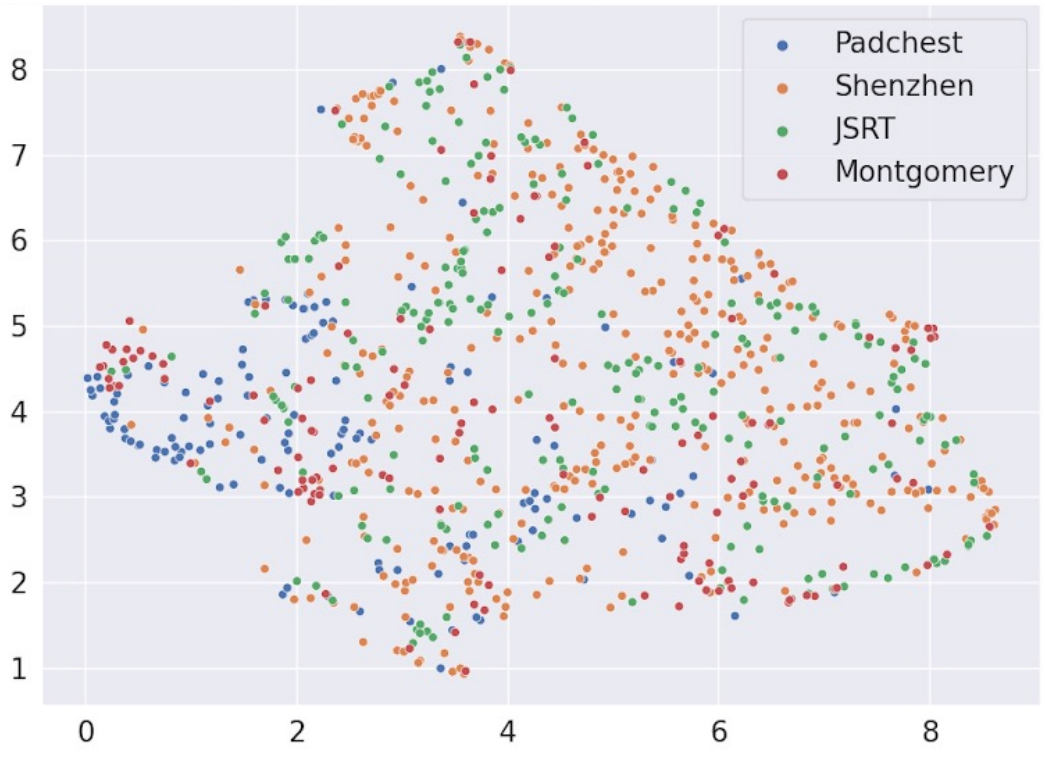




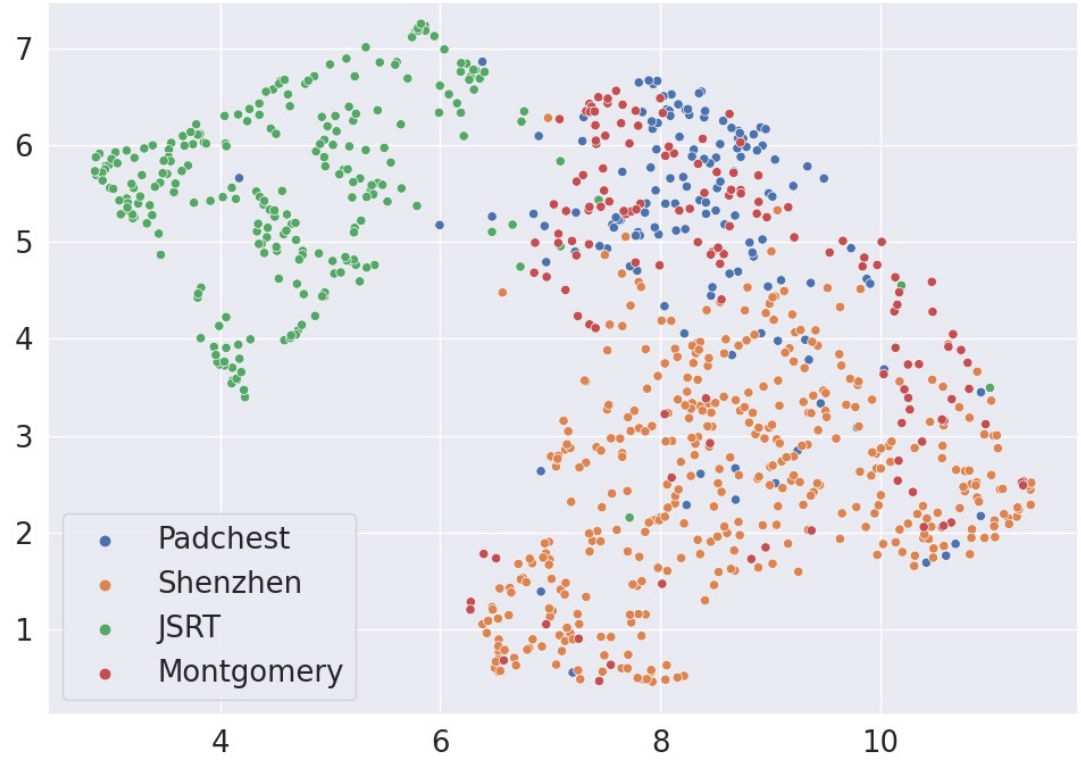


Multi-center anatomical segmentation with heterogeneous labels

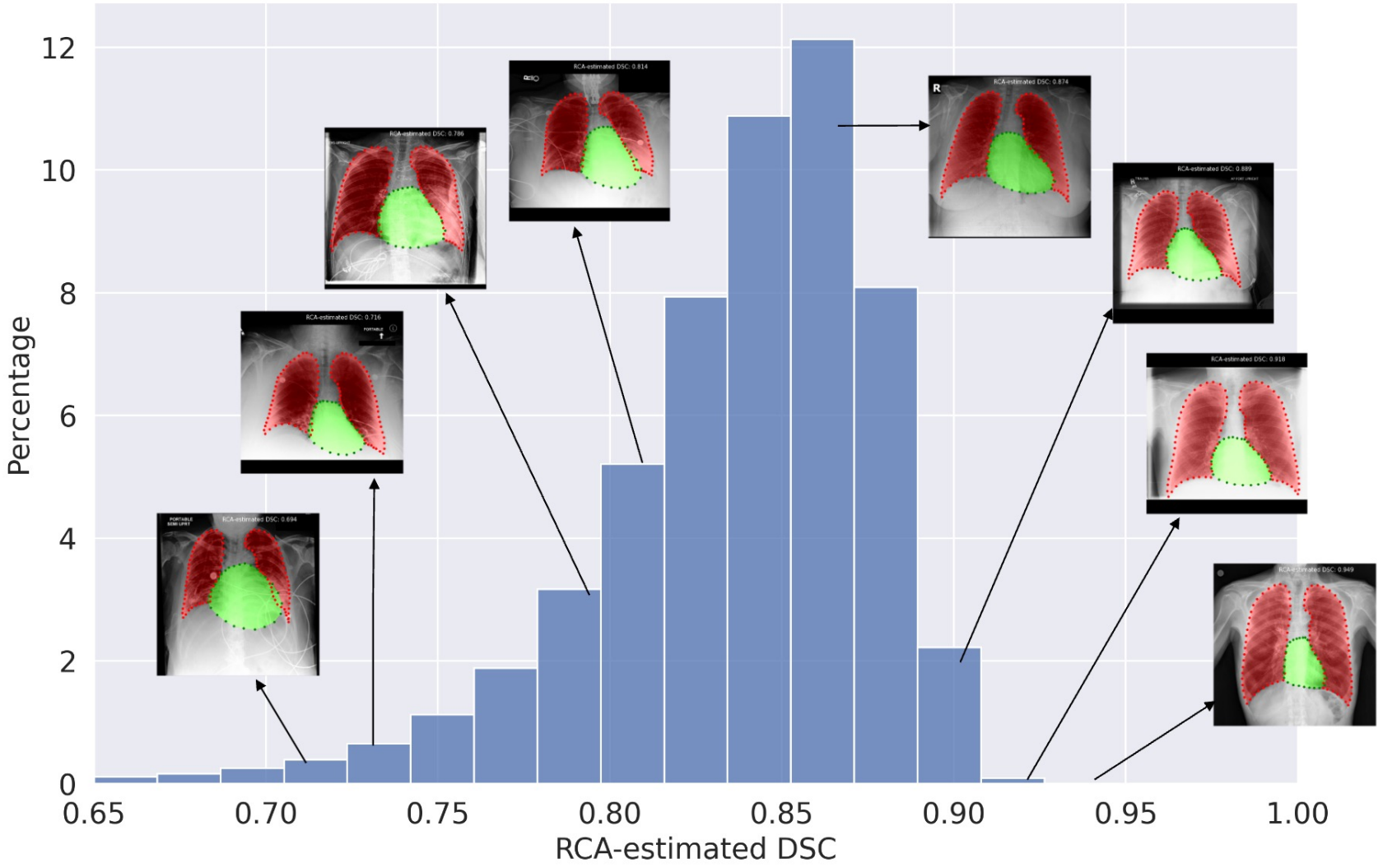
HybridGNet



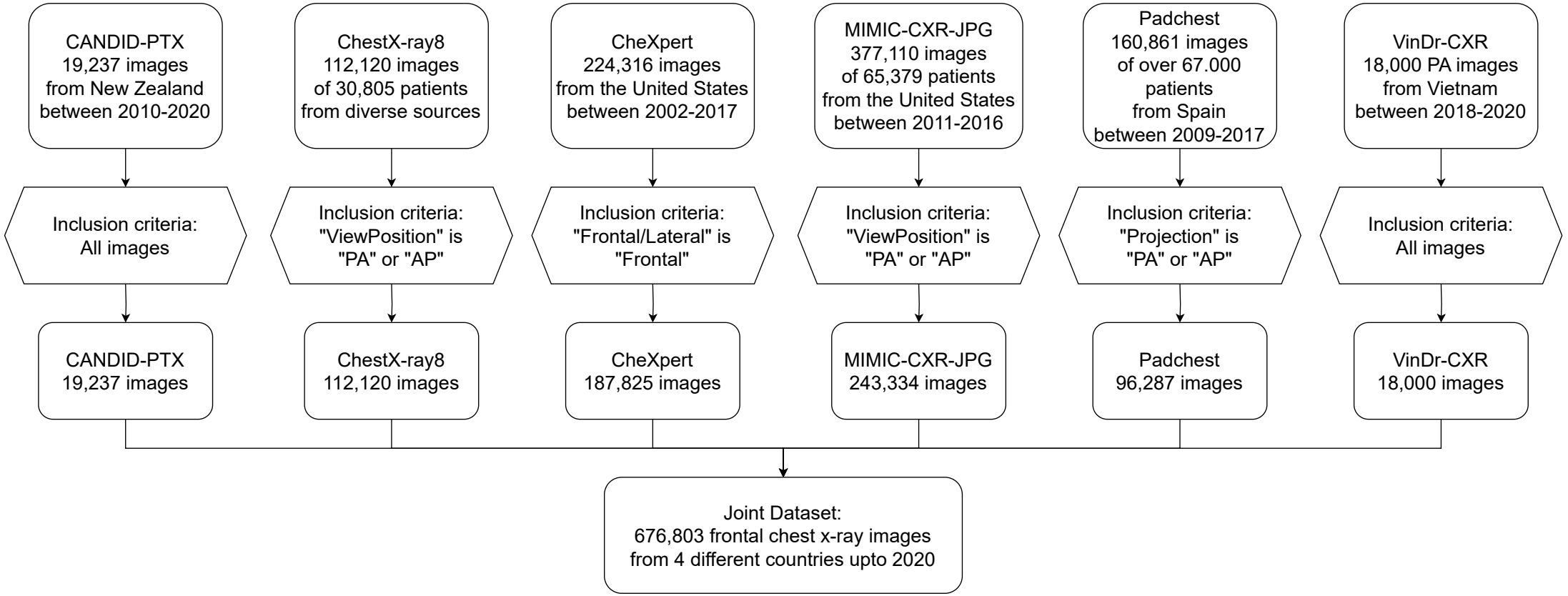
UNet



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



UNL CONICET sinc(i)

HOSPITAL ITALIANO de Buenos Aires

https://tinyurl.com/chexmask

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

Nicolás Gaggion[1], Candelaria Mosquera[2, 3], Lucas Mansilla[1], Martina Aineseder[2], Diego H. Milone[1], and Enzo Ferrante[1,*]

[1]Institute for Signals, Systems and Computational Intelligence, sinc(i) CONICET-UNL, Santa Fe, S3002, Argentina
[3]Health Informatics Department at Hospital Italiano de Buenos Aires, Buenos Aires, CP, Argentina
[3]Universidad Tecnológica Nacional, Buenos Aires, CP, Argentina
[*]corresponding author: Enzo Ferrante (eferrante@sinc.unl.edu.ar)

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 all datasets. Rigorous validation, including expert physician evaluation and automatic quality control, masks. Additionally, we provide individualized quality indices per mask and an overall source for the broader scientific community, streamlining the CheXmask dataset is publicly available

Under evaluation

Nicolás Gaggión

https://github.com/ngaggion/HybridGNet

github.com/ngaggion/HybridGNet

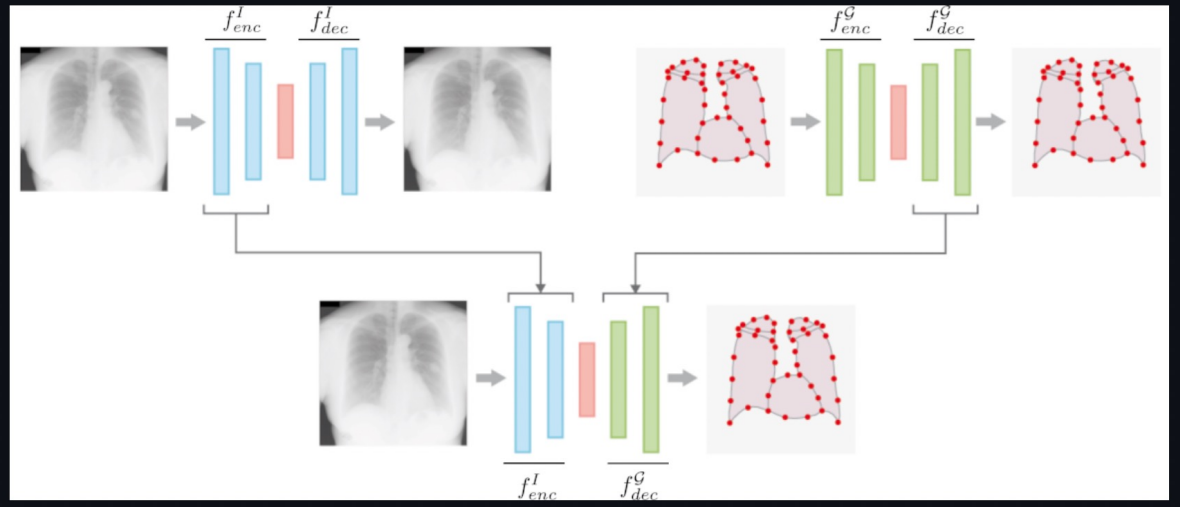
README.md

HybridGNet: Hybrid graph convolutional neural networks for landmark-based anatomical segmentation

Nicolás Gaggion[1], Lucas Mansilla[1], Diego Milone[1], Enzo Ferrante[1]

[1] Research Institute for Signals, Systems and Computational Intelligence (sinc(i)), FICH-UNL, CONICET, Ciudad Universitaria UNL, Santa Fe, Argentina.

Full-paper accepted at MICCAI 2021.
Pre-print available at https://arxiv.org/abs/2106.09832



Installation:

Bias in AI for medical image analysis

nature

Subscribe Search Login

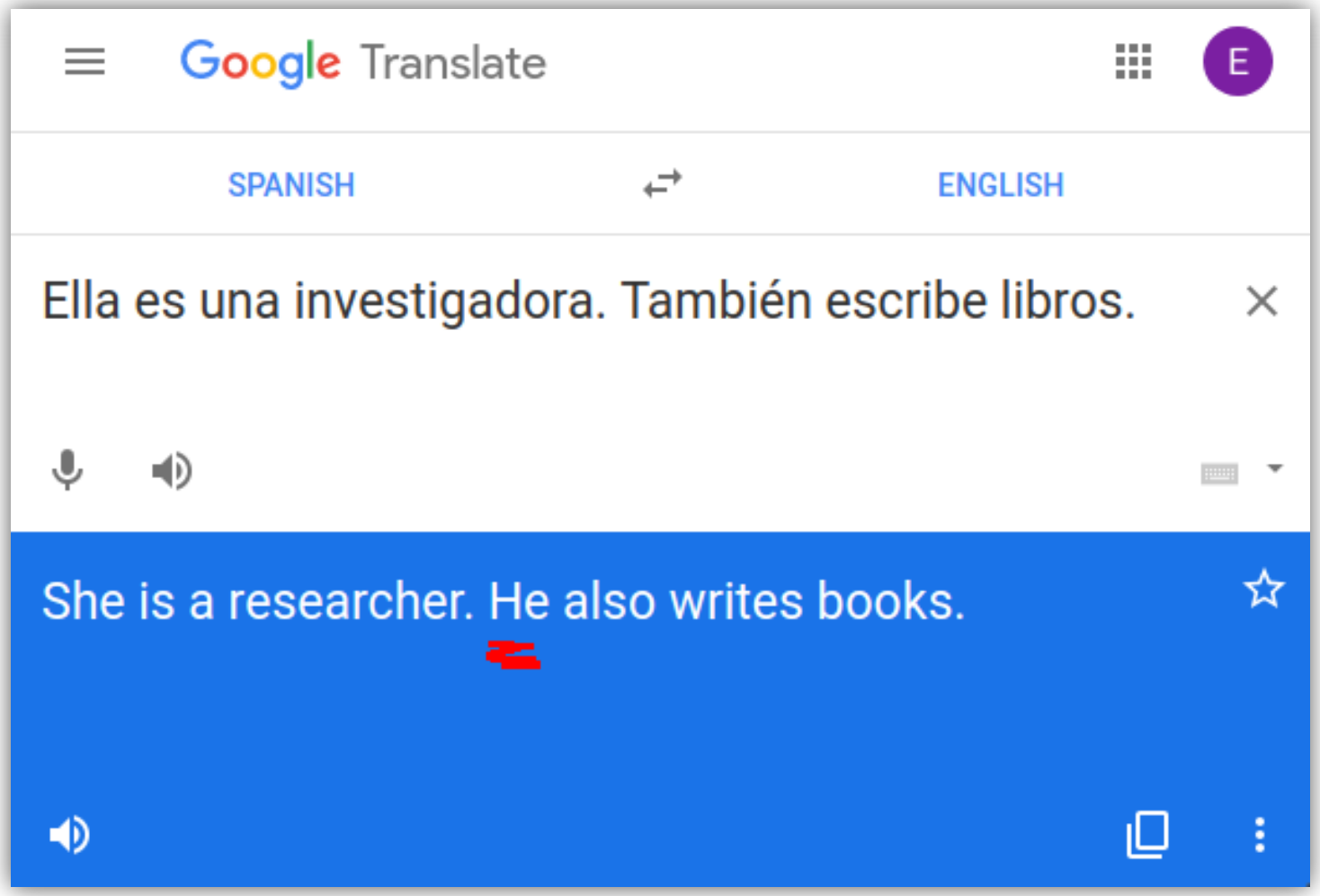
COMMENT · 18 JULY 2018

AI can be sexist and racist — it's time to make it fair

Computer scientists must identify sources of bias, de-bias training data and develop artificial-intelligence algorithms that are robust to skews in the data, argue James Zou and Londa Schiebinger.

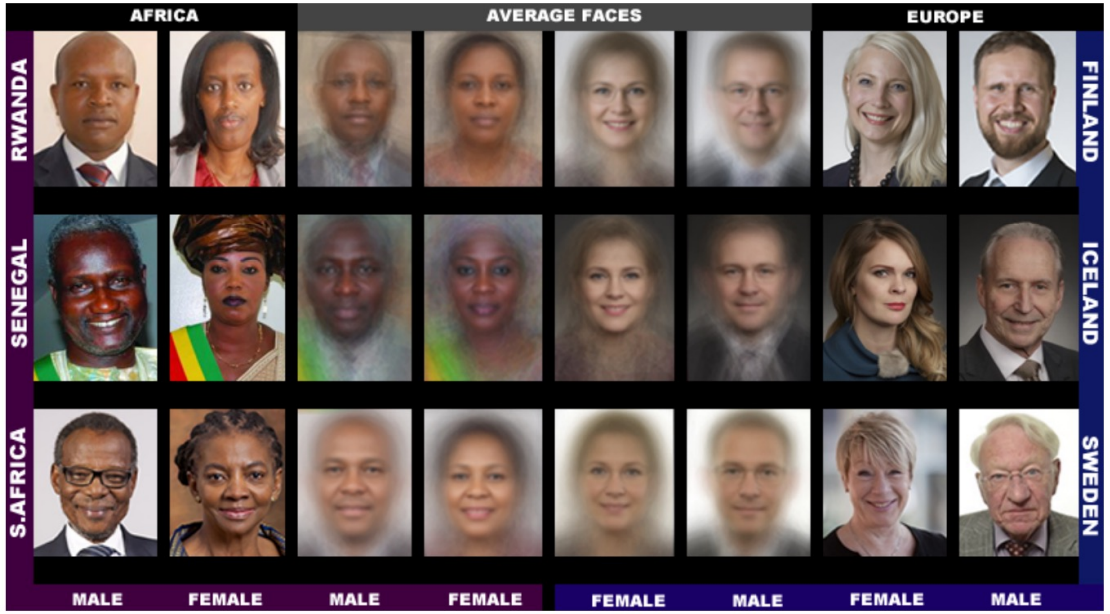
James Zou & Londa Schiebinger

Gender bias in AI systems



Racial bias in AI systems

Face recognition



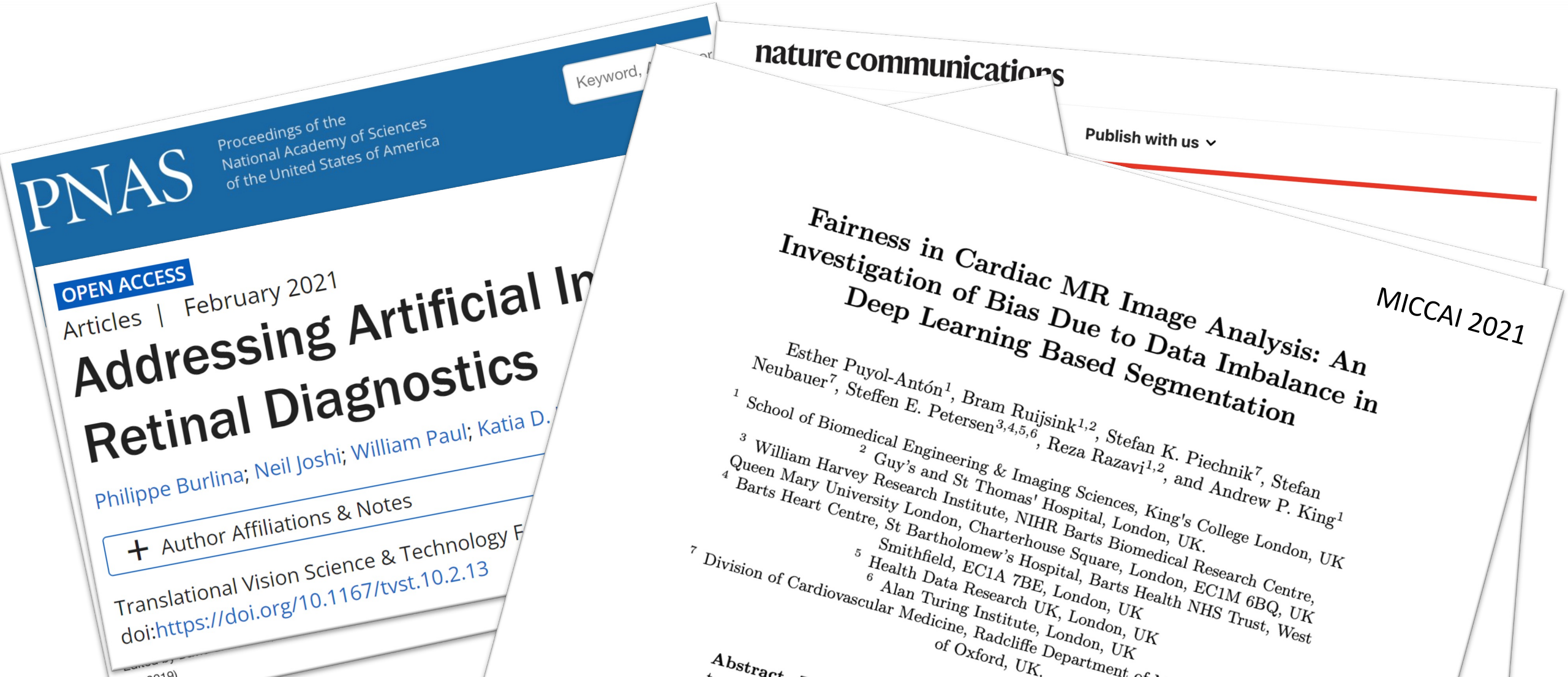
Publicly available commercial face recognition online services provided by Microsoft, Face++, and IBM respectively are found to suffer from achieving much lower accuracy on females with darker skin color(see Fig4, Buolamwini and Gebru, 2018).

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Source: Buolamwini & Gebru, 2018. Conference on Fairness, Accountability and Transparency

Fairness of AI for medical image analysis

Deep learning models for medical image analysis can exhibit bias with respect to specific sub-populations

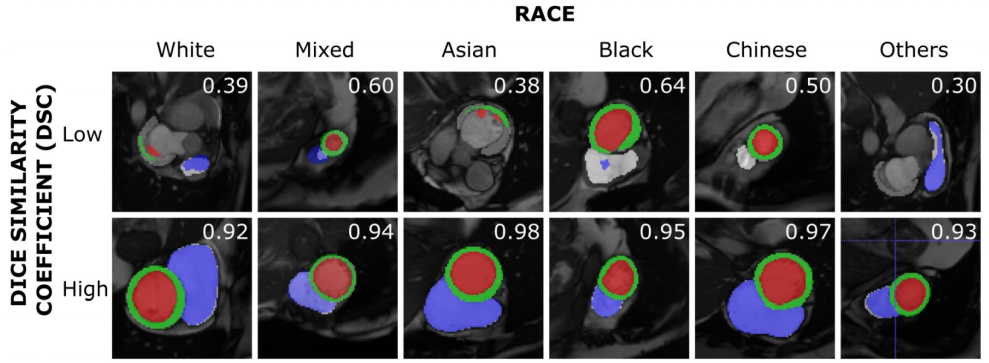


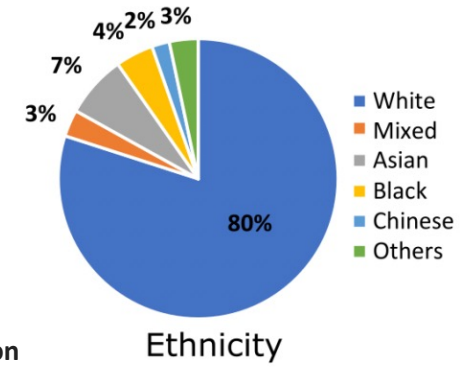
Bias in deep segmentation networks

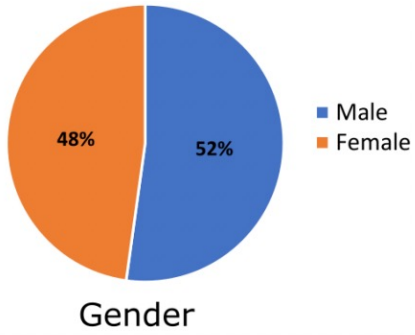
Ethnicity bias in cardiac segmentation

DSC (%) for Baseline —Fairness through unawareness

	ED			ES			Avg
	LVBP	LVM	RVBP	LVBP	LVM	RVBP	
Total	93.48	83.12	89.37	89.37	86.31	80.61	**87.05**
Male	93.58	83.51	88.82	90.68	85.31	81.00	**87.02**
Female	93.39	82.71	89.90	89.59	86.60	80.21	**87.07**
White	97.33	93.08	94.09	95.06	90.58	90.88	**93.51***
Mixed	92.70	78.94	86.91	86.70	82.54	79.32	**84.52***
Asian	94.53	87.33	90.51	90.13	88.94	81.94	**88.90***
Black	92.77	85.93	89.49	89.42	85.74	71.91	**85.88***
Chinese	91.81	74.51	85.74	86.39	85.12	79.34	**83.82***
Others	91.74	78.94	89.50	88.53	84.96	80.27	**85.66***



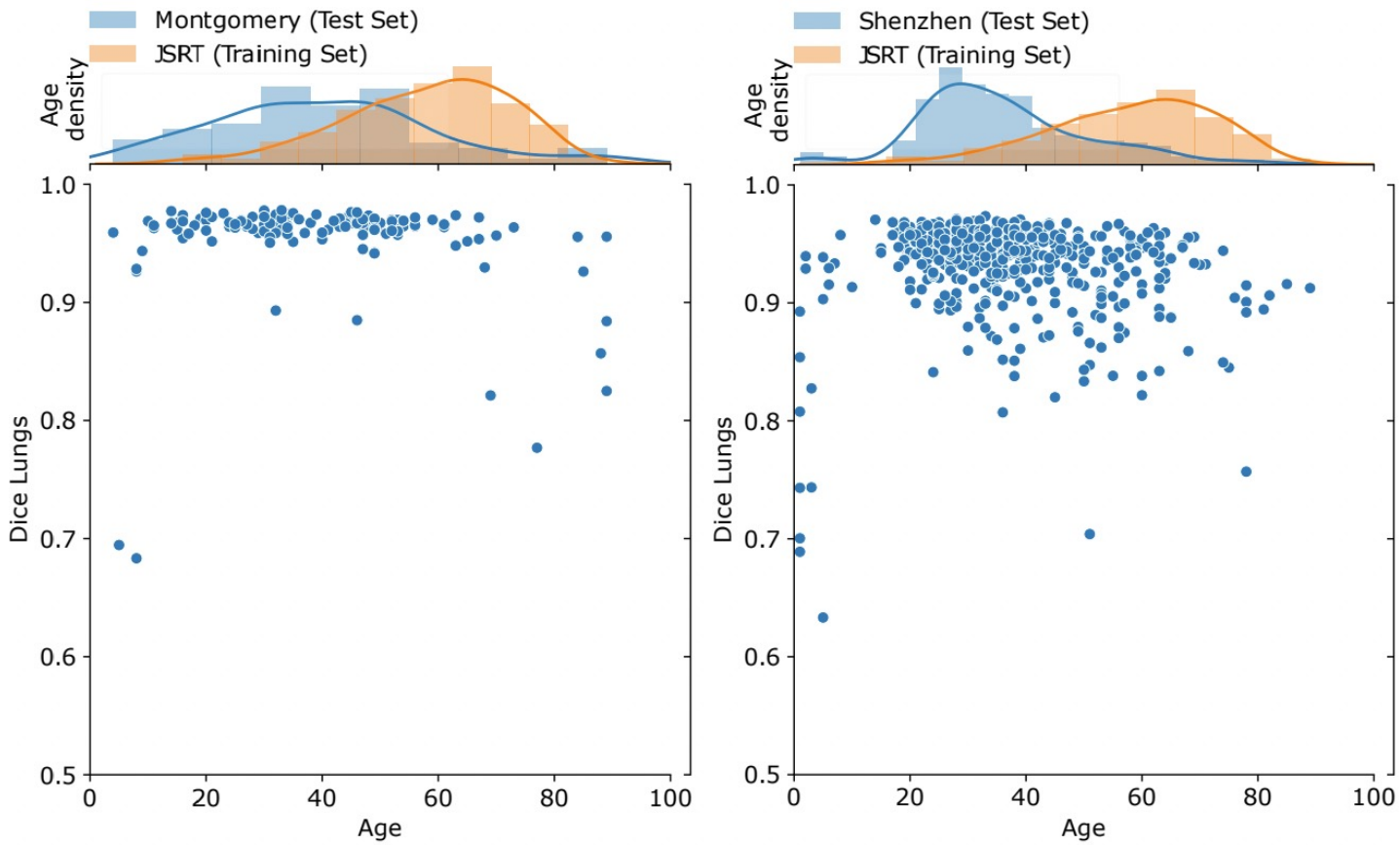




Fairness in cardiac MR image analysis: an investigation of bias due to data imbalance in deep learning based segmentation
Puyol-Antón E, Ruijsink B, Piechnik SK, Neubauer S, Petersen SE, Razavi R, King AP.
MICCAI 2021

Bias in deep segmentation networks

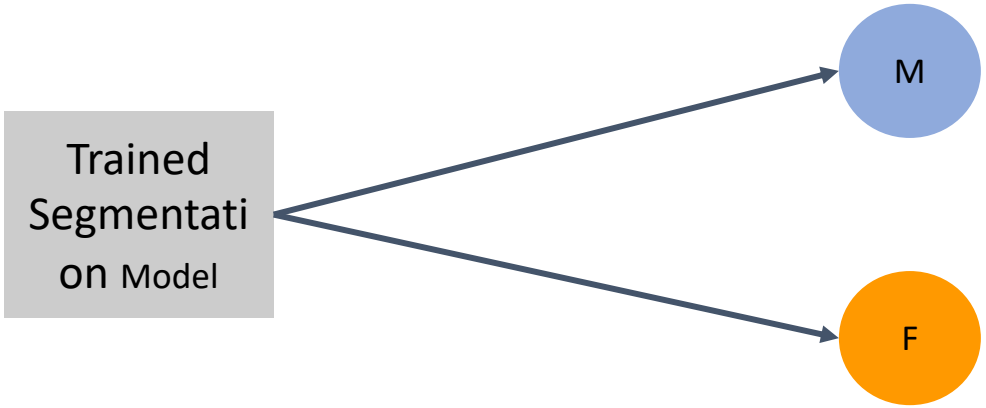
Age bias in lung segmentation



Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis.
IEEE Transactions on Medical Imaging. 2022 Nov 24;42(2):546-56.
Gaggion N, Mansilla L, Mosquera C, caMilone DH, **Ferrante E.**

Auditing fairness in medical image analysis model

Evaluate on subgroups



Compute fairness metrics

- Dice gap
- Dice STD
- Dice Skewed error rate (SER)

$$SER = \frac{\max_g(1-DSC_g)}{\min_g(1-DSC_g)}$$

- Hausdorff based metrics
- Etc

Limitation: we require ground-truth annotations to compute most fairness metrics for segmentation

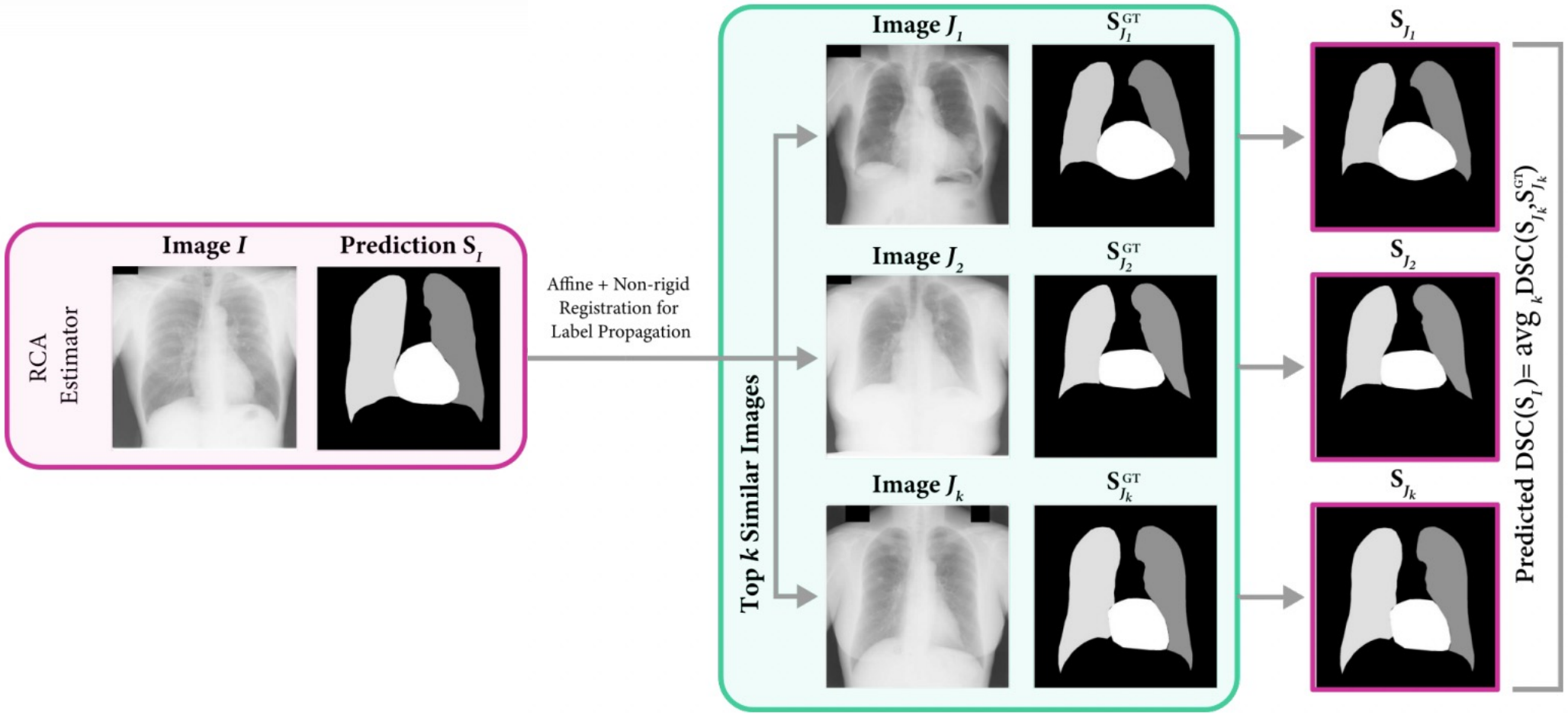
Unsupervised Bias Discovery (UBD)

in deep segmentation networks

Can we **anticipate bias** for segmentation
in new populations
without ground-truth annotations?

Proposed solution

UBD based on Reverse Classification Accuracy (RCA)



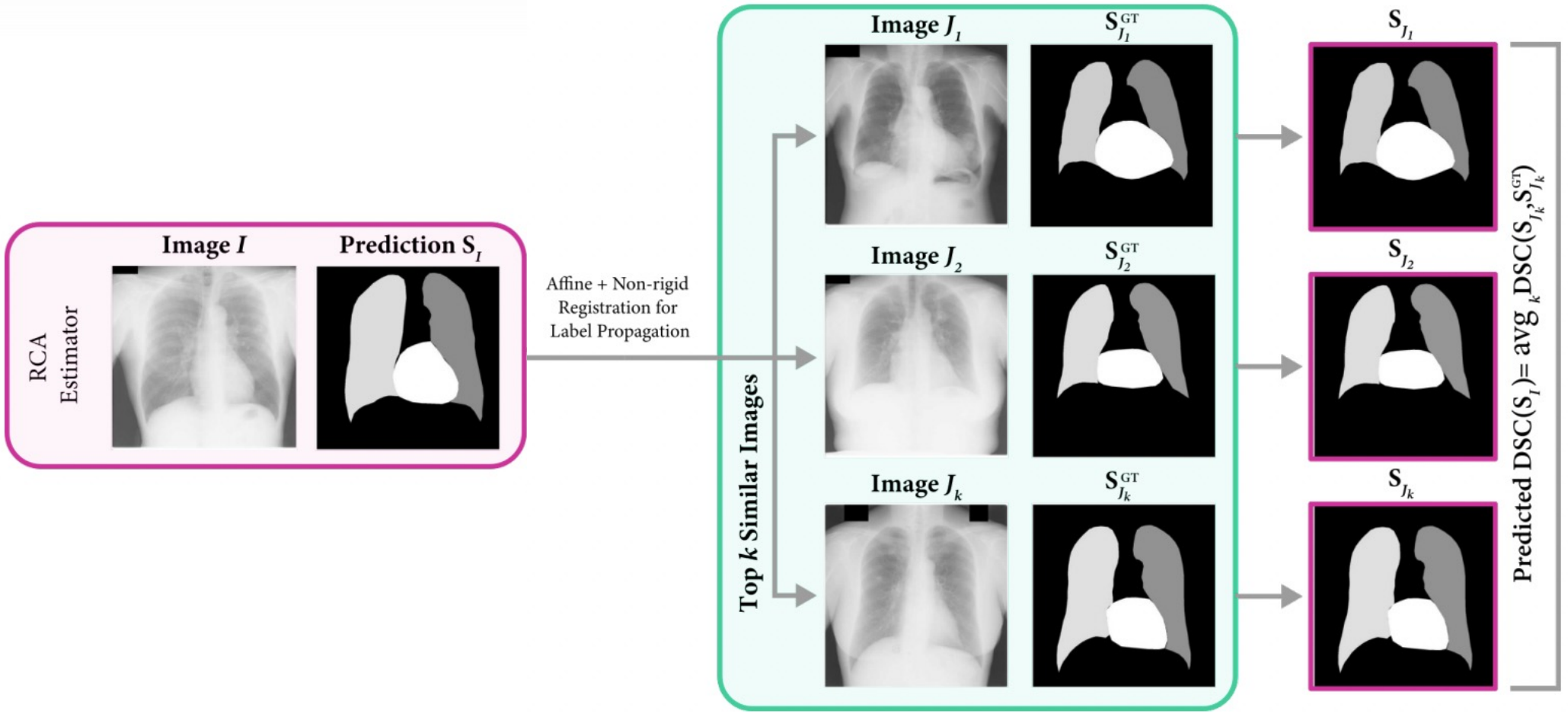
Reverse classification accuracy: predicting segmentation performance in the absence of ground truth.

IEEE transactions on medical imaging. 2017

Valindria VV, Lavdas I, Bai W, Kamnitsas K, Aboagye EO, Rockall AG, Rueckert D, Glocker B.

Proposed solution

UBD based on Reverse Classification Accuracy (RCA)



We will use RCA to estimate the signed gap in terms of DSC (or HD) between different demographic subgroups

$$\Delta DSC^{RCA} = DSC^{RCA}_{A=M} - DSC^{RCA}_{A=F}$$

Experimental validation of RCA based UBD

Synthetic experiment

- Mix of 4 different x-ray datasets (comprising a total of 911 images) including JSRT, Montgomery, Shenzhen and a minor subset of the Padchest dataset.
- UNet model trained via a compound soft Dice and cross-entropy loss
- We saved 12 different Unet versions from intermediate training checkpoints.

M_1 M_2 …. M_{12}

Experimental validation of RCA based UBD

Synthetic experiment

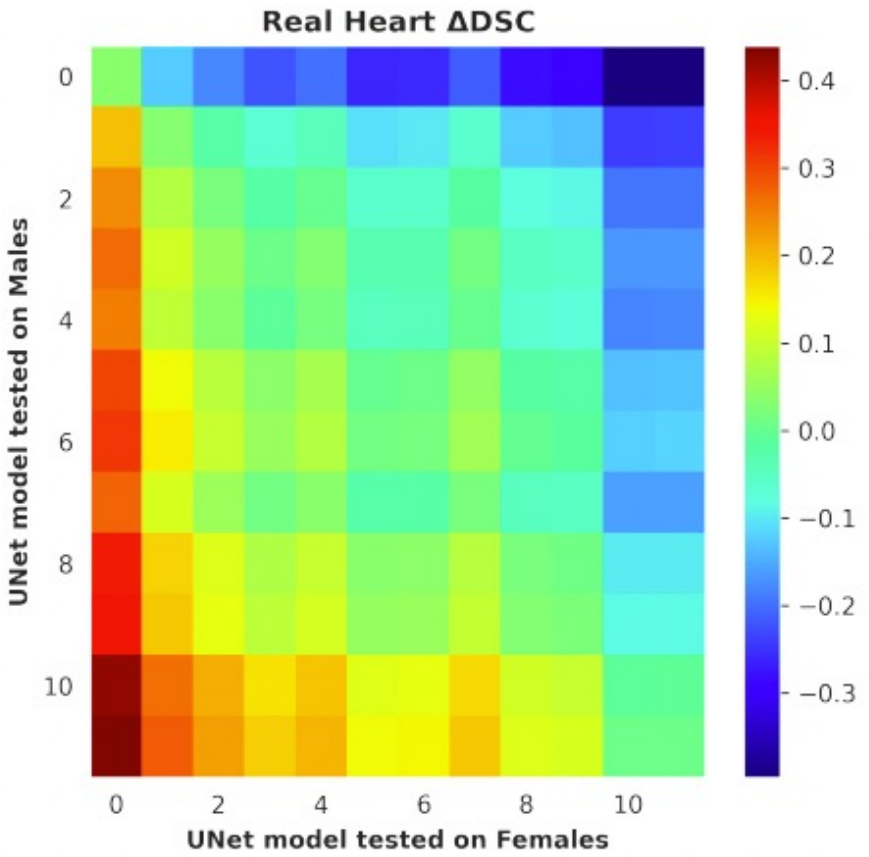
Simulated scenario: the segmentation quality varies based on sex, with either male or female patients exhibiting superior performance.

We selected pairs of UNet models (M_i , M_j) to segment the male patients (M_i) and female patients (M_j)

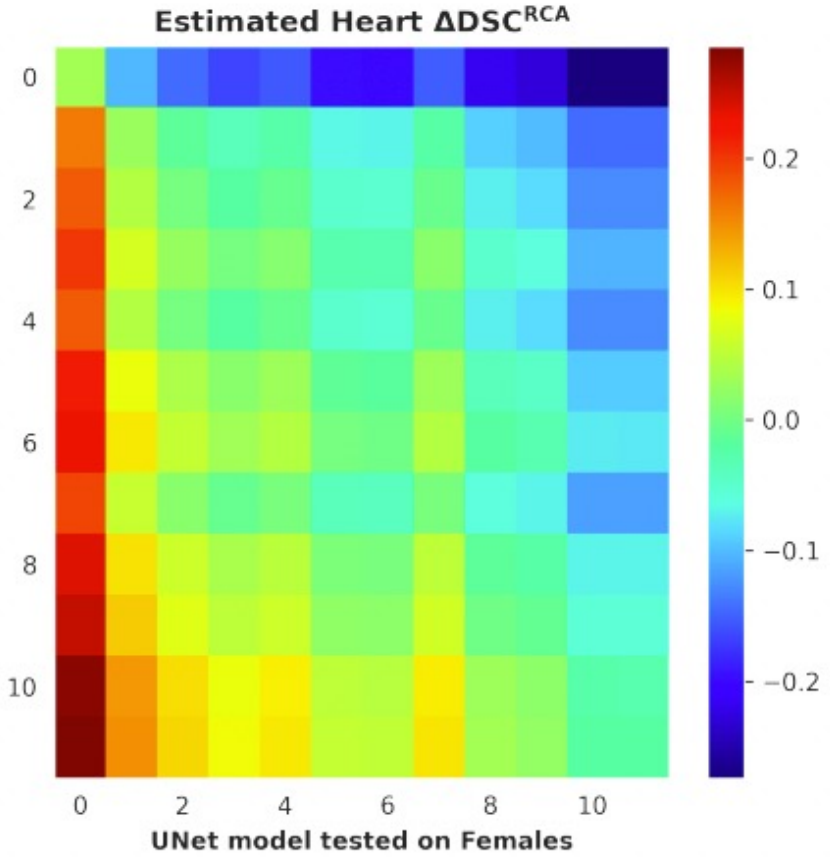
Masks were coming from different models, but in this synthetic experiment **we consider them to be generated by a single fictitious model** whose fairness we would aim to audit.

Results: simulated scenario

DSC estimated with GT

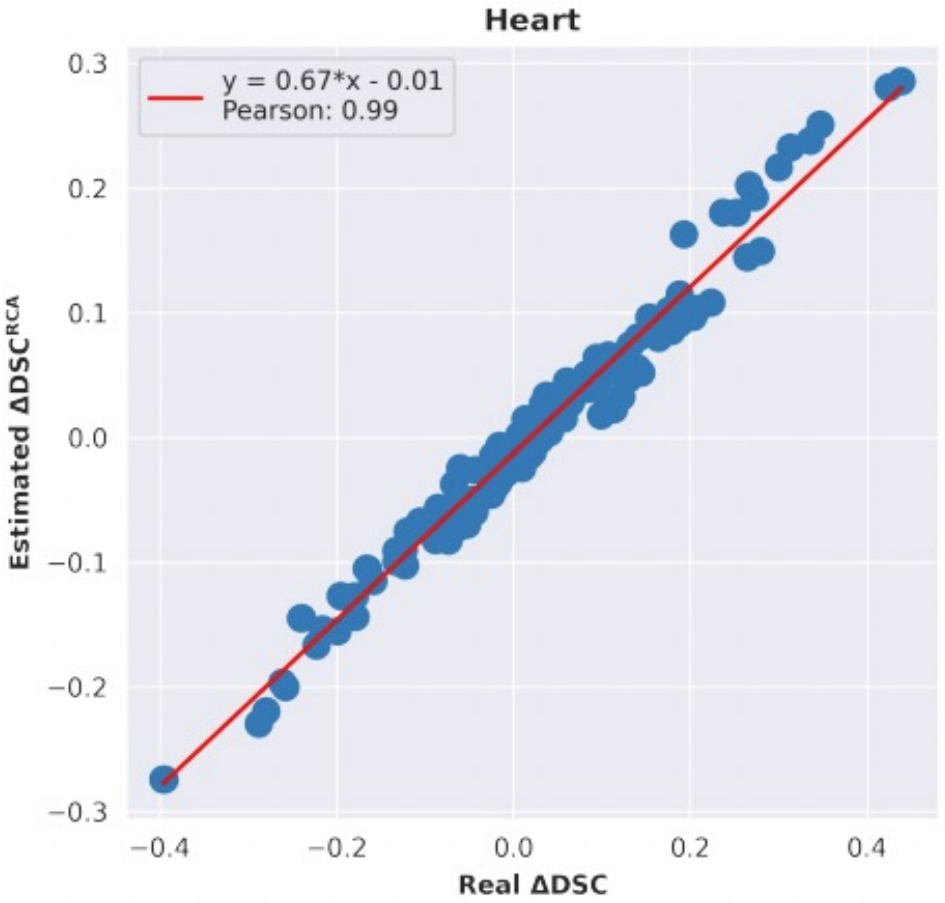


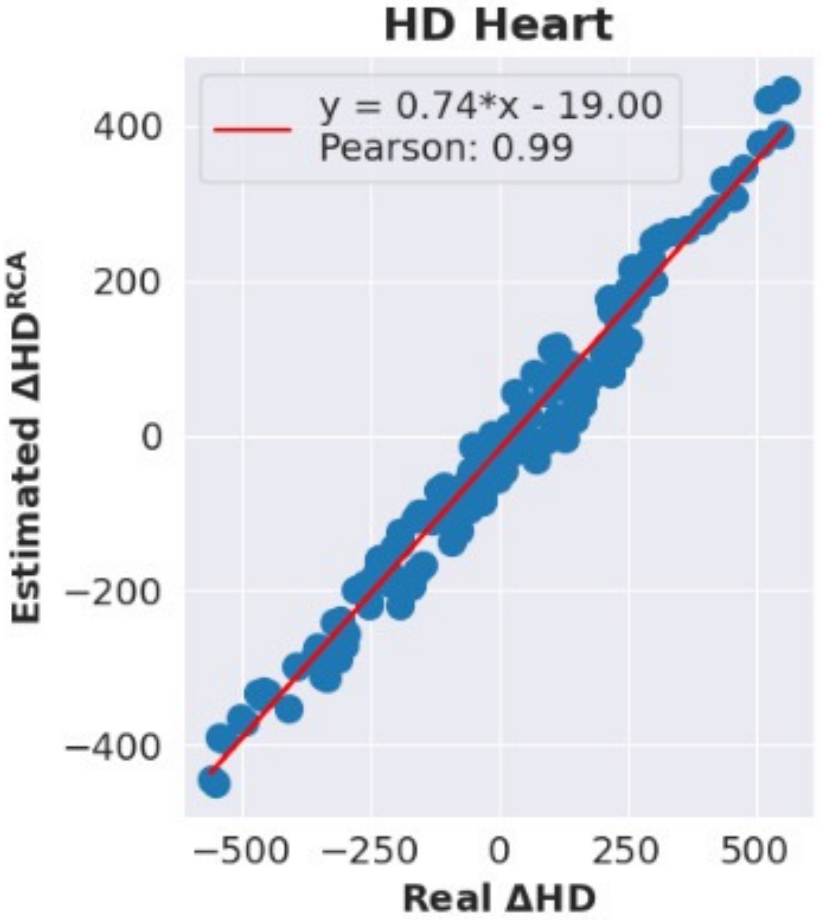
DSC estimated without GT



$$\Delta DSC^{RCA} = DSC^{RCA}_{A=M} - DSC^{RCA}_{A=F}$$

Results: simulated scenario





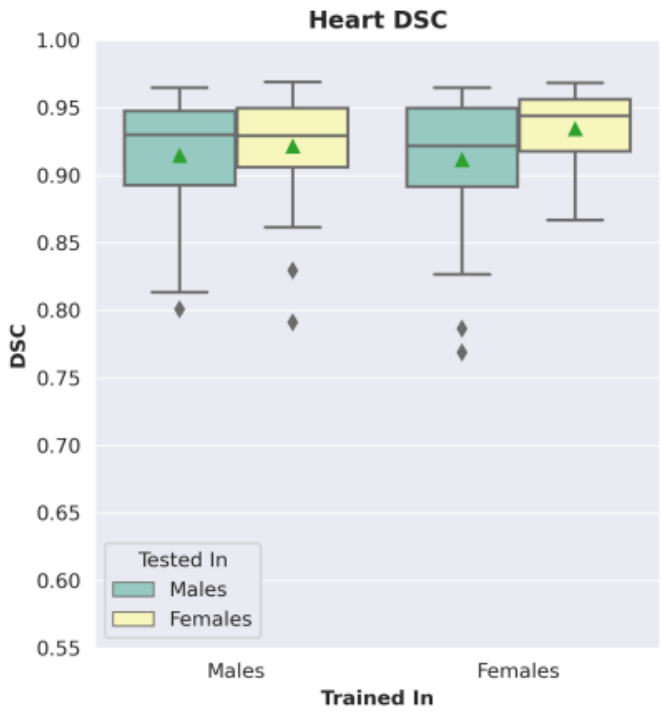
Results: auditing real models

We consider one model trained on 100% males and another on 100% females.

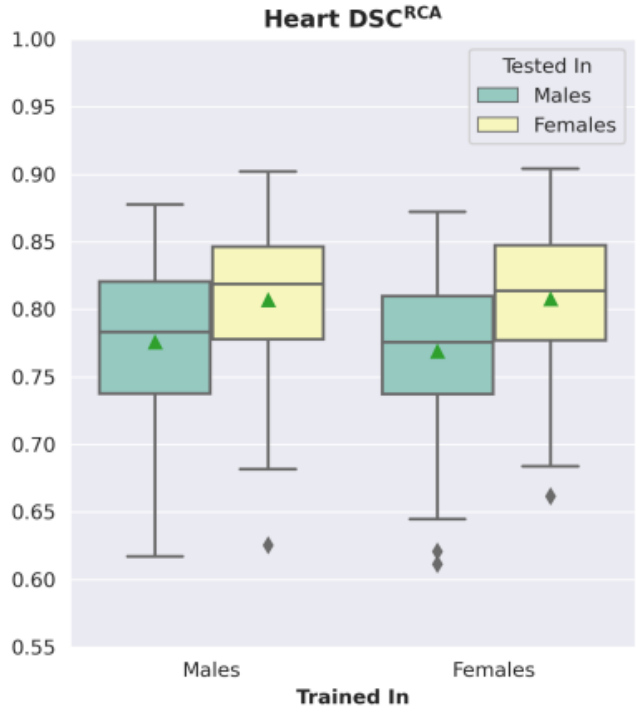
We tested both models on male and female images separately

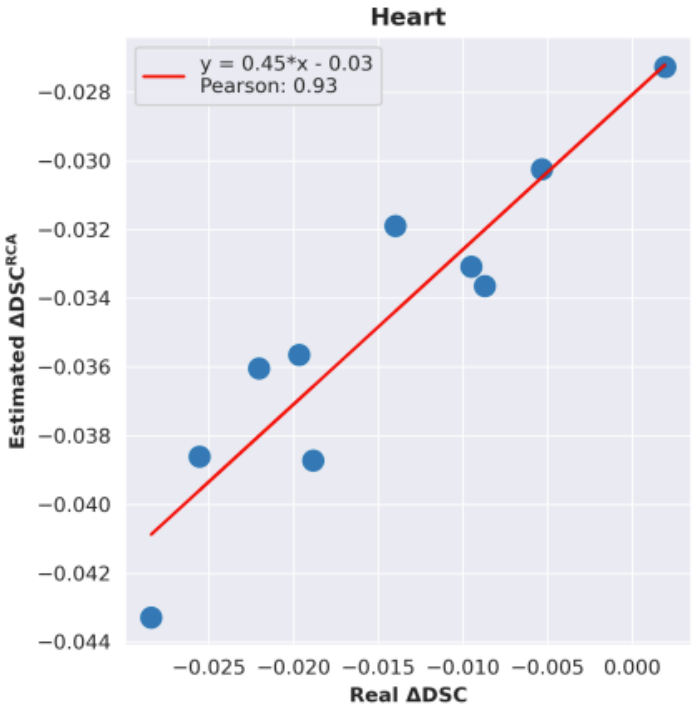
Surprisingly, we found that both models tend to perform better on female than male patients

DSC estimated with GT

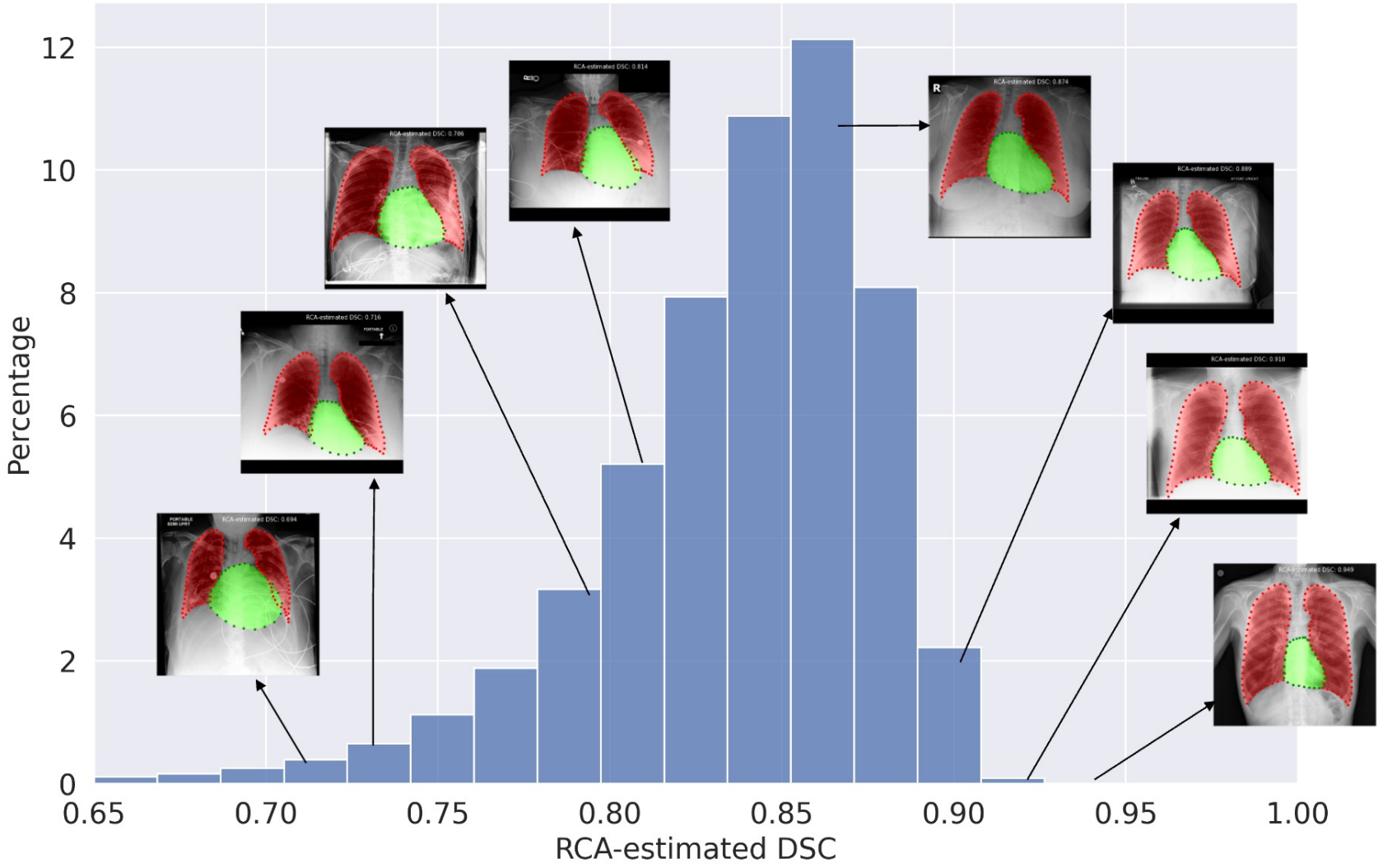


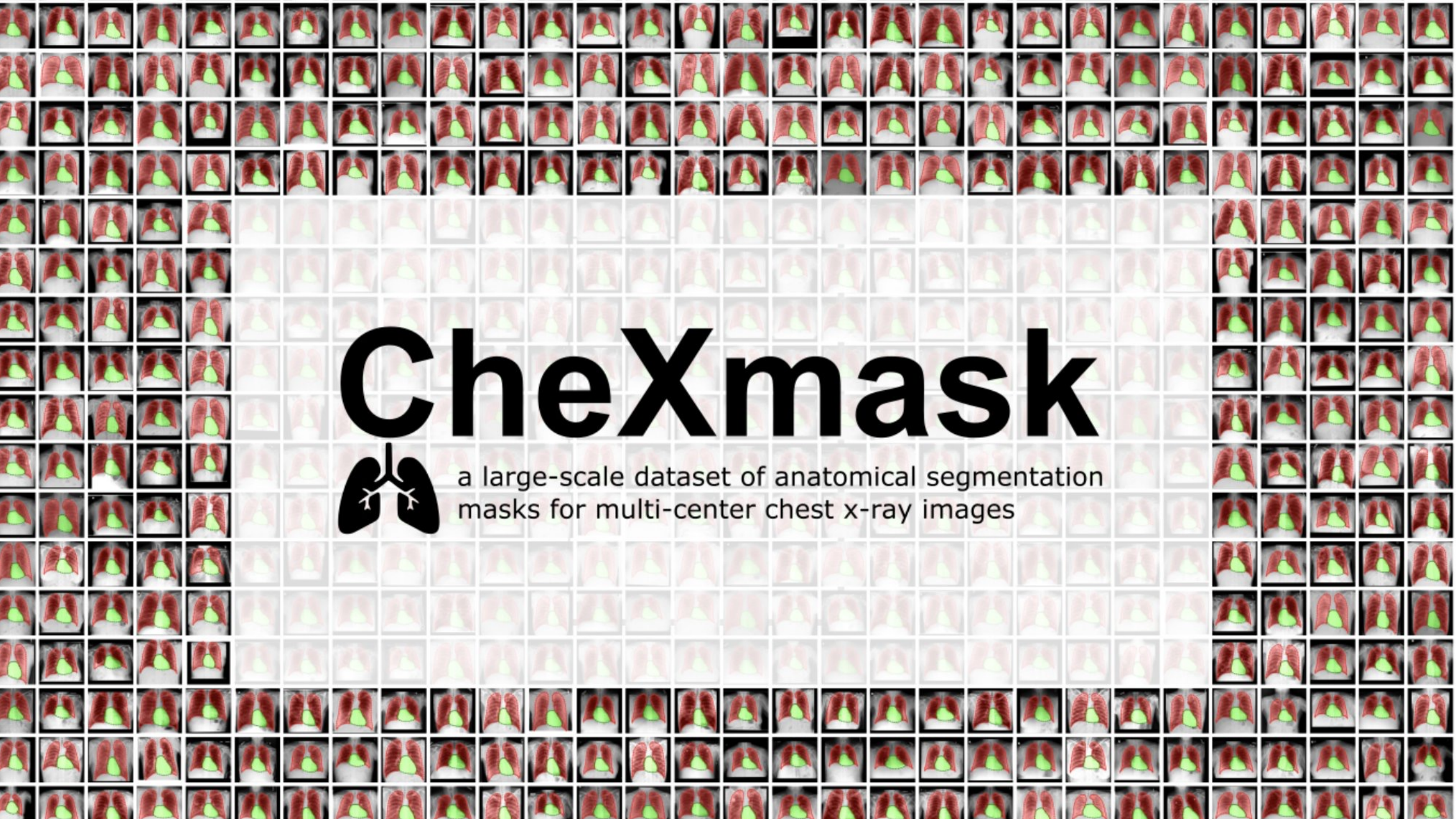
Estimated without GT





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



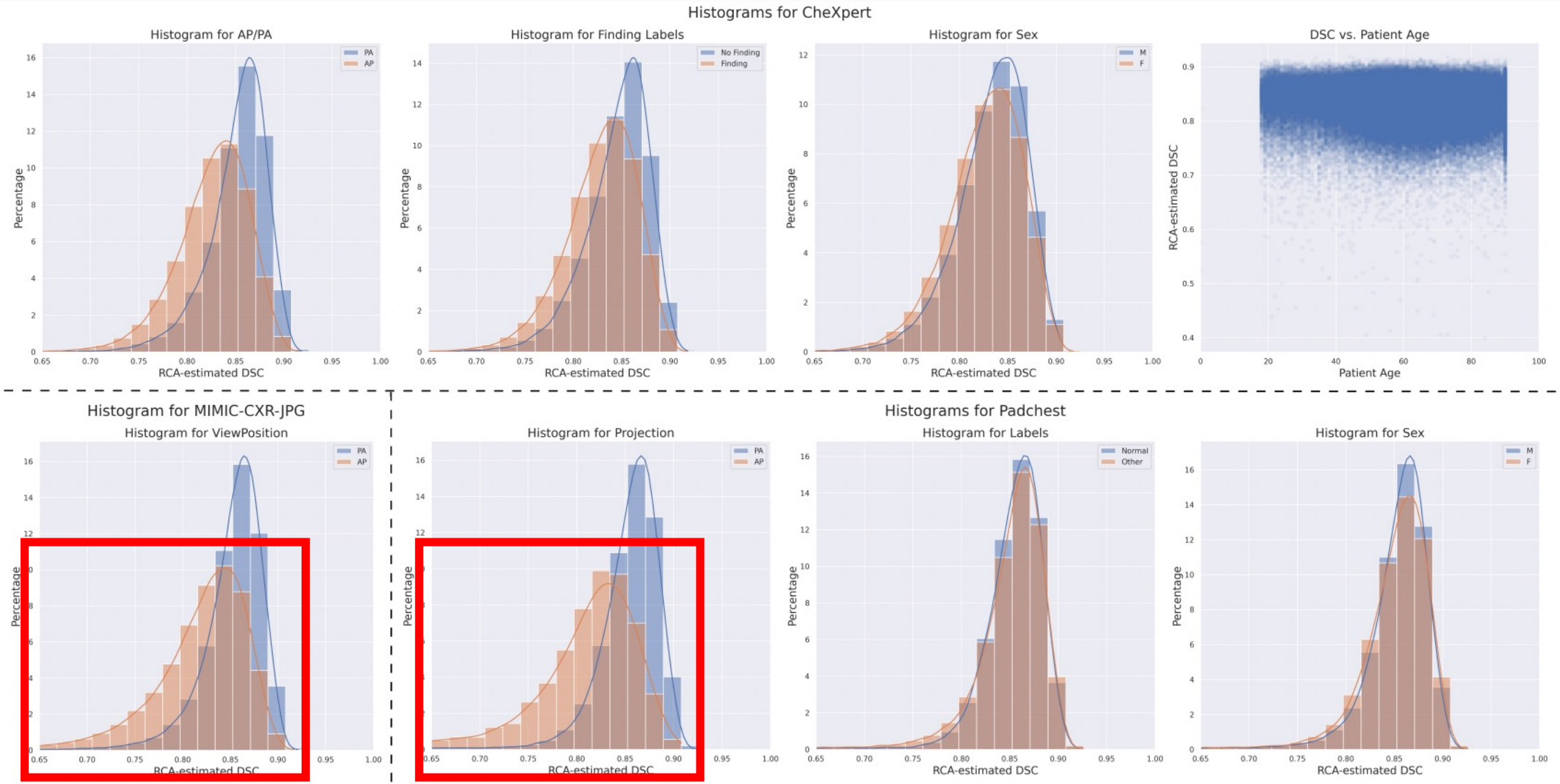

CheXmask
a large-scale dataset of anatomical segmentation
masks for multi-center chest x-ray images

UNL CONICET sinc(i)

HOSPITAL ITALIANO de Buenos Aires

https://tinyurl.com/chexmask

UBD at large scale: CheXmask



AP images tend to come from hospitalized patients, who are more **difficult to position in standard views** and usually **include artifacts** or cables

Active Research Areas

- Anatomical segmentation
- Domain Adaptation and Generalization
- Model Calibration in Biomedical Image Analysis
- Fairness in ML for Biomedical Image Analysis
- Learning representations of life

Domain adaptation and generalization

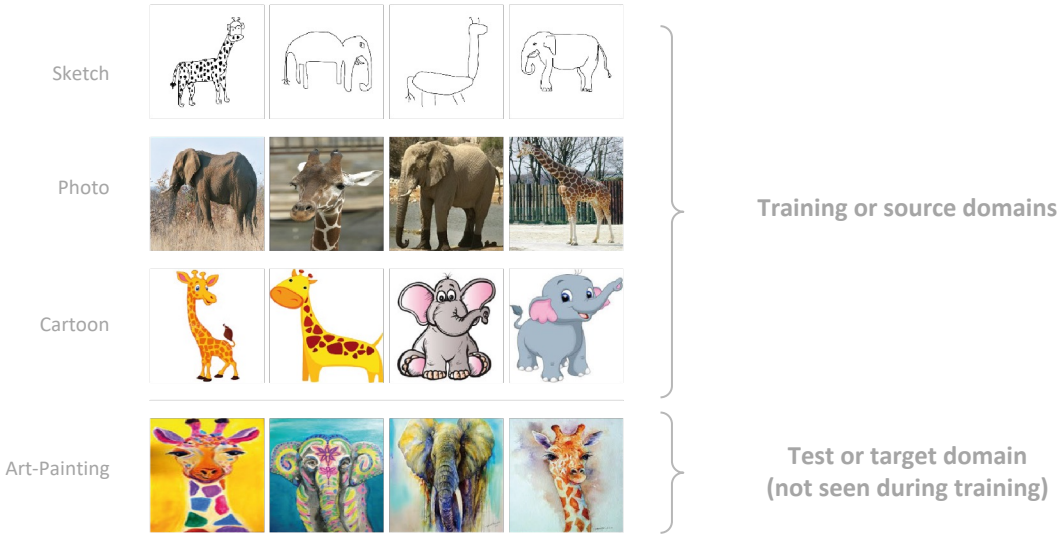
How can we obtain models that generalize to unseen image domains?

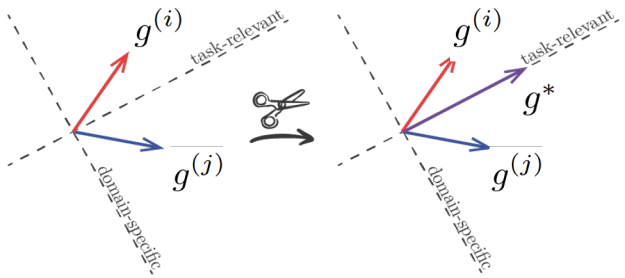
LIAA
Laboratorio de Inteligencia Artificial Aplicada

Gradient Surgery for Domain Generalization

Mansilla L, Echeveste R, Milone D, ***Ferrante E.***

ICCV 2021

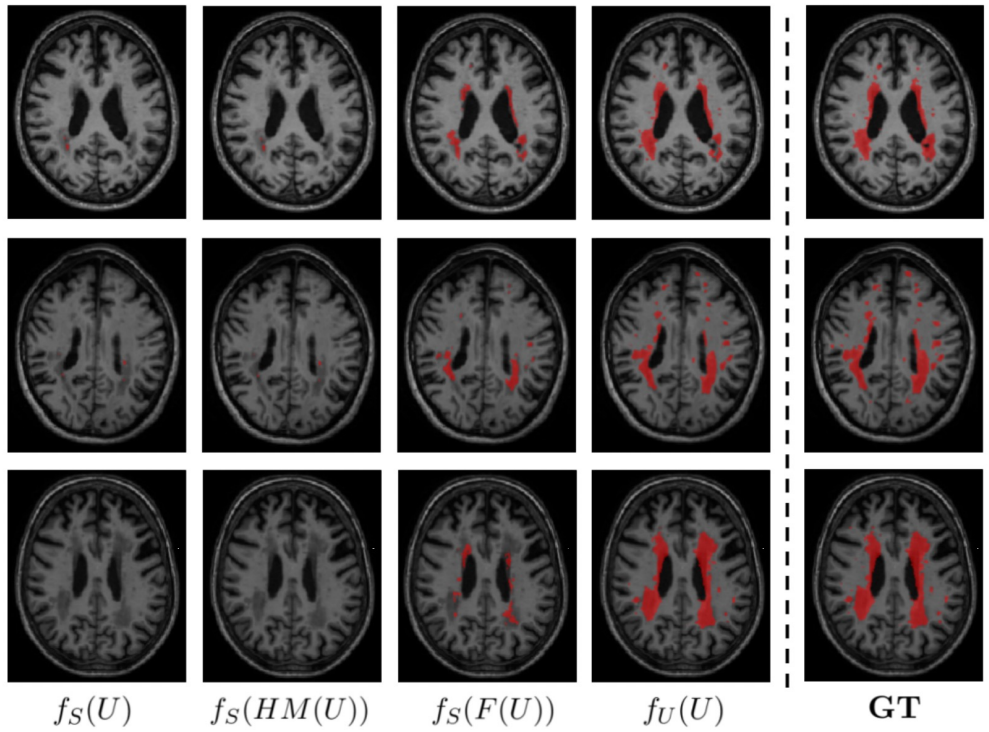




Unsupervised Domain Adaptation via CycleGAN for White Matter Hyperintensity Segmentation in Multicenter MR Images

Palladino J, Fernandez Slezak D, ***Ferrante E.***

SIPAIM-MICCAI Symposium 2020



Model Calibration in Biomedical Image Analysis

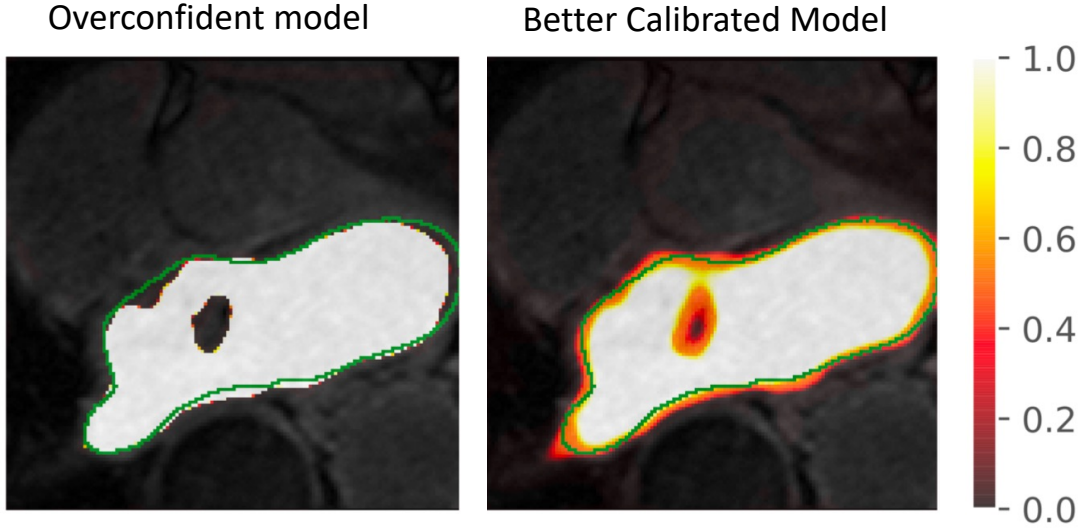
How can we obtain calibrated posteriors when training ML models for classification and segmentation?

ÉTS Le génie pour l'industrie | ÉCOLE DE TECHNOLOGIE SUPÉRIEURE Université du Québec

Orthogonal Ensemble Networks for Biomedical Image Segmentation

Larrazabal A, Martinez C, Doltz J, ***Ferrante E***.

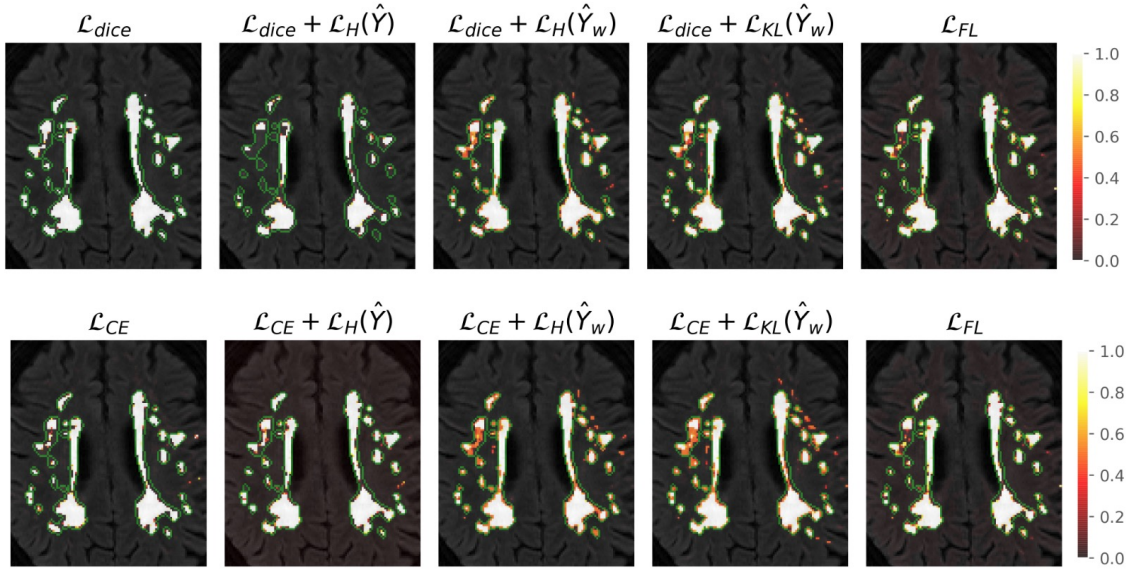
MICCAI 2021



Maximum Entropy on Erroneous Predictions: Improving model calibration for medical image segmentation

Larrazabal A, Martinez C, Doltz J,* ***Ferrante E****.

MICCAI 2023



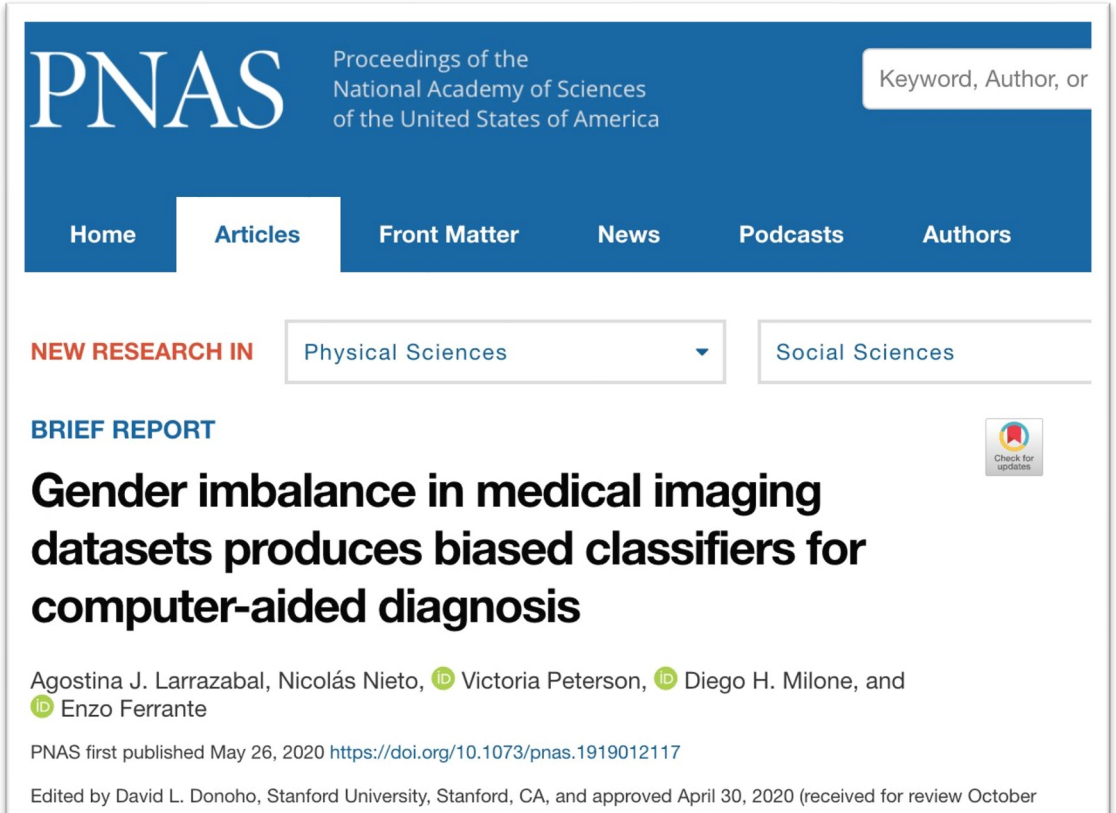
Fairness in ML for Biomedical Image Analysis

Characterizing and mitigating fairness issues in ML models for biomedical image analysis

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Larrazabal A., Nieto N., Peterson V., Milone D., ***Ferrante E.***

Proceedings of the National Academy of Sciences (PNAS) 2020

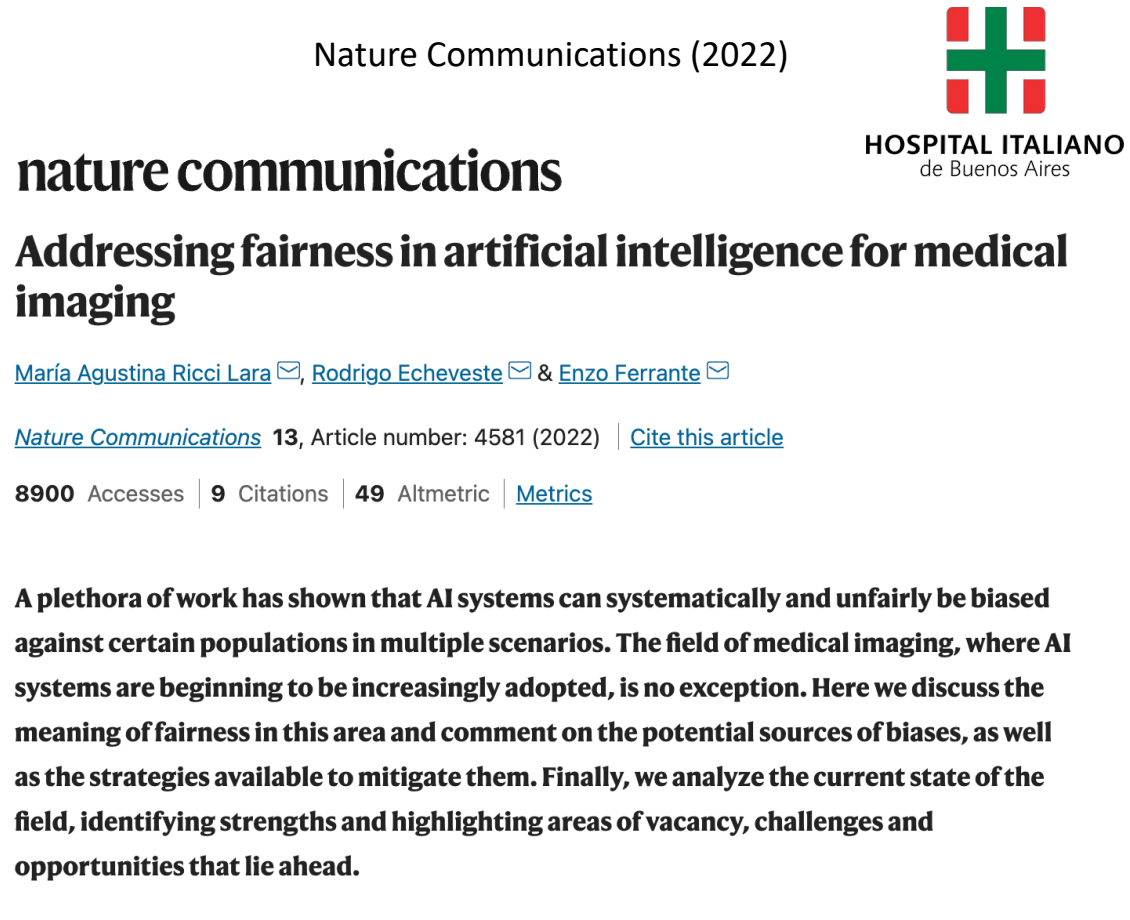


Addressing fairness in artificial intelligence for medical imaging

Ricci Lara A., Echeveste R., ***Ferrante E.***

Nature Communications (2022)

HOSPITAL ITALIANO
de Buenos Aires



Learning representations of life

Linking image derived phenotypes with genetic information via ML methods (imaging genetics)

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

Gaggion N, Ariel F, Daric V, Lambert E, Legendre S, Roule T, Camoirano A, Milone D, Crespi M, Blein T, ***Ferrante E.***

GigaScience

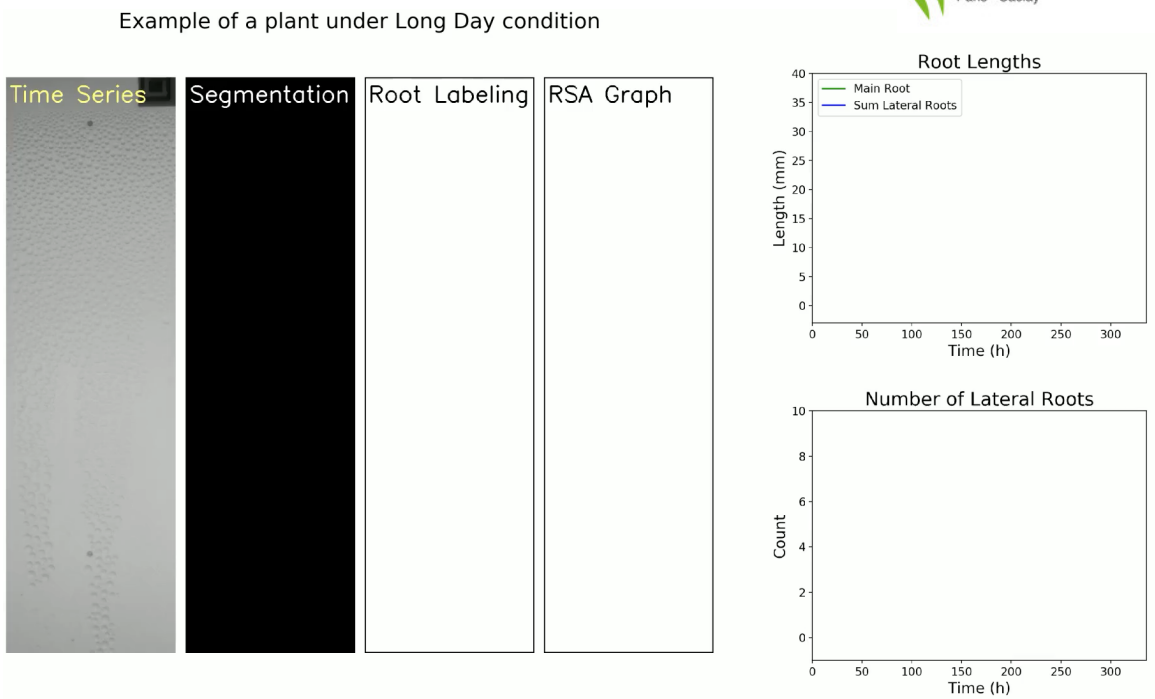
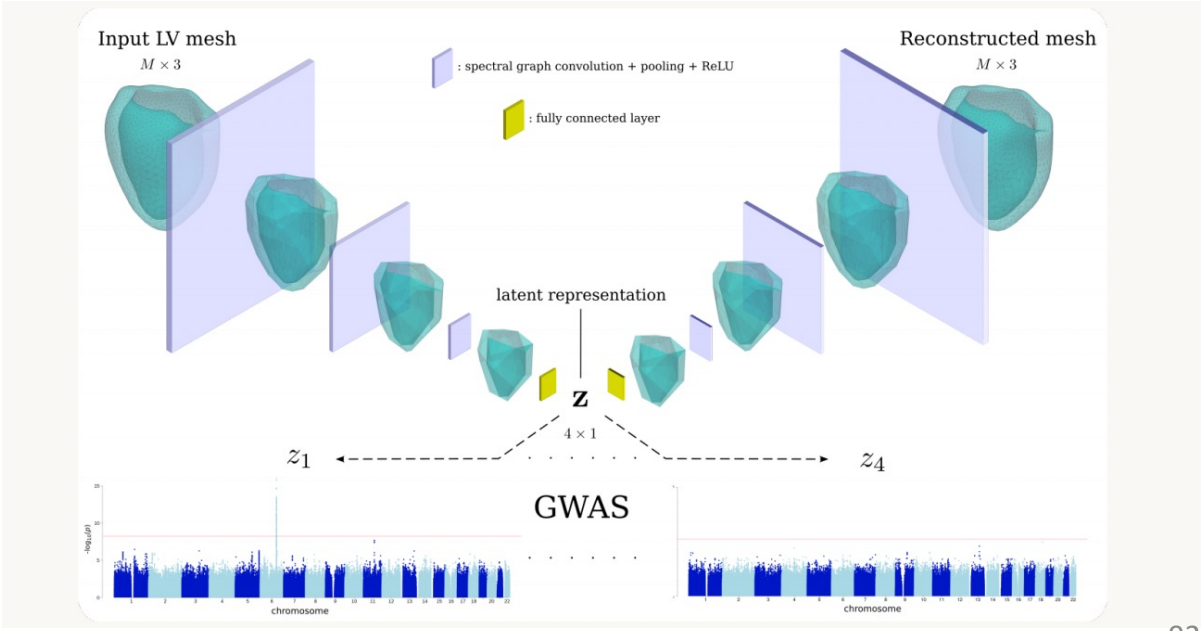


Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in CMR images

Bonazzola R, ***Ferrante E ,****Ravikumar N, Attar R, Syeda-Mahmood T, Frangi A.*

MICCAI 2023



Research supported by: