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 training of deep learning models with heterogeneous labels from public and multi-center datasets.

Index Terms— anatomical segmentation, landmark-based models, missing annotations, graph neural networks

1. INTRODUCTION

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

Several methods have been proposed to independently address the problems of domain shift [5, 8, 9] and heterogeneous labels [10, 6, 11] in medical image segmentation. As for the joint problem, Dorent and coworkers [7] proposed a framework which combines a variational formulation to cope with heterogeneous labels, with conventional techniques based on data augmentation, adversarial learning, and pseudo-healthy image generation to address domain shift. In this work, we argue that landmark based segmentation methods like the HybridGNet [12, 13] can naturally handle these scenarios, as they incorporate prior knowledge about the expected anatomy, without additional burden related to data augmentation, adversarial training, or image generation. We first provide empirical evidence showing how widely used pixel-level approaches drastically fail to learn robust segmentation models using heterogeneous labels from multicentric datasets, while HybridGNet can naturally handle this problem, avoiding memorization issues. Further analysis of the latent spaces learned by the different architectures, indicates that generative landmark-based approaches like HybridGNet tend to learn more invariant representations, which helps to improve the robustness with respect to domain memorization.

2. METHODS AND EXPERIMENTS

Problem statement and experimental setup: We explore anatomical segmentation of lung, heart and clavicles in a multi-center database of chest X-ray images, with heterogeneous labels. The database is composed of 4 publicly available datasets (JSRT [14], Padchest [15], Montgomery [16] and Shenzhen [17]), which originally provide pixel level annotations. Since the proposed method employs landmark-based annotations, we adopted the publicly available *Chest X-ray landmark dataset*, which provides landmarks for 3 different organs from the aforementioned databases (github.com/ngaggion/Chest-xray-landmark-dataset).

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
N. Gaggion, M. Vakalopoulou, D. H. Milone & E. Ferrante; "Multi-center anatomical segmentation with heterogeneous labels via landmark-based models"
arXiv preprint arXiv:2211.07395, 2022.

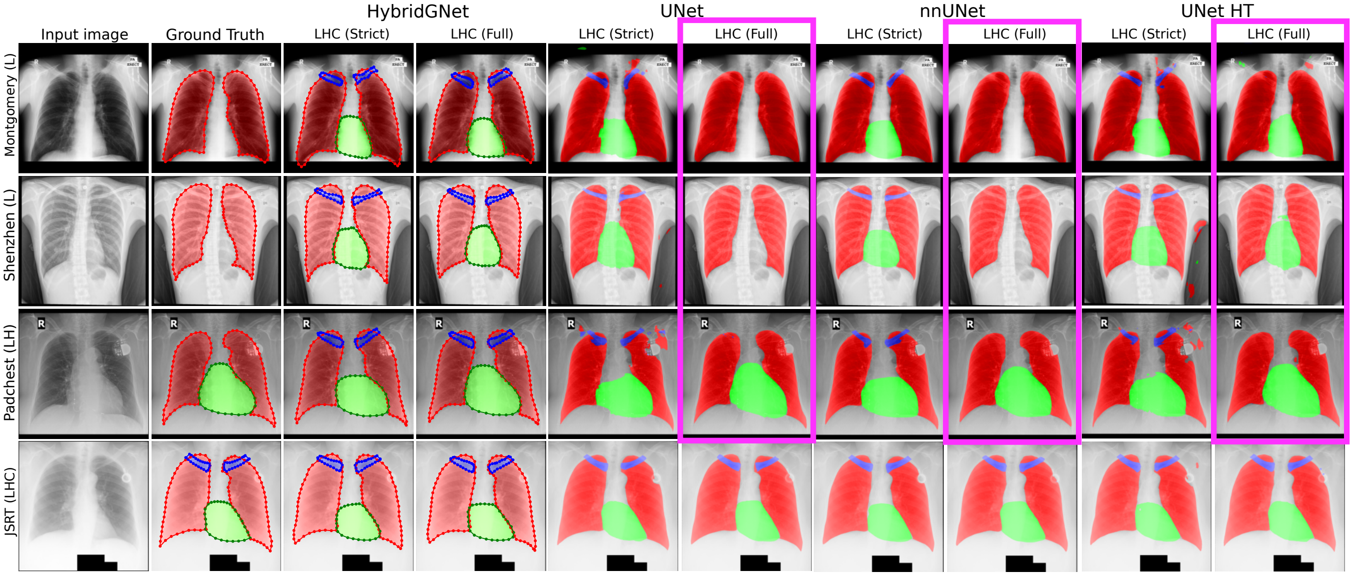

Fig. 1. **Qualitative examples**. Examples from 4 datasets (rows) segmented with different methods and annotation settings (columns). Note how pixel UNet and nnUNet (in pink) fail completely to segment structures that are not presented in the corresponding dataset when trained with heterogeneous labels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

Images from JSRT (246 subjects) include annotations for lungs, heart and clavicle (LHC); Padchest (137 subjects) include lungs and heart (LH); while Montgomery (138 subjects) and Shenzhen (390 subjects) include only lung (L). To evaluate each domain separately, we divide the datasets into 80% train/val and 20% test partitions.

HybridGNet: HybridGNet is a landmark-based segmentation model, where the ROIs are encoded as anatomical graphs representing the organ contour. It follows an encoder-decoder architecture that combines standard convolutions for image encodings, with graph generative models to extract anatomically plausible representations directly from images. The model is trained to minimize the mean squared error (MSE) between the predicted node positions and the ground truth coordinates. Pixel-level masks are then generated by filling in the contours. See [12, 13] for more details about HybridGNet.

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

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

Model	Trained in	Montgomery			Shenzhen			Padchest						JSRT								
		Lungs			Lungs			Lungs			Heart			Lungs			Heart			Clavicles		
		MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD	MSE	Dice	HD
HybridGNet	L (Both)	128.4	0.97	27.4	142.9	0.97	32.5	152.1	0.96	36.6	-	-	-	120.2	0.97	33.0	-	-	-	-	-	-
	LH (Strict)	295.6	0.95	38.5	264.4	0.95	43.1	201.8	0.95	41.0	351.7	0.94	36.8	155.7	0.97	36.3	382.4	0.94	34.5	-	-	-
	LH (Full)	104.4	0.97	27.8	145.9	0.97	32.8	187.0	0.96	38.8	342.0	0.94	36.0	124.8	0.97	31.0	375.8	0.94	33.8	-	-	-
	LHC (Strict)	571.5	0.94	49.2	486.0	0.93	51.1	480.2	0.92	56.8	1317.0	0.87	70.9	139.4	0.97	33.9	413.1	0.93	35.7	83.4	0.84	20.8
	LHC (Full)	110.7	0.97	26.2	148.7	0.96	33.2	172.0	0.96	37.1	235.0	0.94	30.5	122.2	0.97	31.1	390.8	0.94	34.2	101.1	0.82	22.9
UNet	L (Both)	-	0.98	53.3	-	0.98	53.3	-	0.96	85.2	-	-	-	-	0.95	99.9	-	-	-	-	-	-
	LH (Strict)	-	0.96	73.4	-	0.96	131.2	-	0.96	89.4	-	0.93	80.5	-	0.95	101.1	-	0.94	51.7	-	-	-
	LH (Full)	-	0.97	70.0	-	0.96	74.0	-	0.96	94.8	-	0.89	75.2	-	0.95	103.8	-	0.94	51.8	-	-	-
	LHC (Strict)	-	0.90	182.4	-	0.91	217.6	-	0.89	227.1	-	0.87	220.8	-	0.97	67.5	-	0.94	67.6	-	0.93	44.8
	LHC (Full)	-	0.97	72.5	-	0.97	72.7	-	0.96	106.1	-	0.91	76.7	-	0.98	63.0	-	0.94	54.5	-	0.91	49.2
nnUNet	L (Both)	-	0.98	26.1	-	0.97	32.3	-	0.96	40.2	-	-	-	-	0.98	31.1	-	-	-	-	-	-
	LH (Strict)	-	0.97	45.1	-	0.96	53.3	-	0.96	40.5	-	0.95	32.9	-	0.98	26.3	-	0.95	29.0	-	-	-
	LH (Full)	-	0.98	34.0	-	0.97	35.2	-	0.96	37.4	-	0.94	34.5	-	0.98	28.6	-	0.95	30.9	-	-	-
	LHC (Strict)	-	0.93	120.5	-	0.92	121.0	-	0.93	83.6	-	0.89	80.4	-	0.98	35.0	-	0.95	29.5	-	0.95	14.1
	LHC (Full)	-	0.98	25.3	-	0.97	35.2	-	0.96	39.3	-	0.95	33.3	-	0.98	35.6	-	0.95	31.1	-	0.93	38.1
UNet HT	L (Both)	-	0.97	46.9	-	0.97	78.7	-	0.96	70.0	-	-	-	-	0.95	106.1	-	-	-	-	-	-
	LH (Strict)	-	0.96	74.8	-	0.96	131.6	-	0.96	74.9	-	0.94	80.2	-	0.95	105.6	-	0.94	58.4	-	-	-
	LH (Full)	-	0.98	60.5	-	0.97	57.8	-	0.96	87.9	-	0.93	125.6	-	0.95	100.2	-	0.94	59.4	-	-	-
	LHC (Strict)	-	0.91	168.5	-	0.91	204.8	-	0.90	223.6	-	0.87	199.0	-	0.98	78.1	-	0.94	82.2	-	0.94	28.3
	LHC (Full)	-	0.97	72.6	-	0.97	77.1	-	0.96	66.1	-	0.93	87.8	-	0.98	55.2	-	0.94	47.6	-	0.94	43.6

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

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

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

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

3. RESULTS, DISCUSSION AND CONCLUSIONS

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

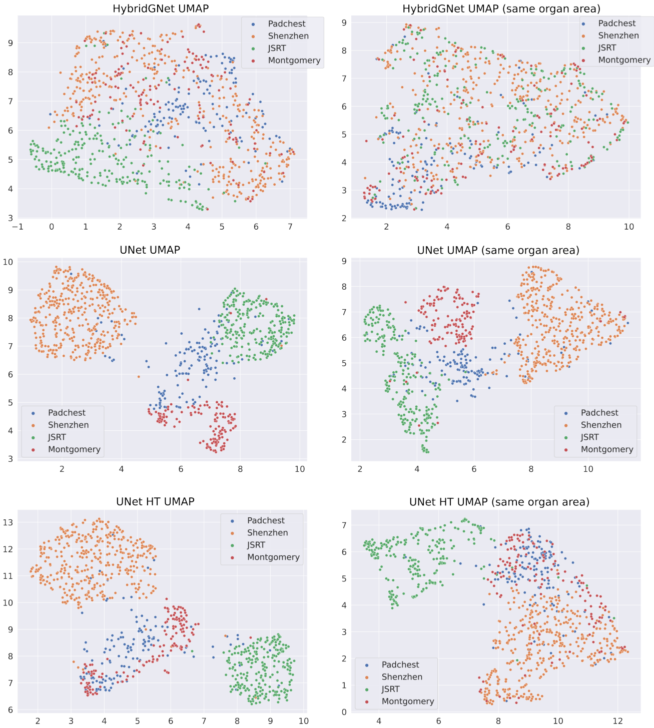

Fig. 2. **Latent space inspection via UMAP embeddings:** Different datasets are shown in colors, allowing to see how both UNet and UNet HT models tend to clusterize images per dataset, while HybridGNet doesn't, explaining the improved robustness to domain-label memorization.

lung labels. On the contrary, HybridGNet is forced to predict all ROIs by construction, using the anatomical priors encoded in the learned latent space to infer the organ position, even if it was not present in that particular dataset.

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

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

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

Table 1 shows quantitative results for the different training scenarios, models and test sets. In general, we see that incorporating the Full dataset through heterogeneous labels improve performance when compared to the Strict case, which only employs images where all annotations are available. Performance is only evaluated for ROIs available as ground-truth in every dataset. As shown in Figure 1, for the heterogenous settings (**LHC (Full)** and **LH (Full)**) UNet and nnUNet drastically fail at segmenting structures that were not present during training in the corresponding dataset. This is due to the memorization issues and conflicting background/organ labels in different domains. This would imply a Dice of 0 for UNet and nnUNet if ground-truth were available for evaluation.

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

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

4. COMPLIANCE WITH ETHICAL STANDARDS

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

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] Enzo Ferrante, Puneet Kumar Dokania, Rafael Marini Silva, and Nikos Paragios, "Weakly supervised learning of metric aggregations for deformable image registration," *IEEE journal of biomedical and health informatics*, vol. 23, no. 4, pp. 1374–1384, 2018.

[2] Juan J Cerrolaza et al., "Computational anatomy for multi-organ analysis in medical imaging: A review," *Medical image analysis*, vol. 56, pp. 44–67, 2019.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015.

[4] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, 2021.

[5] Hao Guan and Mingxia Liu, "Domain adaptation for medical image analysis: a survey," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 3, 2021.

[6] Jana Kemnitz et al., "Combining heterogeneously labeled datasets for training segmentation networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 276–284.

[7] Reuben Dorent, Thomas Booth, Wenqi Li, Carole H Sudre, Sina Kafiabadi, Jorge Cardoso, Sebastien Ourselin, and Tom Vercauteren, "Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets," *Medical image analysis*, vol. 67, pp. 101862, 2021.

[8] Julián Alberto Palladino, Diego Fernandez Slezak, and Enzo Ferrante, "Unsupervised domain adaptation via cyclegan for white matter hyperintensity segmentation in multicenter mr images," in *16th International Symposium on Medical Information Processing and Analysis*. SPIE, 2020, vol. 11583, p. 1158302.

[9] Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy, "Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology," in *MICCAI – MICCAI 2022*. 2022, pp. 120–129, Springer Nature Switzerland.

[10] Olivier Petit et al., "Handling missing annotations for semantic segmentation with deep convnets," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 20–28. Springer, 2018.

[11] Gregory Filbrandt et al., "Learning from partially overlapping labels: Image segmentation under annotation shift," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*, pp. 123–132. Springer, 2021.

[12] Nicolás Gaggion, Lucas Mansilla, Diego H. Milone, and Enzo Ferrante, "Hybrid graph convolutional neural networks for landmark-based anatomical segmentation," in *MICCAI*. 2021, Springer International Publishing.

[13] Nicolás Gaggion, Lucas Mansilla, Candelaria Mosquera, Diego H. Milone, and Enzo Ferrante, "Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis," *IEEE Transactions on Medical Imaging*, 2022.

[14] Junji Shiraishi et al., "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, 2000.

[15] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *Medical image analysis*, vol. 66, pp. 101797, 2020.

[16] Sema Candemir and others., "Lung segmentation in chest radiographs using anatomical atlases with non-rigid registration," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014.

[17] Stefan et al Jaeger, "Automatic tuberculosis screening using chest radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, 2014.

[18] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision*. IEEE, 2016.

[19] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.