Deep Transfer Learning for Segmentation of Anatomical Structures in Chest Radiographs

Hugo Oliveira, Jefersson A. dos Santos
Department of Computer Science
Universidade Federal de Minas Gerais, Brazil
Belo Horizonte, Minas Gerais, 31270–901
Email: {oliveirahugo,jefersson}@dcc.ufmg.br

***Abstract*—Segmentation of anatomical structures in Chest Posterior-Anterior Radiographs is a classical task on biomedical image analysis. Deep Learning has been widely used for detection and diagnosis of illnesses in several medical image modalities over the last years, but the portability of deep methods is still limited, hampering the reusability of pre-trained models in new data. We address this problem by proposing a novel method for Cross-Dataset Transfer Learning in Chest X-Ray images based on Unsupervised Image Translation architectures. Our Transfer Learning approach achieved Jaccard values of 88.20% on lung field segmentation in the Montgomery Set by using a pre-trained model on the JSRT dataset and no labeled data from the target dataset. Several experiments in unsupervised and semi-supervised transfer were performed and our method consistently outperformed simple fine-tuning when a limited amount of labels is used. Qualitative analysis on the tasks of clavicle and heart segmentation are also performed on Montgomery samples and pre-trained models from JSRT dataset. Our secondary contributions encompass several experiments in anatomical structure segmentation on JSRT, achieving state-of-the-art results in lung field (96.02%), heart (89.64%) and clavicle segmentation (87.30%).**

I. Introduction

Radiology has been an useful tool for the detection of several kinds of illnesses for over a century. Radiation allows physicians to peek through human tissue without having to perform invasive procedures. Computer-Aided Diagnosis (CAD) systems have followed the advances in radiology during the last decades, providing automated second opinions for physicians.

An important radiological exam on a physician's toolkit is the Posterior-Anterior (PA) Chest Radiograph, more commonly called Chest X-Ray (CXR). CXRs are the single most acquired medical image modality [1]. These images are important for the detection and diagnosis of several pulmonary diseases, such as tuberculosis, interstitial lung disease, pulmonary embolism and lung cancer [1]. Lung field segmentations can be used as important markers for the detection of such illnesses, as they often affect the lung's shape and volume. Therefore, the automation of lung segmentation tasks is an active research area in biomedical image analysis. Other health abnormalities which may be detected using CXRs are bone fractures in the thoracic area and health diseases that affect heart shape and/or volume. Therefore, we explore three tasks on CXR images: lung field segmentation, clavicle segmentation and heart segmentation. Examples of ground truths for these three tasks can be seen in Figure 1.

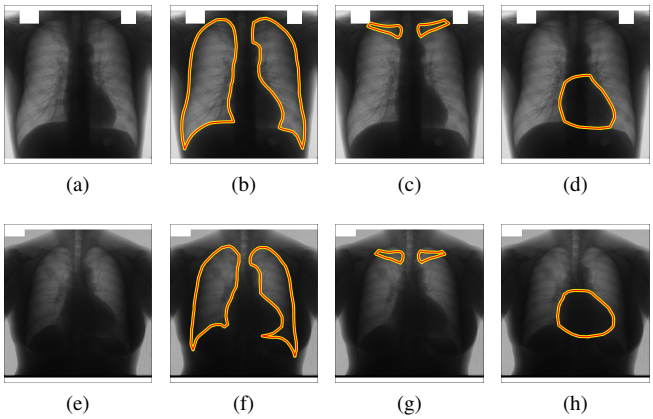

Fig. 1. Example of ground truths for two CXRs. (a,e) Original image. (b,f) Lungs. (c,g) Clavicles. (d,h) Heart.

As in most Computer Vision applications, Deep Learning has been widely used in the area of biomedical image analysis, including CXRs. A major problem in Deep Neural Networks (DNNs) is the lack of generalization between different datasets, image modalities and/or tasks. There are a lot of CXR data publicly available on the Internet, but very few labeled datasets are within reach in the public domain. Therefore, a technique that transfers the knowledge obtained in a labeled dataset to the images of another and is also capable of learning from unlabeled data would be helpful for producing more general representations over CXR images.

The main contribution of this work is the proposal of a novel Transfer Learning technique for knowledge transfer, as detailed in Section IV-C. Secondary contributions include:

- Evaluation of DNN architectures for semantic segmentation of lungs, clavicles and heart in supervised, semi-supervised and unsupervised settings;
- State-of-the-art results in CXR segmentation tasks;
- Assessment of the superiority in stability of Deep Learning methods compared to classical machine learning approaches and to secondary human annotations.

The remaining sections of this work are organized as follows: Section II shows related research projects in CXR

segmentation, including deep-based ones; Section III presents the DNN architectures used in our work; Section IV describes the proposed approach for cross-dataset knowledge transfer; Section V shows the experimental procedures, datasets and evaluation metrics chosen for this work; Section VI presents and compares the results obtained in our experiments; and Section VII contains our final remarks about the results.

II. Related Work

As it is a classical problem in the area of biomedical image analysis, CXR segmentation has been approached by using traditional image processing techniques. Ginneken *et al.* [1] performed tests with several automated segmentation algorithms and comparisons of the gold standard with other human-generated labels. Their best automated results were a mean Jaccard value of 94.9% for lung field segmentation, 87.8% for heart segmentation and 73.6% for the clavicles, while human observers achieved 94.6%, 87.8% and 89.60% for these tasks, respectively. Ginneken *et al.* [1] found no significant differences between accuracies of the automated methods and the labels obtained by the human observers in most tasks, which highlighted inter-observer inaccuracies. Candemir *et al.* [2] used Atlas to perform lung field segmentation in CXRs, achieving Jaccard scores of 95.4%. Hogeweg *et al.* [3] focused on clavicle segmentations using Hybrid Dynamic Programming/Active Shape Model/Pixel Classification (HDAP), achieving 86% Jaccard.

During the last year, some works already introduced DNNs in the area of CXR segmentation [4], [5]. Dai *et al.* [4] developed an adversarial architecture for the task and reported Jaccard scores of 94.7% for lung field segmentation and 86.6% for heart segmentation. Novikov *et al.* [5] proposed a change in the existing U-net architecture [6] (Section III-A2) – called InvertedNet – to perform segmentation on CXR data, achieving Jaccards of 95.1% for the lungs, 87.1% for the heart and 87% for the clavicles.

Semi-supervised transfer using regression objectives for the unlabeled data/domain have been proposed in the literature [7]. These methods rely mostly on supervised loss functions being used together with unsupervised losses. Early experimental evaluations of this work pointed that Chen and Chien's approach [7] for transfer in classification settings achieved subpar segmentation results in the CXR image domain. We observed that the source domain achieves good segmentation results, while the target domain is not optimized correctly.

III. Deep Semantic Segmentation

Most DNNs for image analysis have been based on convolution operations. Vanilla implementations of Convolutional Neural Networks (CNNs) [8] are essentially stackings of three types of layers: Convolutional, Pooling and Fully Connected (FC) layers. Convolutional and Pooling layers are often stacked in the beginning of these networks and serve as learnable feature extractors, while FC layers play the role of the classifier at the end of the network.

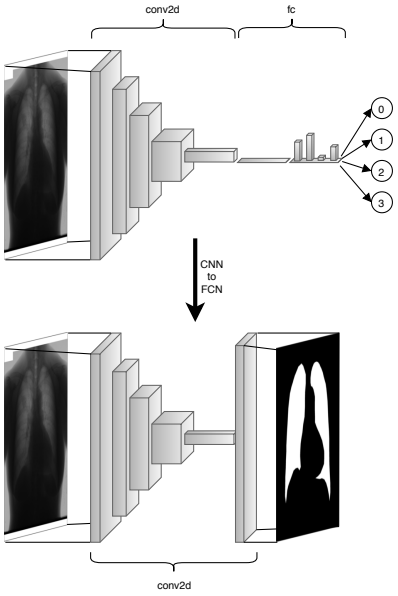

Fig. 2. Transforming fully connected layers into convolution layers enables a classification net to output dense predictions. Each conv2d box corresponds to multi-channel convolutions followed by downsampling (pooling). Each vector represents a feature vector in an FC layer. Adapted from [9].

A. Deep Semantic Segmentation

Segmentation has been an active research topic in the area of biomedical image analysis for decades, as it is a rather common preprocessing and evaluation tool for several medical applications. Traditionally this field of research uses several active contour, clustering, atlas and interactive methods. More recently, with the advent of DNNs, segmentation in Computer Vision has become dominated by semantic deep-based methods. Therefore, several algorithms comprising the state-of-the-art of deep semantic segmentation were used in our experimental setup. Most of these architectures are discriminative models based on improvements over CNNs and Fully Convolutional Networks (FCNs) [9].

1) Fully Convolutional Networks: The most basic segmentation architectures are the FCNs [9], which are often based on CNN models like AlexNet [8] and VGG [10] adapted to dense prediction (Figure 2). An FCN can be understood as a patchwise approach, wherein each pixel in an image is a sample. Whole image fully convolutional training is identical to patchwise training where each batch consists of all the pixels in an image or set of images. Replacing fully connected layers in a CNN by convolutional layers and adding a spatial loss produces an efficient machine for end-to-end dense learning [9], as can be seen in Figure 2.

2) U-nets: Ever since the appearance of FCNs, several attempts to mitigate the vanishing gradient problem have been proposed, most of them relying in alternative paths for information flow [11], [12]. Skip connections are the most common way to create these alternative paths, serving as highways for backpropagation to reach earlier layers in the network without passing through all the layers in front of

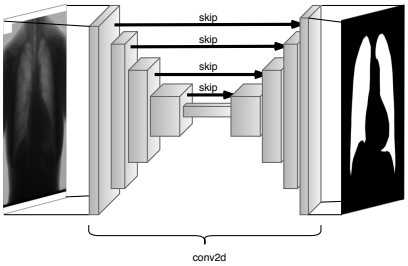

Fig. 3. U-net architecture. Each conv2d box corresponds to multi-channel convolutions followed by downsampling or upsampling. Arrows denote the skip connections between symmetric layers. Adapted from [6].

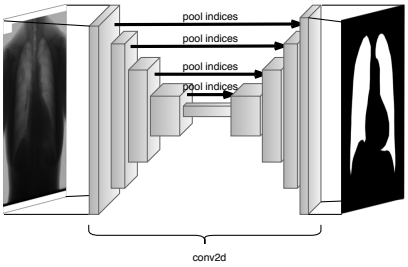

Fig. 4. An illustration of the SegNet architecture. Arrows denote the passage of pooling indices to forward layers. Adapted from [13].

them. U-nets [6] take advantage of skip connections to map higher semantic-level information to low semantic-level pixel information. These networks are Encoder-Decoder architectures wherein the downsampling half (Encoder) is symmetrical to the upsampling half (Decoder), as shown in Figure 3. There is also a larger amount of feature channels in the upsampling layers, which allows for more information to be propagated to higher resolution layers [6].

3) SegNets: SegNets [13], like U-nets, are Encoder-Decoder architectures for segmentation with symmetric layers. The Encoder half of the network is composed of VGG-like 3×3 convolutional layers. The construction of the Decoder network is accomplished by mirroring the Encoder layers and replacing the pooling layers for upsampling components, as can be seen in Figure 4. One main advantage of SegNet compared to other segmentation architectures is the use of the pooling indices in the upsampling process. SegNet uses the max pooling indices to upsample (without learning) the feature maps and deconvolves with a trainable decoder filter bank [13].

IV. Deep Transfer Learning

DNNs are powerful overcomplete statistical models that can learn to extract features from and infer over unstructured data such as images, sounds or texts. One great limitation for DNNs is the amount of data available for feeding these models, as generalizing patterns over unstructured can be an exceptionally hard task. Transfer Learning is an important tool in this context, as it allows for less data to be used in the training procedure of more specific domains. We used fine-tuning [14] as our baseline, as it is currently the most common method for Transfer Learning in the field of Computer Vision. It is easy to find several pre-trained neural network architectures in large datasets such as ImageNet[1]. CV applications can benefit from this strategy [14], as fine-tuning a pre-trained net only requires a small fraction of the data needed to train a DNN from scratch. This section introduces the proposed methodology for Cross-Dataset Transfer Learning.

A. Image-to-Image Translation

Isola *et al.* [15] proposed the first DNN architecture for domain-agnostic image-to-image translation. Before this work, image translation tasks were tackled with special-purpose methodologies [16]–[20], but the problem remains the same in all of these settings: mapping pixels to pixels. The main contribution of [15] was, therefore, to provide a general architecture – henceforth referred to as pix2pix[2] – and loss for this kind of task.

Before pix2pix, the literature had already discovered that naive approaches for image translation losses – such as using Euclidean Distance – tend to produce blurry results, as the network tries to minimize sample means [20], [21]. The solution to this problem was to introduce an adversarial loss to the pipeline by using a Generative Adversarial Network (GAN) [22] architecture. Adversarial losses tend to produce more photorealistic images than traditional losses, as the discriminator is able to identify blurry images and force the generator to produce images with sharper edges.

The generator network is normally an Encoder-Decoder network such as U-net [6], which receives the image in the source domain and translates it to the target domain. The discriminator network is a traditional architecture for image classification, such as a CNN [8], [10]. The discriminator has the job of determining if the image is a natural sample from the specific domain or if it is a translated sample originally from another domain.

Samples are fed to the network during the training phase in a supervised manner and, therefore, pix2pix requires paired images in the source and target domains. The need for paired samples represents a serious hampering for many real world applications of pix2pix, including biomedical ones. These difficulties in obtaining paired samples from different domains encouraged the creation of Unpaired Image-to-Image Translation models [23]–[25], further detailed in Section IV-B.

B. Unpaired Image-to-Image Translation

Unpaired Image-to-Image Translation [23]–[25] can be achieved by using the property of Cycle Consistency (Figure 5). When using paired networks as pix2pix, one can simply compare images from the source domain to images from the target domain, but this strategy does not work for unpaired samples, thus the need for cycle consistent losses.

Cycle Consistency is based on the premise that image translation can be modeled as follows. Let A and B be two image domains and $G_{AB} : A \rightarrow B$ and $G_{BA} : B \rightarrow A$ two

[1] www.image-net.org/

[2] https://phillipi.github.io/pix2pix/

(a) (b)

Fig. 5. Typical adversarial architecture for Unpaired Image-to-Image Translation based on Cycle Consistency. (a) Translation of a sample $A \rightarrow B \rightarrow A$. (b) Translation of a sample $B \rightarrow A \rightarrow B$.

translation functions between these domains. Given two samples $a \in A$ and $b \in B$, it is possible to derive an adversarial loss based on the following comparisons: $a \approx G_{BA}(G_{AB}(a))$ and $b \approx G_{AB}(G_{BA}(b))$. This objective function enforces that G_{AB} and G_{BA} are inverses of each other, that is, $G_{AB} = G_{BA}^{-1}$ and $G_{BA} = G_{AB}^{-1}$. Generative networks such as G_{AB} and G_{BA} are generally implemented as Encoder-Decoder architectures, similar to U-nets [6] and SegNets [13].

The counterparts of the generative networks in GANs are discriminative networks, which are trained to identify if an image is natural from the domain or a translated sample originally from another domain. D_A and D_B will be henceforth referred to as the discriminative networks for datasets A and B, respectively. Discriminators are normally traditional supervised networks, such as CNNs [8], [10], which are trained in the classification task of distinguishing real images from fake images generated by the generators.

Even though there are several architectural differences between the methods of Unpaired Image-to-Image Translation, the core of the idea of Cycle Consistency can be seen in Figure 5. Specific architectures of G_{AB}, G_{BA}, D_A and D_B, as well as customly designed losses can grant different translation methods special characteristics such as different encodings for style and content in an image [25].

C. Proposed Method

A recent survey on Cross-Dataset Transfer Learning [26] foresees the use of Image-to-Image Translation Networks [15], [23], [25] for Unlabelled Target Dataset Transfer. The Transfer Learning DNN architecture proposed in this paper builds up on this prediction and can be used in unsupervised and semi-supervised settings, that is, with few or no target labels.

With only simple modifications, one can adapt the Unpaired Image Translation architecture shown in Figure 5 in order to perform Cross-Dataset Transfer Learning. Let A be a labeled dataset and B be a weakly labeled or unlabeled dataset. We propose the architecture shown in Figure 6 for transfering knowledge from A to B. The unsupervised part (circled in green) is simply an unpaired translation network, such as [23]. The supervised section (in red) uses a model M_A pre-trained on A to enforce discriminative translations by G_{AB} and G_{BA} – that is, translations from B to A that preserve the visual features important for the class discrimination in M_A. As shown in Figure 6b, if there are any labels for the dataset B, they are also taken into account by the architecture, allowing for a better training of G_{BA}.

Discriminative and generative models in GANs are trained intermittently. At first, the generators are frozen while both discriminators are trained simultaneously using backpropagation. Later the inverse occurs: the discriminative networks are frozen and both generators are trained at the same time. Our method adds a third optimization procedure to this pipeline, wherein M_A is fine-tuned and backpropagates the training errors to G_{AB} and G_{BA}, while D_A and D_B are frozen. These training steps will be henceforth called generative, discriminative and supervised steps.

If convergence is met, it is possible to forward an image $b \in B$ to G_{BA}, get its counterpart in A and forward it to M_A, as G_{BA} was enforced to preserve the visual features important for M_A. If there are no labels for B samples, only the A labels are used in the supervised part of the network. Therefore, our architecture can use A labels to train a model for B samples in a completely unsupervised setting. Contrary to fine-tuning, our method uses the whole B dataset to transfer the knowledge, not only the labeled samples in B.

As the proposed Transfer Learning architecture is built on top of a generic Unpaired Image Translation architecture (Figure 5), it is agnostic to the choice of Cycle Consistency network. That is, one could easily shift between implementations of CycleGANs [23], UNIT [24] or MUNIT [25].

V. Experimental Setup

A. Chest X-Ray Datasets

The most used CXR datasets are the Japanese Society of Radiological Technology (JSRT [27][3]), the Montgomery/Shenzhen Sets [28][4] and the ChestX-ray8 [29][5].

JSRT contains 247 PA Chest Radiographs, while the Montgomery Set is composed of 138 cases. JSRT has pixel-level labels for lung field segmentation tasks as well as heart and clavicle ground truths, while the Montgomery Set only contains ground truths for the lungs. ChestX-ray8 and the Shenzhen Set do not provide pixel-level labels. Therefore our

[3]http://db.jsrt.or.jp/eng.php

[4]https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets/

[5]https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/37178474737

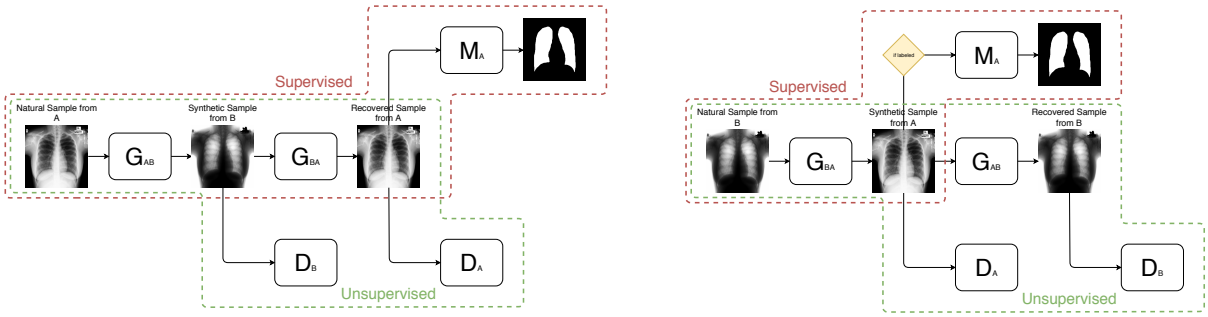

Fig. 6. Proposed architecture for unsupervised and semi-supervised Transfer Learning. Unsupervised modules correspond to a cycle consistency-based translation network. (a) Training procedure for a sample $A \rightarrow B \rightarrow A$. (b) Training procedure for a sample $B \rightarrow A \rightarrow B$. One should notice that most or all labels for $b \in B$ samples might be missing. This is the reason why there is a conditional in the $B \rightarrow A \rightarrow B$ diagram.

quantitative experimental procedure only took into account the JSRT and Montgomery Sets, as they are the only ones with pixel-level annotations. Qualitative assessments of heart and clavicle Transfer Learning results using models pre-trained in JSRT are presented for the Montgomery Set, which does not have ground truths for these tasks.

B. Segmentation Experimental Procedure

We resized all images and ground truths to 256×256 pixels in order to lower GPU memory requirements and due to exploratory tests that showed no benefits in using larger image sizes. We built all architectures in pytorch[6], using the pytorch-semantic-segmentation[7] implementation as a basis on an NVIDIA Titan X Pascal with 12GB of memory.

Results were obtained using a 5-fold cross-validation methodology over the datasets. For each test fold, one of the other 4 training folds was not used in training and served as a second validation step in order to select the epoch with the best results. All networks were trained for 150 epochs using 4 images per batch, the Adam [30] solver and Cross Entropy loss. Momentum was set to 0.9, while a Learning Rate of 1×10^{-4} was used and a weight decay of 5×10^{-4}.

C. Transfer Learning Experiments

The exploratory tests for the proposed Transfer Learning architecture were performed using MUNIT [25]. This architecture was chosen because it is designed to split the encoding of content and style information in the images. This allowed our Transfer Learning method to encode images from the unlabeled/semi-labeled dataset B (Montgomery) with the style of the labeled dataset A (JSRT), while still preserving the content – that is, the basic shape – of the original image $b \in B$.

In order to prevent the vanishing gradients problem in G_{BA} and G_{AB}, we chose for M_A a segmentation architecture with skip connections: a U-net [6]. This U-net was pre-trained on a training fold comprised of 60% of the samples in the JSRT dataset according to the parameters shown in Section V-B. The other 40% of the JSRT images were used as validation (20%) and test sets (20%) for the U-net. For convenience, we matched the fold used in the Transfer Learning experiments with one of the 5-folds used in the segmentation tests (Section V-B), aiming to use one of the pre-trained segmentation models in the transfer experiments. The Montgomery Set was also divided in a 60%-20%-20% configuration. The knowledge acquired by the pre-trained U-net was then transferred to the Montgomery Set using both fine-tuning and our Transfer Learning method.

Due to GPU memory constraints, we needed to use rather shallow models for the generators (G_{AB} and G_{BA}) and for the discriminators (D_A and D_B). Each generative network only contained a total of 4 layers (2 in the Encoder and 2 in the Decoder), while each discriminative network only had 2 convolutional layers and one FC layer.

We noticed that trying to transfer the knowledge in M_A since the first training epoch was detrimental to the convergence of the translation model, probably due to competing supervised and unsupervised objectives. Therefore, we first trained the generators and discriminators for 20 iterations and made sure they converged via visual assessment. Only then we started training the supervised part of the model coupled with the unsupervised translation method. This strategy also allowed us to train one single translation model in a completely unsupervised fashion for the first 20 epochs, only then starting the supervised training steps for different tasks using the same pre-trained G_{AB}, G_{BA}, D_A and D_B.

VI. Results and Discussion

This section presents the results from both supervised semantic segmentation (Section VI-A) and unsupervised/semi-supervised Transfer Learning (Section VI-B). Supervised results are shown according to Jaccard $\ddot{J}$ (Intersection over Union – IoU) and Dice $\ddot{D}$ (F1-score) metrics, while Transfer Learning only contains $\ddot{J}$ comparisons, as it is the most common metric in the literature.

A. Segmentation Results

Segmentation metrics for the JSRT dataset can be seen in Table I. This table compares the fully supervised semantic

[6]http://pytorch.org/

[7]https://github.com/ZijunDeng/pytorch-semantic-segmentation

TABLE I
JACCARD AND DICE RESULTS FOR THE JSRT DATASET ON THE LUNG FIELD, HEART AND CLAVICLE SEGMENTATION TASKS. BOLD RESULTS SHOW THE BEST VALUES FOR EACH METRIC IN EACH TASK.

Lung Field

Methods	$\ddot{J}$	$\ddot{D}$
FCN	95.05 ± 0.20	97.45 ± 0.11
U-Net	**96.02 ± 0.33**	**97.96 ± 0.17**
SegNet	95.54 ± 0.32	97.71 ± 0.17
InvertedNet [5]	95.10	97.50
SCAN [4]	94.70 ± 0.40	97.30 ± 0.20
Atlas [2]	95.40 ± 1.50	96.70 ± 0.80
Hybrid Voting [1]	94.90 ± 2.00	-
Human Observer [1]	94.60 ± 1.80	-

Heart

Methods	$\ddot{J}$	$\ddot{D}$
FCN	89.25 ± 0.56	94.24 ± 0.35
U-Net	89.21 ± 1.32	94.16 ± 0.88
SegNet	**89.64 ± 0.91**	**94.44 ± 0.55**
InvertedNet [5]	87.10	93.10
SCAN [4]	86.60 ± 1.20	92.70 ± 0.20
Hybrid Voting [1]	86.00 ± 0.56	-
Human Observer [1]	87.80 ± 5.40	-

Clavicles

Methods	$\ddot{J}$	$\ddot{D}$
FCN	75.52 ± 1.03	85.90 ± 0.68
U-Net	86.54 ± 0.99	92.58 ± 0.80
SegNet	87.30 ± 0.67	**93.08 ± 0.49**
InvertedNet [5]	87.00	93.00
ASM [1]	73.40 ± 13.70	-
Hybrid Voting [1]	73.60 ± 10.60	-
Human Observer [1]	**89.60 ± 3.70**	-

segmentation results obtained by our networks with the results of Atlas-based methods [2], InverseNets [5], GANs [4] and the techniques used by [1].

One can see that our DNNs achieved better results than all other methods in all tasks but in clavicle segmentation, where humans observers obtained better results. This is likely due to the fact that clavicle segmentation is an extremely unbalanced task, as these bones are quite small when compared to CXR image dimensions. SegNets and U-nets performed better than FCNs in all tasks, mainly in clavicle segmentation, where FCNs obtained much lower results ($\ddot{J} = 75.52 \pm 1.03$ and $\ddot{D} = 85.90 \pm 0.68\%$), while other DNNs achieved Jaccard values close greater than to 86.5%.

In heart and lung field segmentation, DNNs surpassed the results of human observers, which is an indication that the inter-observer variability is larger than the variability between automatic segmentations. In other words, the DNNs are able to match the labeling characteristics of the golden standard observer – which was used as ground truth for the dataset – better than other human observers. This conclusions is aligned with what [1] reported in their conclusions.

DNNs obtained considerably lower standard deviations than shallow methods. This is evidence that deep methods tend to produce lower magnitude errors and more reliable predictions.

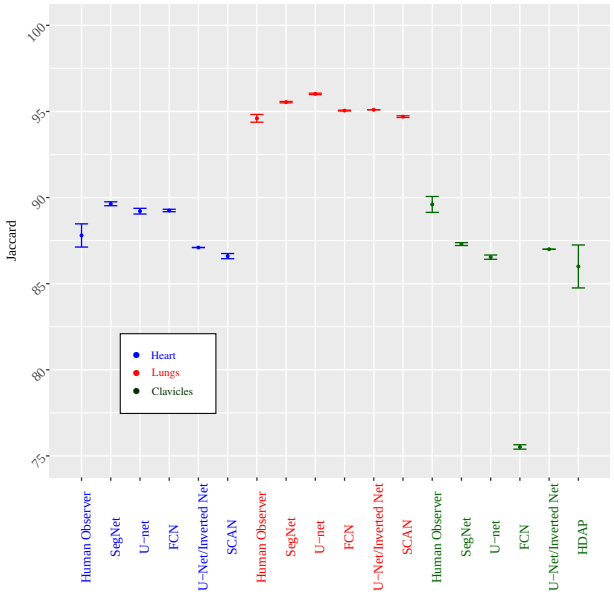

Fig. 7. Confidence Intervals (CIs) for the JSRT dataset [27] in lung field, heart and clavicle segmentation. Vertical axis represent Jaccard metrics for $p \leq 0.05$. The lower end of the plot was trimmed at 75% to improve visualization.

TABLE II
TRANSFER LEARNING RESULTS FOR THE MONTGOMERY SET [28] IN AN U-NET PRE-TRAINED IN THE JSRT DATASET [27]. BOLD VALUES INDICATE THE BEST RESULTS FOR EACH LINE.

Label %	Our Method	Fine-Tuning	From Scratch
0%	**88.20 ± 9.80**	4.30 ± 4.13	–
1.25%	**88.83 ± 9.81**	78.94 ± 13.32	54.23 ± 13.37
2.5%	**88.25 ± 10.19**	83.32 ± 12.32	56.01 ± 13.76
5%	**90.79 ± 7.05**	83.46 ± 8.60	55.10 ± 14.42
10%	**89.18 ± 9.18**	83.66 ± 9.69	87.80 ± 6.78
20%	**91.26 ± 7.20**	88.71 ± 8.73	89.50 ± 7.65
50%	92.15 ± 5.90	**93.78 ± 5.42**	89.82 ± 4.34
100%	93.18 ± 5.47	**94.81 ± 5.15**	94.16 ± 4.57

Confidence Intervals (CIs) for $p \leq 0.05$ regarding the results presented in Table I can be seen in Figure 7.

B. Transfer Learning Results

Table II shows the results obtained by the proposed Transfer Learning method compared with normal fine-tuning and training the networks from scratch with the limited labels. Figure 8 shows the CIs using $p \leq 0.05$ for the results in Table II. The horizontal axis represents the amount of labels kept by the experiment, while the vertical axis denote Jaccard values achieved in these settings. It is clear that our Transfer Learning method significantly surpasses the effectiveness of fine-tuning when using between 0% and 20% of the labels from the target training set. When using 50% and 100% of the target labels, fine-tuning marginally surpassed our method, even though the difference was not statistically significant.

When using no labeled data in the target dataset, it can be seen that $\ddot{J}$ drops to only 4.30%. This result renders it infeasible to interchange models between CXR datasets

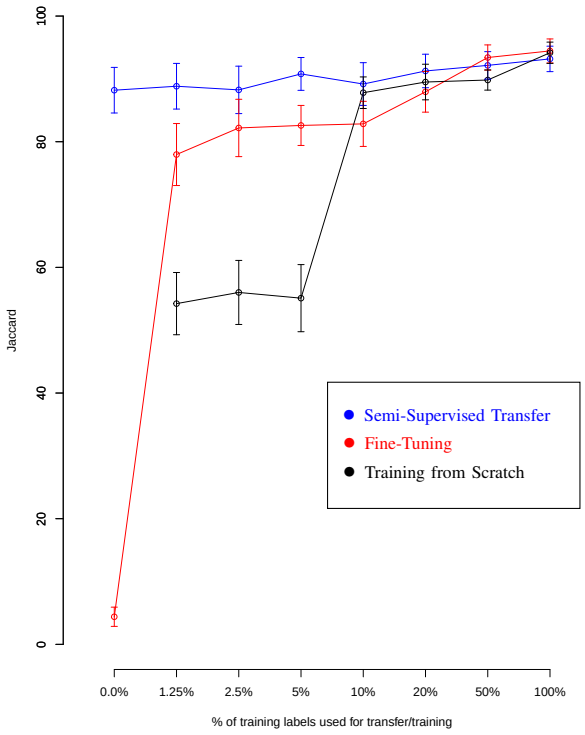

Fig. 8. CIs for the Montgomery Set [28] in lung field segmentation using a model pre-trained in the JSRT dataset [27].

without labeled data in the target dataset using traditional transfer methods. Our method achieves a Jaccard of 88.20% even without labeled data in the target set, as it uses all the unlabeled target samples to perform the transfer and the source labels to ensure visual feature preservation by G_{AB} and G_{BA}.

1) Qualitative Assessment of Other Tasks: As the Montgomery Set [27] does not have pixel-level labels for clavicle and heart segmentation, we performed qualitative tests on our Transfer Learning method for these tasks using the unsupervised case, that is, without labeled data in the target dataset. One can see in Figure 9 that clavicles and heart regions were accurately recognized. In most Montgomery images the algorithm correctly identified the clavicles, with only 4 cases of inadequate segmentations in one or both clavicles among the 27 images tested for this task. One example of misidentification of the clavicle area is shown in Figure 9d. Most heart segmentations were near perfect, but, as the Montgomery Set contains more diverse samples, hearts with abnormal shapes were not fully identified, as can be see in Figure 9h.

VII. Conclusion

CXR segmentation results showed that DNNs yielded better metrics than all baselines, both shallow and deep. Our use of classic semantic segmentation architectures [6], [9], [13] for CXR tasks obtained better results than customly made DNNs [4], [5]. This is probably due to hyperparameter optimization,

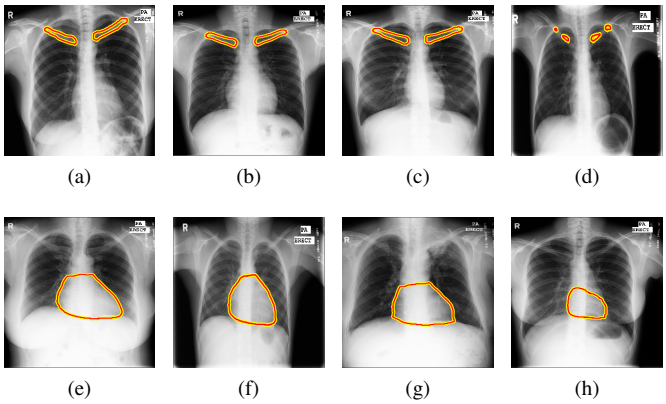

Fig. 9. Segmentation results for the Montgomery Set [28] in the tasks of (a-d) clavicle and (e-h) heart segmentation from a model pre-trained in JSRT [27] and transferred with 0% of labeled data in the target dataset.

as [4] reported much worse segmentations with FCNs than the ones achieved in our experimental results.

Transfer Learning experiments showed the superiority of our semi-supervised methodology over both fine-tuning and training from scratch when dealing with few labels on the target dataset. These results could be explained by the fact that out method uses both the labeled and unlabeled data of the target domain, while fine-tuning and training from scratch only consider labeled images. This hypothesis will be investigated in future works. The proposed method produced good Jaccard results close to 90% even when no target labels were used. Qualitative assessments also proved that the transfer was successful in segmenting heart and clavicle regions even with 0% of labeled data on the target dataset. The Montgomery Set is a much more challenging dataset than JSRT, as the former contains much more lung abnormalities due to diseases that alter the shape and size of the lungs [4]. The Montgomery Set also contains a broader range of imaging quality, further assessing the generalization capabilities of our method.

Disadvantages of the proposed method when compared with fine-tuning are: extended training time required for the translation and tuning procedure of M_A and larger GPU memory requirements. Stability was also a major concern during the training procedure, as sometimes the translation network did not converge properly, hampering the transfer procedure. In these early tests we simply repeated the training procedure when visual comparisons and objective metrics showed that G_{AB} and G_{BA} did not properly achieved convergence. Therefore, the lack of stability in our training procedure is a pressing issue that will be addressed on future iterations of this work. At last, as our model is agnostic to the unpaired translation GAN architecture, future advances in this front should be compatible and further benefit our approach.

As explained in Section V-C, the discriminator and generator networks used in the transfer procedure contained only 2 and 4 layers, respectively. As shallower models allow for simpler semantic representations, using deeper models should further improve the results presented in Section VI-B.

Therefore, future works include testing the proposed method with deeper models for both generative and discriminative networks. Future experiments will also feature other semantic segmentation architectures with skip connections, such as SegNets [13], ResNets [11] and DenseNets [12].

It was mentioned in Section V-C that we chose MUNIT [25] as a basis for the Transfer Learning tests because it has the ability to split content information from style information in an image. One should also expect the model M_A (see Figure 6) to act as a regularizer for enforcing content preservation between translated domains. This effect will be further evaluated in future tests with the use of CycleGANs [23] and UNIT [24], which do not possess the ability to explicitly separate style from content. MUNIT is also not ideal for Cross-Dataset Transfer Learning tasks, as it uses only one image $a \in A$ for extracting the dataset's style. If the chosen image is an outlier or if it simply does not represent correctly the general style of A samples, the transfering procedure can be compromised. Authors intent to solve this problem by extracting one single style for each dataset by adding a style regularization loss component between samples from the same dataset.

No aspect of our method ties it only to CXR images, as all network components are general purpose semantic segmentation and image translation DNNs, therefore, tests in different biomedical domains are planned. Cross-Domain Transfer Learning in different but similar domains (i.e. Magnetic Resonance and Computerized Tomography) are also planned for future iterations of this work.

ACKNOWLEDGMENT

Authors would like to thank NVIDIA for the donation of the GPUs that allowed the execution of all experiments in this paper. We also thank CAPES, CNPq, and FAPEMIG (APQ-00449-17) for the financial support provided for this research project.

REFERENCES

[1] B. Van Ginneken, M. B. 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.

[2] 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.

[3] L. Hogeweg, C. I. Sánchez, P. A. de Jong, P. Maduskar, and B. van Ginneken, "Clavicle segmentation in chest radiographs," *Medical image analysis*, vol. 16, no. 8, pp. 1490–1502, 2012.

[4] W. Dai, N. Dong, Z. Wang, X. Liang, H. Zhang, and E. P. Xing, "Scan: Structure correcting adversarial network for organ segmentation in chest x-rays," 2018.

[5] A. A. Novikov, D. Lenis, D. Major, J. Hladůvka, M. Wimmer, and K. Bühler, "Fully convolutional architectures for multi-class segmentation in chest radiographs," *IEEE Transactions on Medical Imaging*, 2018.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[7] H.-Y. Chen and J.-T. Chien, "Deep semi-supervised learning for domain adaptation," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, June 2015, pp. 3431–3440.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[12] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *CVPR*, vol. 1, no. 2, 2017, p. 3.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *TPAMI*, vol. 39, no. 12, pp. 2481–2495, 2017.

[14] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1717–1724.

[15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[16] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 60–65.

[17] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 5, p. 124, 2009.

[18] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001, pp. 341–346.

[19] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[20] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European Conference on Computer Vision*. Springer, 2016, pp. 649–666.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[23] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.

[24] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 700–708.

[25] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *arXiv preprint arXiv:1804.04732*, 2018.

[26] J. Zhang, W. Li, and P. Ogunbona, "Transfer learning for cross-dataset recognition: A survey," 2017.

[27] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

[28] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[29] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3462–3471.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.