Pattern Recognition Letters 140 (2020) 10–17

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

From 3D to 2D: Transferring knowledge for rib segmentation in chest X-rays☆

Hugo Oliveira[a,*], Virginia Mota[a], Alexei M.C. Machado[b,c], Jefersson A. dos Santos[a]

[a] Computer Science Department, Universidade Federal de Minas Gerais, 31270-901, Antônio Carlos Av., Belo Horizonte, 6627, Brazil
[b] Department of Anatomy and Imaging, School of Medicine, Universidade Federal de Minas Gerais, 30130-100, Alfredo Balena Av., Belo Horizonte, 192, Brazil
[c] Computer Science Department, Pontifícia Universidade Católica de Minas Gerais, 30535-901, Dom José Gaspar Av., Belo Horizonte, 500, Brazil

ARTICLE INFO

Article history:
Received 21 September 2019
Revised 7 August 2020
Accepted 21 September 2020
Available online 22 September 2020

Keywords:
Chest X-rays
Deep domain adaptation
Rib segmentation
Generative adversarial networks

ABSTRACT

Chest X-rays are the most common type of biomedical radiologic exam, being widely adopted for the diagnosis of a myriad of illnesses in the thoracic region. Computed Tomography – even though being more expensive and rare – is also a useful tool for the detection of several illnesses and surgery planning, providing volumetric information. This paper proposes a methodology aiming to leverage the larger amounts of spatial information and lack of occlusion in tomographic images to aid in the rib segmentation of 2D X-ray images by means of Domain Adaptation. We perform extensive quantitative and qualitative experiments to test the capabilities of this methodology in segmenting ribs in 7 X-ray datasets with distinct visual features, using 6 different metrics and without any use of rib segmentation labels from the target image sets. In order to encourage reproducibility, all data and code used in this research is publicly available online, including a new 2D Digitally Reconstructed Radiograph generated from tomographic data and a new pixel-level label map for the JSRT Chest X-ray dataset. We also publicize our generalizable pretrained models for both rib segmentation in Chest X-rays and lung field segmentation in Digitally Reconstructed Radiographs. Results show that the proposed pipeline outperforms shallow rib segmentation baselines in almost all quantitative metrics and produce higher fidelity pixel-map predictions than simply using the pretrained Neural Networks on the flattened 3D data, mainly in datasets where domain shift is more pronounced. The use of Conditional Domain Adaptation also allows the method to perform inference on all 7 X-ray datasets using one single model, achieving over 0.856 of AUC on OpenIST and 0.934 of AUC on JSRT, with Dice scores of 0.68 and 0.69 in these two datasets, respectively.

1. Introduction

Deep Neural Networks (DNNs) have achieved state-of-the-art results in both computer vision and biomedical image analysis [1]. DNNs are powerful overcomplete statistical models that can learn to extract features from unstructured data such as image, audio, video and text. These models can be understood as ensembles of perceptrons organized in layers with increasing semantic capabilities, being able to extract and select features and to perform inference conjointly. In general, deeper models are able to encode information with higher semantics, while shallower models can only optimize low-level semantic information. Convolutional Neural Networks (CNNs) [2] and their variants [3,4] are the most popular architectures used for performing inference (i.e. classification, segmentation and detection tasks) over images, including biomedical settings.

Despite recent efforts [5–7] in acquiring large labeled datasets – mainly for diagnosis, that is, classification tasks – most biomedical image domains suffer from a lack of labeled data. Therefore, it is highly desirable to acquire the most amount of knowledge possible with the few labeled data available in the literature, while also using the vast amounts of unlabeled data present in some domains. Aiming to lessen the requirements for labeled data in visual recognition, Domain Adaptation (DA) [8–11] is the research area that comprises the theoretical background and methods for knowledge transfer between distinct tasks and/or data.

Chest X-rays (CXRs) are the most common type of radiological exam acquired nowadays, mainly due to their ability to aid in the detection and diagnosis of several kinds of ailments [6]. Several diverse health conditions such as pulmonary nodules [12], tuberculosis [13], pulmonary effusion, pneumonia and cardiomegaly [6], as well as bone fractures, can be assessed by CXRs in a quick and

☆ Handle by Associate Editor-in-Chief Gabriella Sanniti di Baja, PhD.

* Corresponding author.
E-mail address: oliveirahugo@dcc.ufmg.br (H. Oliveira).

https://doi.org/10.1016/j.patrec.2020.09.021
0167-8655/

radiation-efficient exam. With the advent of large labeled datasets [6,7], automation efforts have been proposed for aiding in the diagnosis of most of these illnesses. Computed Tomography (CT) exams yield volumetric images that allow physicians to perform 3D analysis of the thorax, but require more expensive hardware and submit the patient to between one and two orders of magnitude larger doses of radiation.

As argued by Zhang et al. [14] the understanding of anatomical objects in CXRs is useful for several clinical applications, such as pathological diagnosis, treatment evaluation, surgical planning and as an automated preprocessing step for Computer-Aided Detection (CAD) systems. Van Ginneken et al. [15,16] point that the delineation of rib borders is part of this anatomical registration process whose automation can help physicians by providing a frame of reference for the location of abnormalities [14,16] and help surgery planning. However, computerized analysis is still the largest beneficiary from algorithms for rib cage detection, being useful to mitigate both false positives [17] and false negatives [18] in nodule or rib fracture detection, which may also indicate other issues as damage to lung tissue, hemorrhage signs of osteoporosis or even an underlying cancer [19]. For instance, Austin et al. [18] report that between 82% and 95% of undetected lung cancers in CXRs were obscured by foreground bones such as ribs or clavicles. Yet, mainly due to the great burden involved in pixel-level annotations, there is only a tiny amount of publicly available labeled samples for the task of rib segmentation in CXRs, as further discussed in Section 3.3.

Given the problems related to acquiring pixel-level labels for rib segmentation and the usefulness of these data, the main contribution of this paper is a pipeline based on Conditional Domain Adaptation [20] for rib cage segmentation. Secondary contributions of this work include validating the use of volumetric data for CXR bone segmentation, presenting a new rib segmentation label set for the JSRT dataset [12], and defining a standard quantitative and qualitative comparison procedure for rib cage segmentation.

The remaining sections of this paper are organized as follows. Section 2 describes the current state of the literature on rib segmentation and suppression methods. Section 3 presents the proposed method for rib segmentation from CT data, while also discussing the experimental setup (i.e. datasets, metrics, etc) used in the tests. Section 4 shows the quantitative and qualitative results yielded by the experimental setup, comparing them to the baselines of rib segmentation and with pretrained DNNs. At last, Section 5 concludes the paper with our final remarks and future works.

2. Related work

This section presents the current literature on rib segmentation and suppression methods. It also describes the basis of Unsupervised Image-to-Image Translation, which is the theoretical basis for the proposed method.

2.1. Rib segmentation

The problem of rib segmentation has been classically tackled with geometric models globally fitted in the radiography, which is often filtered in order to highlight the ribs from lung background pixels [15,19]. Curve and geometry fitting models, while having the advantage of being compact, computationally fast and usually unsupervised, are often too simple to comprise the highly complex spectrum of rib shapes [21]. Thus, the literature in the past decades has favored statistical and/or learning-based models [21,22] able to generalize to more diverse data. Pixel Classification (PC) based on local gray-level variation [22], texture [22,23] and shape [24] information has become, therefore, a common methodology in the field.

Instead of just using training and testing images from the same domain, it is possible to leverage the larger amount of information available in 3D images (i.e. CT-scans or Magnetic Resonance Imaging – MRI) to perform useful tasks on 2D data, such as CXRs [14,21]. As all the information in one whole column of the Posterior-Anterior (PA) axis is compiled into one single pixel in PA CXRs, there is a lot of occlusion and tissue overlay in these images, thus, making some tasks as rib segmentation considerably harder than in 3D imaging. Data generated by flattening the PA axis in CT images to acquire 2D shapes resembling CXRs are known in this context as Digitally Reconstructed Radiographs (DRRs) [14]. Depending on the flattening operation used (i.e. average, maximum, linear combination of percentiles, etc), one might yield distinct resulting images which can, for instance, resemble a CXR or serve as bone-segmentation masks.

To the best of our knowledge, the present work is the first one that fully leverages this larger amount of information in 3D space to acquire 2D labels for performing domain adaptation to CXRs. The proposed method differs significantly from Zhang et al. [14] since it does not use bone labels from CT scans for supervised learning, but rather generates the labels from the 3D data itself. Only public datasets and well-known metrics are used for validation. The experiments are focused on the generalization capabilities of the pipeline, presenting extensive results for several datasets (see Section 3.3) that are achieved by one single shared model.

2.2. Unsupervised image-to-image translation

Unsupervised Image-to-Image (I2I) translation networks are based on the concept of Cycle-Consistency, which models the translation process between two image domain as an invertible process represented by a cycle. This cyclic structure allows for Cycle-Consistent losses to be used together with the adversarial loss components of traditional GANs.

A Cycle-Consistent loss can be formulated as follows: let A and B be two image domains containing unpaired image sample sets X_A and X_B. Consider then two functions $G_{A\to B}$ and $G_{B\to A}$ that perform the translations $A \to B$ and $B \to A$ respectively. Then a loss $\mathcal{L}_{cyc}$ can be devised by comparing the pairs of images $\{X_A^{(i)}, X_{A\to B\to A}^{(i)}\}$ and $\{X_B^{(i)}, X_{B\to A\to B}^{(i)}\}$. In other words, the relations $X_A^{(i)} \approx G_{B\to A}(G_{A\to B}(X_A^{(i)}))$ and $X_B^{(i)} \approx G_{A\to B}(G_{B\to A}(X_B^{(i)}))$ should be maintained in the translation process. 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 translated samples originally from other domains. D_A and D_B are referred to as the discriminative networks for datasets A and B, respectively. Discriminative networks are normally traditional supervised networks, such as CNNs [2], which are trained in the classification task of distinguishing real images from fake images generated by the generators. Unsupervised I2I translation have already proven to be useful in the generation of synthetic samples for DA in biomedical images [20,25–31].

3. Methodology

In this section, we present the proposed methodology for rib cage segmentation in CXRs by using Unsupervised Domain Adaptation (UDA) from DRRs. Our method leverages the capabilities of Conditional DA [20] to transfer the knowledge learned from synthetically flattened CT-scans to 2D CXRs. Sections 3.4 and 3.3 present a standardized set of metrics for rib segmentation evaluation and the datasets used in this research, respectively.

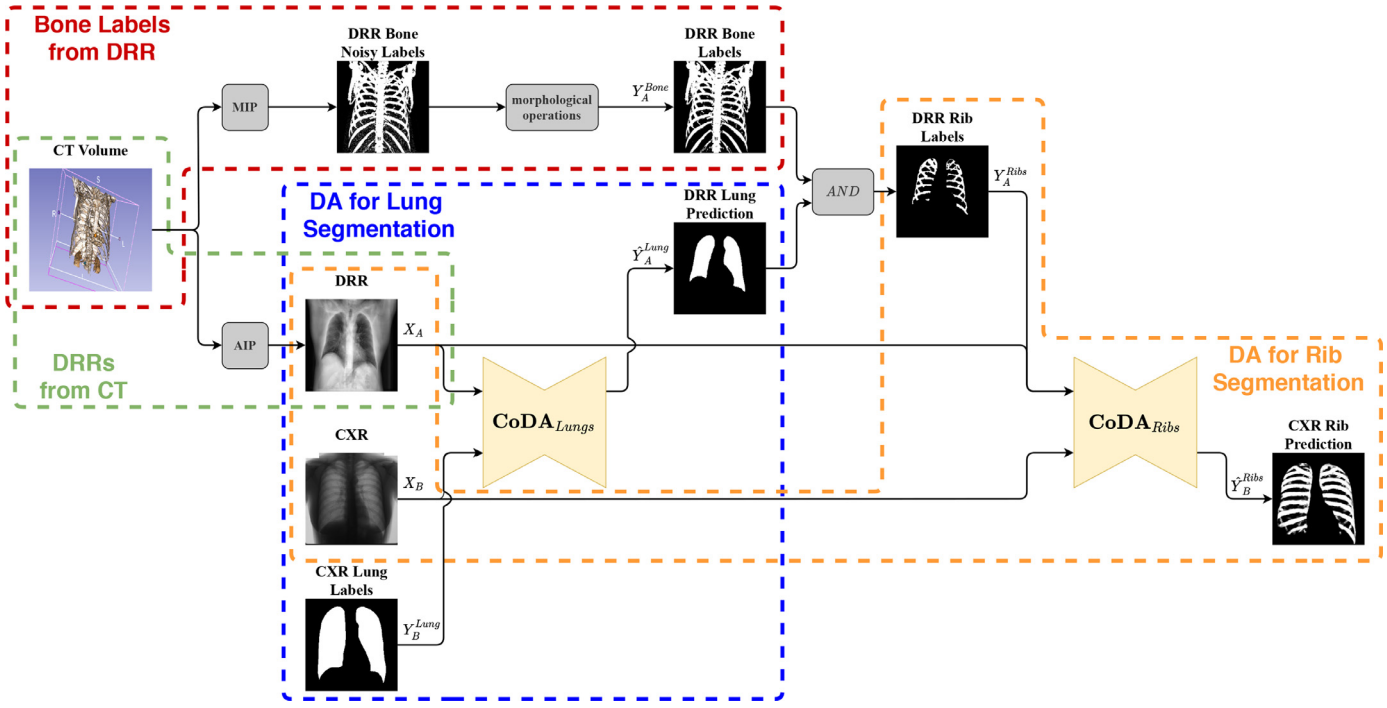

Fig. 1. Proposed rib segmentation pipeline. There are four submodules highlighted in the image: (1) the MIP procedure for acquiring bone labels from CT-scans in red; (2) AIP flattening in the PA axis to produce highlighted in green; (3) Conditional DA for DRR lung segmentation from CXR labels delineated in blue; and (4) CoDAGAN for segmenting CXR ribs in orange.

3.1. Rib segmentation pipeline

Conditional Domain Adaptation Generative Adversarial Networks (CoDAGANs) are Deep Learning architectures that can perform Unsupervised, Semi-Supervised or Fully Supervised Domain Adaptation. One can find the official CoDAGAN implementation publicly available in their project's webpage.[1] These DNNs are built upon the mathematical framework of unsupervised I2I translation [32] and Conditional GANs (CGANs) [33] for leveraging both the labeled and unlabeled data from several datasets in one single model. CoDAGANs have been shown to work on biomedical image segmentation tasks, being adaptable to model multimodal data from several distinct datasets [20] without scalability issues. Therefore, the methodology proposed in this paper is based on Conditional DA for its adaptability in modeling complex relationships among multimodal data distributions. This feature allows for the experimental procedures to be carried out from one single DRR source of labels to multiple CXR target datasets, as presented in Section 3.3.

In order to extract unsupervised knowledge useful for 2D images such as CXRs from volumetric CT-scans, the proposed pipeline for rib segmentation begins with two operations for flattening 3D volumes into 2D planes. These flattening operations can take the form of Average Intensity Projection (AIP) and Maximum Intensity Projection (MIP), being further described by Candemir et al. [21]. Applying AIP and MIP on CT volumes results in the images X_A and Y_A^{Bone} – that is, an averaged 2D projection of the 3D data and a noisy mask for the bones, respectively. These flattening procedures were previously referred in the literature as a way to generate useful 2D representations that can be used to extract knowledge for CXRs, especially for delineating anatomical structures such as bones and organs [14,21]. The proposed pipeline that contains AIP and MIP operations can be seen in Fig. 1.

Bone masks generated by the MIP operation on the CT-scans (Y_A^{Bone}) yield an acceptable yet noisy segmentation of the bones in the resulting DRR. Simple morphological filtering was observed to fix the noise introduced by the MIP operation in the labels. The computation of bone labels from the MIP operations in CT volumes can be seen delineated in red in Fig. 1. As bones and other natural/artificial structures prominently appear in DRRs when flattening is done using the MIP operation, undesirable objects as scapula and humerus bones, other anatomical features, and even pacemakers are often present as false positives in the label maps acquired for the DRRs. These artifacts are often located outside of the lung field area in the 2D projection of the DRRs, implying that an efficient lung segmentation could remove most of them from the training label set. Thus, in order to filter all these undesirable artifacts from the labels, we first used CoDAGANs to perform UDA for lung field segmentation from labeled and unlabeled CXR datasets to DRRs, as highlighted in blue in Fig. 1. These networks yielded semantic prediction maps for the lung pixels and allowed us to filter the noisy labels acquired from the noisy MIP operation on the PA axis of CT-scans. The resulting masks Y_A^{Ribs} are, therefore, computed according to:

$$Y_A^{Ribs} = Y_A^{Bone} \;\&\; \hat{Y}_A^{Lung}, \tag{1}$$

where the operator & represents the pixel-wise AND operation.

One should notice that in this pipeline we make two distinct uses of Conditional DA [20]:

1. **CoDA**$_{Lungs}$ for acquiring lung segmentation predictions $\hat{Y}_A^{Lungs}$ for DRRs from labeled CXR source datasets (X_B and Y_B) – highlighted in blue in Fig. 1;
2. **CoDA**$_{Ribs}$ for translating the knowledge from the filtered rib segmentation masks Y_A^{Ribs} for DRRs in order to use them in CXR data (X_B), resulting in the prediction $\hat{Y}_B^{Ribs}$ – as delineated in orange in Fig. 1.

[1] http://patreo.dcc.ufmg.br/codagans.

Table 1
Experimental setup for the 8 datasets used in our experiments. The sole DRR dataset (LIDC) was used only as a source dataset, while the other CXR data were used as targets. Due to the UDA nature of our methodology, no labels for CXRs were used during training. In order to perform quantitative data on the very few labeled samples in the CXR data, 12 samples from JSRT and 15 samples from OpenIST were used as testing sets, while experiments in other CXR datasets cannot yield objective evaluations due to lack of labels.

Division\dataset	LIDC	JSRT	OpenIST	Montgomery	Shenzhen	NIH	NLMCXR	PadChest
# **unlabeled training**	0	228	260	110	529	3390	2204	1590
# **labeled training**	835	0	0	0	0	0	0	0
# **unlabeled test**	0	0	0	28	133	270	854	381
# **labeled test**	0	12	15	0	0	0	0	0

In the experiments these two networks were trained separately, as they had distinct objectives and the experimental procedure was conceived to be relatively lightweight in GPU memory usage.

3.2. Architecture and hyperparameters

In order to implement $CoDA_{Lungs}$ and $CoDA_{Ribs}$, we adapted the original source code for CoDAGANs from Oliveira et al. [20]. This adversarial architecture, instead of being composed of a pair of generators ($G_{A \rightarrow B}$ and $G_{B \rightarrow A}$) and discriminators (D_A and D_B) – as most other unsupervised I2I translation DNNs [32,34,35] – has only one universal G and one universal discriminator D. G is composed of an Encoder ($G_{\mathbb{E}}$) with two downsampling and two residual layers, while the Decoder ($G_{\mathbb{D}}$) has two residual and two upsampling layers. This architecture has been shown to be effective for both pairwise I2I translation [34,35] and UDA for dense labeling in CXRs [20,25,31]. All experiments in this work were performed with the MUNIT [35] backbone for CoDAGANs, as it was shown to be more stable during training [20].

The state-of-the-art Adam [36] optimizer was used with a learning rate of 10^{-4}. More details regarding the training parameters of the CoDAGANs in this work can be found both in the original implementation of that method[2] and in this project's webpage.[3] Code for both baselines and the proposed pipeline were implemented with the PyTorch Deep Learning framework[4] using one single NVIDIA Titan X Pascal with 12 GB of internal memory.

3.3. Datasets

For testing the generalization capabilities of the methodology in data acquired using distinct digitization conditions, the pipeline was tested on 1 DRR source dataset (LIDC [37]) and 7 PA CXR target datasets: JSRT [12], OpenIST, Montgomery and Shenzhen sets [13], NIH Chest X-ray 8 [6], PadChest [7] and NLMCXR [5]. Labeled and unlabeled data distributions of these datasets are shown in Table 1. Links for the download pages of all datasets are shown in this project's website.

One should notice that the flattened 2D DRR dataset was not precomputed by the LIDC. Instead, the data and label DRR sets were generated from the CT-scans, resulting on a new dataset which is being made publicly available in the project's website for future research.

Another important consideration is that only one small set of labels for rib segmentation in real CXRs[5] could be found for the OpenIST dataset. In order to allow for a more complete quantitative evaluation, rough labels were generated for the JSRT dataset according to the Bone Shadow Suppression version of JSRT (the BSE-JSRT[6] [38]) using a simple difference operation between the bone suppression and the original datasets. Twelve of these images were curated to produce pixel-level label maps for the ribs by using manual thresholding, filling in the gaps between missing rib sections and removing false positive pixels. These 12 new ground truths for JSRT are also available for download at the project's website, as well as the pipeline's predictions for all CXR dataset test samples.

3.4. Evaluation metrics

There is no sole evaluation metric adopted by the literature of rib segmentation, therefore the quantitative results are presented using a wide variety of metrics aiming to standardize future experimental procedures in the field. The Receiver Operating Characteristic (ROC) curves are a common evaluation metric for detection/segmentation tasks according to probabilistic models, allowing for threshold-independent evaluation of model predictions. The Area Under Curve (AUC) can be computed for each ROC, yielding a quantitative value for comparing ROC curves generated by distinct algorithms. AUC is the main threshold-independent metric used in the literature of rib segmentation [15,21].

By selecting a threshold probability one can also compute the Accuracy, Sensitivity and Specificity of a segmentation model. Accuracy, Sensitivity and Specificity are, therefore, the main set of metrics used by the rib segmentation literature [15]. The Dice and Jaccard metrics were also used, as they are common metrics in the general field of radiological image segmentation [20,25,39].

4. Results and discussion

Before discussing the rib cage segmentation results, we present a subset of segmentation predictions by $CoDA_{Lungs}$ in DRRs from LIDC-IDRI (Fig. 2), also showing the poor lung segmentation generalization achieved by the baselines. These intermediate results are important because inaccurate segmentation predictions on the DRRs should result in poor bone mask filtering for separating the ribs from other bones in the MIP outputs. Most segmentations from $CoDA_{Lungs}$ are near perfect in identifying lung pixels, while Pre-trained U-Nets cannot properly compensate for the domain shifts between source (CXR) and target (DRR) data, completely missing the lungs in the general case. Domain-to-Domain (D2D) approaches for dense visual DA [25,31] show better domain shift compensation capabilities, but still fail to correctly delineate the whole lung area in most samples. The D2D approach used as a baseline is derived on the architecture proposed by Oliveira and dos Santos [25] and is virtually equal to the method described by Yang et al. [31]. Readers can appreciate the full predictions for lungs in DRRs in the project's webpage.

In order to assess the capabilities of the proposed method in rib segmentation, a systematic set of experiments was carried out in order to answer the following questions: (1) Is the proposed methodology effective for learning knowledge applicable to CXR data from synthetic DRRs? (2) How does Conditional DA perform in comparison with both deep and shallow baseline methods? (3)

[2] https://github.com/hugo-oliveira/CoDAGANs.
[3] https://sites.google.com/view/virginiafernandes/datasets/lidc-idri-drr.
[4] https://pytorch.org/.
[5] https://www.kaggle.com/viktorivanovio/testerv11.
[6] https://www.kaggle.com/hmchuong/xray-bone-shadow-supression.

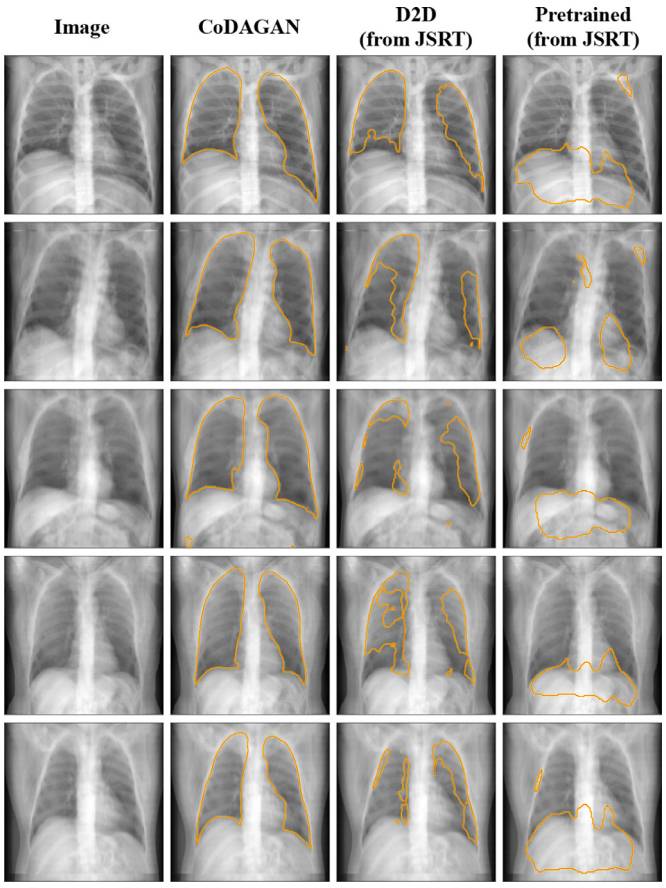

Fig. 2. Lung field predictions from $CoDA_{Lungs}$ and two deep baselines (D2D [25,31] and Pre-trained U-Nets [3]) in DRRs computed from LIDC-IDRI.

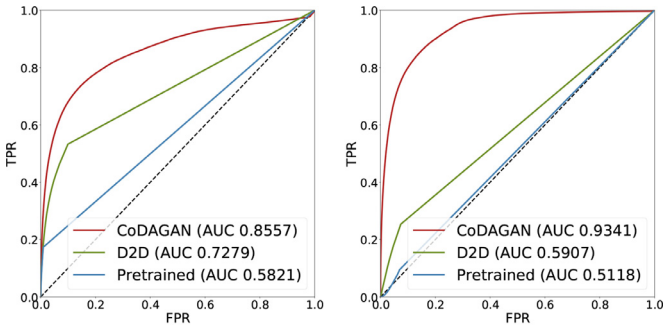

Fig. 3. ROC curves for CoDAGANs, D2D and Pre-trained U-Net in the OpenIST (a) and JSRT (b) image sets.

Is it possible to train a single CoDAGAN that is able to generalize to different data distributions acquired with diverse settings and digitization equipment? Questions (1) and (2) are addressed in Section 4.1, which presents quantitative results using the evaluation metrics described in Section 3.4 for the JSRT and OpenIST datasets. We compare Conditional DA [20] with both common shallow baselines in the area of rib segmentation [15,21,22] and modern Deep Learning approaches for UDA, such as reusing Pre-trained U-Nets [3] from other datasets and D2D. Section 4.2 discusses qualitative results obtained in both labeled and unlabeled test datasets and addresses question (3).

4.1. Quantitative analysis of rib cage segmentation

Fig. 3 presents the overlayed ROC curves for OpenIST and JSRT of both Pre-trained DNNs and Conditional DA in the task of rib segmentation. One can already see that in Fig. 3a CoDAGAN achieves clearly superior performance than both D2D and Pre-trained U-Nets. However, CoDAGANs excels even more in comparison to the baselines in JSRT (Fig. 3b), with Pre-trained U-Nets achieving almost random chance and D2D not so differently. These variations in the performance of D2D and CoDAGANs can be better understood by analysing the datasets themselves. The DRRs generated by AIP present inconsistent positions and rotations in the patient's orientation, while also showing large variations in radiogram contrast that more closely resembles the OpenIST dataset. In contrast, JSRT was obtained in a much more controlled setting with patients being imaged in a standard position and equipment calibration. In addition, the original samples from JSRT have inverted color maps, with higher density indicated by clearer regions. D2D and Pre-trained U-Nets were not capable of compensating for this domain shift between JSRT sampes and DRRs. These results are somehow compatible with the findings of Oliveira et al. [20].

Table 2 shows the quantitative results according to the metrics described in Section 3.4 of the proposed methodology and common baselines in the literature for the OpenIST and JSRT datasets – which are the only ones with available pixel-map labels. Bold cells highlight the method with the best results among all for their respective datasets and metrics. In the OpenIST dataset, Conditional DA obtained similar results if compared to simply using the pre-trained DNN for image segmentation trained in the DRR data, as this dataset presents rather similar visual features as the DRRs themselves. Pre-trained DNNs showed considerably better results mainly in Sensitivity, Dice and Jaccard, while Accuracy metrics were rather close in both Pre-trained DNNs and CoDAGANs for OpenIST. Specificity, however, presented significantly better results for CoDAGANs, highlighting a larger portion of non-rib pixels being classified correctly by the method, which is backed up by qualitative analysis in Section 4.2. CoDAGANs also presented better AUC results, implying that this method yields a better trade-off between false positives and false negatives along the ROC curve, as can also be seen in Fig. 3a. Pre-trained models presenting higher Sensitivity and lower Specificity compared with CoDAGANs implies that the former method tends to overshoot the prediction of positive pixels, while the latter has a higher certainty while classifying a pixel as pre-training to a rib.

Even with D2D yielding competitive results in OpenIST, the real advantage of using Conditional DA over both shallow [15,21,22] and deep [3,25,31] baseline methods is seen when comparing the results in the JSRT dataset, because these data present a much larger domain shift from the original DRRs. These distinct visual features between samples from the domains are further highlighted in Section 4.2. One can see in Table 2 that CoDAGANs are able to compensate much more efficiently for the visual differences in the two data sources than other methods, achieving state-of-the-art results for AUC, Accuracy, Sensitivity, Dice and Jaccard methods. The only metric wherein the literature reports better performance than Conditional DA is Specificity, as most of the deep and shallow baseline methods tend to overestimate the amount of negatively labeled pixels, being more susceptible to present higher TNRs. This relatively higher propensity of baseline methods to larger amounts of false positive rib pixels is further evidenced by baselines' near-perfect Specificity scores.

Both OpenIST and JSRT CoDAGANs achieved state-of-the-art results in threshold-independent AUC metric, with 0.8557 and 0.9341 areas under the ROC curve. In the Accuracy metric, we also achieved state-of-the-art results in OpenIST and marginally

Table 2
Quantitative results for rib segmentation in the JSRT and OpenIST datasets yielded from shallow [15,21,22] and deep baselines [3,25,31] and the proposed pipeline based on CoDAGANs [20]. This table summarizes all metrics presented in Section 3.4, which were chosen according to the literature and/or due to their large use in segmentation tasks. Accuracy, Sensitivity, Specificity, Jaccard and Dice are presented in the form $\mu \pm \sigma$ with the average and standard deviation values being computed from the distinct test samples in the dataset, while AUC is shown without a standard deviation because it was computed across all samples at once. Blank cells represent metrics that were not reported in the original paper that proposed their respective method and, thus, could not be used in our comparisons. Bold values indicate the best results for a given metric on the corresponding dataset acquired either from the proposed method or baselines.

Dataset	Method	Metric					
		AUC	Accuracy	Sensitivity	Specificity	Jaccard	Dice
OpenIST	**Pretrained U-Nets** [3]	0.5821	0.78 ± 0.03	0.16 ± 0.05	**0.99 ± 0.00**	0.16 ± 0.05	0.27 ± 0.07
	D2D [25,31]	0.7279	0.81 ± 0.03	0.46 ± 0.14	0.93 ± 0.02	0.38 ± 0.10	0.54 ± 0.11
	CoDAGAN	**0.8557**	**0.82 ± 0.02**	**0.73 ± 0.07**	0.86 ± 0.04	**0.52 ± 0.05**	**0.68 ± 0.04**
JSRT	**Model-based** [15]	0.9105	0.74 ± 0.05	0.71 ± 0.08	0.85 ± 0.03	–	–
	PC [22]	–	0.79 ± 0.05	0.71 ± 0.05	0.85 ± 0.03	–	–
	ICPC [22]	–	**0.86 ± 0.06**	0.79 ± 0.09	0.92 ± 0.04	–	–
	Atlas-based [21]	0.8759	**0.86 ± 0.03**	0.75 ± 0.06	0.92 ± 0.02	–	–
	Pretrained U-Nets [3]	0.5118	0.77 ± 0.04	0.09 ± 0.05	0.93 ± 0.02	0.07 ± 0.04	0.12 ± 0.07
	D2D [25,31]	0.5907	0.80 ± 0.03	0.18 ± 0.07	**0.95 ± 0.02**	0.14 ± 0.06	0.25 ± 0.09
	CoDAGAN	**0.9341**	0.85 ± 0.02	**0.85 ± 0.06**	0.85 ± 0.02	**0.53 ± 0.07**	**0.69 ± 0.06**

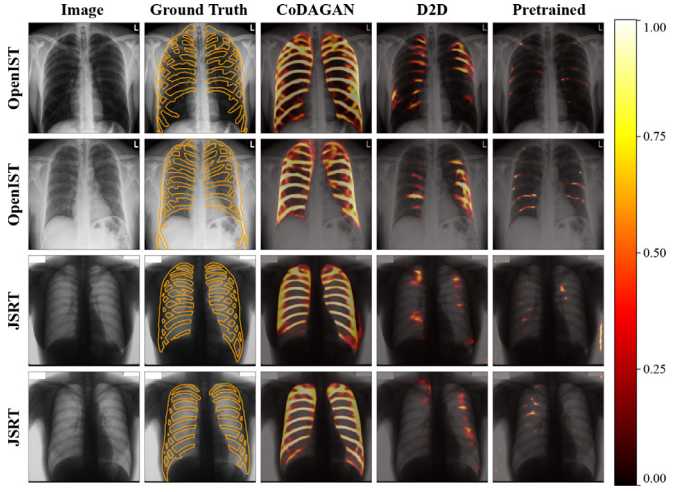

Fig. 4. Sample of probability maps generated by the proposed method and other deep baselines in the datasets with labeled test sets: OpenIST and JSRT. The colormap on the right side indicates the probabilities in the predictions of the three rightmost columns.

worse results than ICPC [22] and Atlas-based [21] segmentations, while also being generalizable to a myriad of datasets. CoDAGANs again achieve Sensitivity, Dice and Jaccard state-of-the-art in both datasets, often with large margins and statistically significant difference between the proposed approach and baselines.

4.2. Generalization qualitative analysis

Fig. 4 shows a small sample of qualitative results in rib segmentation for the JSRT and OpenIST datasets, which are the only ones that have pixel-level labels in our experiments. Visual analysis over the OpenIST samples further reinforce the previously mentioned tendency to overestimate rib pixels of DNNs pre-trained in DRR synthetic samples, while Conditional DA are more conservative in predicting rib pixel labels. It is also evident from the overlayed prediction probability map that the pre-trained models have a much sharper decision boundary than CoDAGANs. That is, the deep baseline method predicts either rib or background pixels with more confidence than Conditional DA, which also results in rougher segmentation boundaries.

Also consistently with the objective results, the most evident advantage of using the proposed pipeline for DA is seen when there is a larger shift between the source and target domains, as can be seen in the two bottom rows of Fig. 4 wherein segmentation results for JSRT are shown. One can clearly see that the Pre-trained U-Net missed the regions in the samples with actual ribs, presenting highly erratic predictions. D2D shows better results than Pre-trained U-Nets, but still fails to recognize most of the ribs in these samples. CoDAGAN, however, are capable of domain generalization, being able to translate the knowledge from the DRR

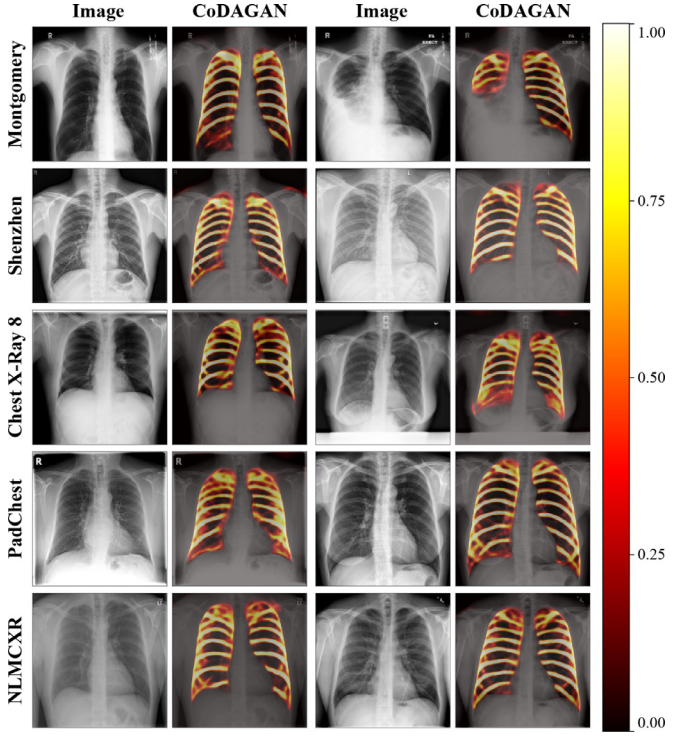

Fig. 5. Sample of probability maps generated by the proposed method in the datasets with unlabeled test sets: Montgomery, Shenzhen, NIH Chest X-ray 8, PadChest and NLMCXR. The colormap on the right side indicates the probabilities in the predictions of the two rightmost columns.

images and noisy labels to JSRT much more effectively, resulting in high quality predictions for the rib semantic maps.

Additionally to the results from the datasets with labeled data, we show a small sample of CoDAGAN's predictions from unlabeled data in 5 additional datasets, as can be seen in Fig. 5. Probability maps predicted by the proposed methodology for all samples in these datasets can be found in this paper's webpage in order to encourage reproducibility.

5. Conclusion

In this paper we presented a novel methodology for UDA applied to the problem of rib segmentation using Conditional Domain Adaptation [20]. The proposed pipeline uses higher dimensional 3D data to acquire two sets of flattened 2D images: DRRs that visually resemble real CXRs – serving as training samples for rib segmentation; and bone segmentation semantic maps that can be curated in order to become pixel-level rib segmentation labels.

We also proposed a novel evaluation procedure for rib segmentation in order to guide future state-of-the-art comparisons in the field, according to a specific set of metrics and public datasets and labels in a conscious effort for reproducibility. One of the label sets used in the standardized experiments for the JSRT dataset was computed according to a bone suppression dataset, curated and made publicly available for other researchers.

Future extensions of this work intend to use similar pipelines to the segmentation of other sets of bones in the human body, also without requiring any annotation from physicians. By modifying the MIP step in Fig. 1, the data from the LIDC-IDRI dataset may allow for similar methodologies for automated unsupervised segmentation of vertebrae, pelvic bones and shoulder bones. In addition to that, further improvements in UDA for dense labeling may allow for fully generalizable segmentation of bony structures in the human body without requiring annotations.

In cases where bone shadows are a problem for detection/diagnosis – as it is the case of lung findings in CXRs – generalizable bone detection can be an important preprocessing step, as the algorithms must know which pixels need suppression. It is also possible to port the proposed pipeline to directly perform rib suppression in CXRs instead of first segmenting the ribs, according to labeled datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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 (424700/2018-2 and 311395/2018-0), and FAPEMIG (APQ-00449-17 and APQ-00519-20 – CAD-COVID-19 Project) for the financial support provided for this research.

References

[1] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[2] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Neural Information Processing Systems, Curran Associates, Inc., 2012, pp. 1097–1105.
[3] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2015, pp. 234–241.
[4] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems (2017) 91–99.
[5] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2015) 304–310.
[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Conference on Computer Vision and Pattern Recognition, 2017.
[7] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, PadChest: A Large Chest X-ray Image Dataset with Multi-Label Annotated Reports, arXiv:1901.07441(2019).
[8] V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: a survey of recent advances, IEEE Signal Process. Mag. 32 (3) (2015) 53–69.
[9] L. Shao, F. Zhu, X. Li, Transfer learning for visual categorization: asurvey, IEEE Trans. Neural Netw. Learn. Syst. 26 (5) (2015) 1019–1034.
[10] J. Zhang, W. Li, P. Ogunbona, Transfer Learning For Cross-Dataset Recognition: A Survey, arXiv:1705.04396(2017).
[11] M. Wang, W. Deng, Deep visual domain adaptation: a survey, Neurocomputing 312 (2018) 135–153.
[12] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, 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, Am. J. Roentgenol. 174 (1) (2000) 71–74.
[13] S. Jaeger, S. Candemir, S. Antani, Y.-X.J. Wáng, P.-X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quant. Imaging Med. Surg. 4 (6) (2014) 475.
[14] Y. Zhang, S. Miao, T. Mansi, R. Liao, Task driven generative modeling for unsupervised domain adaptation: application to X-ray image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2018, pp. 599–607.
[15] B. van Ginneken, B.M. ter Haar Romeny, Automatic delineation of ribs in frontal chest radiographs, in: Medical Imaging 2000: Image Processing, 3979, International Society for Optics and Photonics, 2000, pp. 825–837.
[16] B. Van Ginneken, B.T.H. Romeny, M.A. Viergever, Computer-aided diagnosis in chest radiography: asurvey, IEEE Trans. Med. Imaging 20 (12) (2001) 1228–1241.
[17] E. Soleymanpour, H.R. Pourreza, et al., Fully automatic lung segmentation and rib suppression methods to improve nodule detection in chest radiographs, J. Med. Signals Sens. 1 (3) (2011) 191.
[18] J. Austin, B. Romney, L. Goldsmith, Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect, Radiology 182 (1) (1992) 115–122.
[19] Z. Yue, A. Goshtasby, L.V. Ackerman, Automatic detection of rib borders in chest radiographs, IEEE Trans. Med. Imaging 14 (3) (1995) 525–536.
[20] H.N. Oliveira, E. Ferreira, J.A. Dos Santos, Truly generalizable radiograph segmentation with conditional domain adaptation, IEEE Access 8 (2020) 84037–84062.
[21] S. Candemir, S. Jaeger, S. Antani, U. Bagci, L.R. Folio, Z. Xu, G. Thoma, Atlas-based rib-bone detection in chest X-rays, Comput. Med. Imaging Graph. 51 (2016) 32–39.
[22] M. Loog, B. Ginneken, Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification, IEEE Trans. Med. Imaging 25 (5) (2006) 602–611.
[23] G. Zhang, H. Wu, W. Guo, Rib segmentation in chest radiographs by support vector machine, in: International Conference on Education, Management, Computer and Society, Atlantis Press, 2016.
[24] M. Gargouri, J. Tiern, E. Jolivet, P. Petit, E.D. Angelini, Accurate and robust shape descriptors for the identification of rib cage structures in CT-images with random forests, in: International Symposium on Biomedical Imaging, IEEE, 2013, pp. 65–68.
[25] H.N. Oliveira, J.A. dos Santos, Deep transfer learning for segmentation of anatomical structures in chest radiographs, in: Conference on Graphics, Patterns and Images, IEEE, 2018.
[26] J.P. Cohen, M. Luck, S. Honari, Distribution matching losses can hallucinate features in medical image translation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2018, pp. 529–536.
[27] Y. Zhang, S. Miao, T. Mansi, R. Liao, Task driven generative modeling for unsupervised domain adaptation: application to X-ray image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2018, pp. 599–607.
[28] Z. Zhang, L. Yang, Y. Zheng, Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network, in: Conference on Computer Vision and Pattern Recognition, 2018, pp. 9242–9251.
[29] Y. Tang, Y. Tang, J. Xiao, R.M. Summers, XLSor: a robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation, in: International Conference on Medical Imaging with Deep Learning, 2019.
[30] Y. Tang, Y. Tang, V. Sandfort, J. Xiao, R.M. Summers, TUNA-Net: task-oriented UNsupervised adversarial network for disease recognition in cross-domain chest X-rays, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2019, pp. 431–440.

[31] J. Yang, N.C. Dvornek, F. Zhang, J. Chapiro, M. Lin, J.S. Duncan, Unsupervised domain adaptation via disentangled representations: application to cross–modality liver segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, Springer, 2019, pp. 255–263.
[32] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017.
[33] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784(2014).
[34] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised Image-to-Image Translation Networks, in: Neural Information Processing Systems, 2017, pp. 700–708.
[35] X. Huang, M.-Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image–to-image translation, in: European Conference on Computer Vision, 2018, pp. 172–189.
[36] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Ooptimization, arXiv:1412.6980(2014).
[37] S.G. Armato III, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Med. Phys. 38 (2) (2011) 915–931.
[38] M. Gusarev, R. Kuleev, A. Khan, A.R. Rivera, A.M. Khattak, Deep learning models for bone suppression in chest radiographs, in: Conference on Computational Intelligence in Bioinformatics and Computational Biology, IEEE, 2017, pp. 1–7.
[39] B. Van Ginneken, M.B. Stegmann, M. Loog, Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database, Med. Image Anal. 10 (1) (2006) 19–40.