SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-rays

Wei Dai, Joseph Doyle, Xiaodan Liang, Hao Zhang, Nanqing Dong, Yuan Li, Eric P. Xing

Petuum Inc.

{wei.dai,joe.doyle,xiaodan.,hao.zhang,nanqing.dong,christy.li,eric.xing}@petuum.com

Abstract

Chest X-ray (CXR) is one of the most commonly prescribed medical imaging procedures, often with over 2–10x more scans than other imaging modalities such as MRI, CT scan, and PET scans. These voluminous CXR scans place significant workloads on radiologists and medical practitioners. Organ segmentation is a crucial step to obtain effective computer-aided detection on CXR. In this work, we propose Structure Correcting Adversarial Network (SCAN) to segment lung fields and the heart in CXR images. SCAN incorporates a critic network to impose on the convolutional segmentation network the structural regularities emerging from human physiology. During training, the critic network learns to discriminate between the ground truth organ annotations from the masks synthesized by the segmentation network. Through this adversarial process the critic network learns the higher order structures and guides the segmentation model to achieve realistic segmentation outcomes. Extensive experiments show that our method produces highly accurate and natural segmentation. Using only very limited training data available, our model reaches human-level performance without relying on any existing trained model or dataset. Our method also generalizes well to CXR images from a different patient population and disease profiles, surpassing the current state-of-the-art.

1. Introduction

Chest X-ray (CXR) is one of the most common medical imaging modalities. Due to CXR's low cost and low dose of radiation, hundreds to thousands of CXRs are generated in a typical hospital daily, which creates significant diagnostic workloads. In 2015/16 year over 22.5 million X-ray images were requested in UK's public medical sector, constituting over 55% of the total number of medical images and dominating all other imaging modalities such as computed tomography (CT) scan (4.5M) and MRI (3.1M) [8]. Among X-ray images, 8 millions are Chest X-rays, which translate to thousands of CXR readings per radiologist per year. The shortage of radiologists is well documented in the
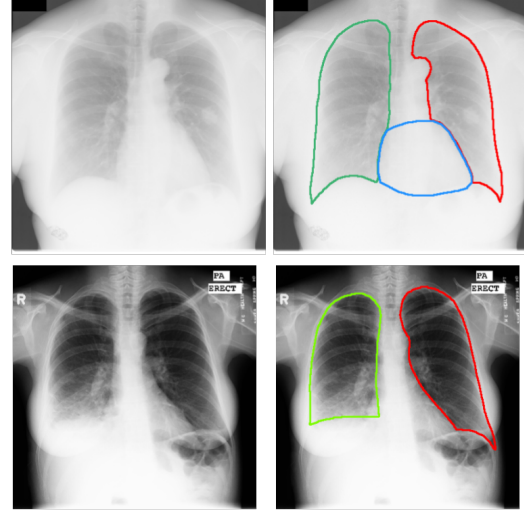

Figure 1: Two example chest X-ray (CXR) images from two dataset: JSRT (top) and Montgomery (bottom). From left to right columns show the original CXR images and the lung fields annotation by radiologists. JSRT (top) additionally has the heart annotation. Note that contrast can vary significantly between the dataset, and pathological lung profiles such as the bottom patient pose a significant challenge to the segmentation problem.

developed world [19, 18], not to mention developing countries [1]. Compared with the more modern medical imaging technologies such as CT scan and PET scans, X-rays pose diagnostic challenges due to their low resolution and 2-D projection. It is therefore of paramount importance to develop computer-aided detection methods for chest X-rays to support clinical practitioners.

An important step in computer-aided detection on CXR images is organ segmentation. The segmentation of the lung fields and the heart provides rich structure information about shape irregularities and size measurements that can be used to directly assess certain serious clinical conditions, such as cardiomegaly (enlargement of the heart), pneumothorax (lung collapse), pleural effusion, and emphysema. Furthermore, explicit lung region masks can also improve interpretability of computer-aided detection, which is important for the clinical use.
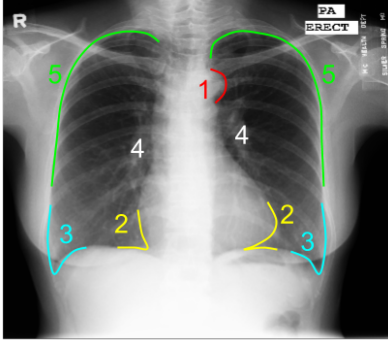
Figure 2: Important contour landmarks around lung fields: aortic arch (1) is excluded from lung fields; costophrenic angles (3) and cardiodiaphragmatic angles (2) should be visible in healthy patients. Hila and other vascular structures (4) are part of the lung fields. The rib cage contour (5) should be clear in healthy lungs.

One major challenge in CXR segmentation is to incorporate the implicit medical knowledge involved in contour determination. In the most basic sense, the positional relationship between the lung fields and the heart implies the adjacency of the lung and heart masks. Moreover, when medical experts annotate the lung fields, they look for certain consistent structures surrounding the lung fields (Fig. 2). Such prior knowledge helps resolve boundaries around less clear regions caused by pathological conditions or poor imaging quality, as can be seen in Fig. 1. Therefore, a successful segmentation model must effectively leverage global structural information to resolve the local details.

Unfortunately, unlike natural images, there is very limited CXR training data with pixel-level annotations, due to the expensive label acquisition involving medical professionals. Furthermore, CXRs exhibit substantial variations across different patient populations, pathological conditions, as well as imaging technology and operation. Finally, CXR images are gray-scale and are drastically different from natural images, which may limit the transferability of existing models. Existing approaches to CXR organ segmentation generally rely on hand-crafted features that can be brittle when applied on a different patient population, disease profiles, and image quality. Furthermore, these methods do not explicitly balance local information with global structure in a principled way, which is critical to achieve realistic segmentation outcomes suitable for diagnostic tasks.

In this work, we propose to use the Structure Correcting Adversarial Network (SCAN) framework that incorporates a critic network to guide the convolutional segmentation network to achieve accurate and realistic chest X-ray organ segmentation. By employing a convolutional network approach to organ segmentation, we side-step the problems faced by existing approaches based on ad hoc feature engineering. Our convolutional segmentation model alone can achieve performance competitive with existing methods. However, the segmentation model alone can not cap-

ture sufficient global structure to produce natural contours due to the limited training data. To impose regularization based on the physiological structures, we introduce a critic network that discriminates between the ground truth annotations from the masks synthesized by the segmentation network. The segmentation network and the critic network can be trained end-to-end. Through this adversarial process the critic network learns the higher order regularities and effectively transfers this global information back to the segmentation model to achieve realistic segmentation outcomes.

We demonstrate that SCAN produces highly realistic and accurate segmentation outcomes even when trained on very small dataset, without relying on any existing models or data from other domains. With the global structural information, our segmentation model is able to resolve difficult boundaries that require a strong prior knowledge. Using intersection-over-union (IoU) as the evaluation metric, SCAN improves the segmentation model by 1.8% absolutely and achieves 94.7% for the lung fields and 86.6% for the heart, both of which are the new state-of-the-art by a single model, competitive with human experts (94.6% for the lungs and 87.8% for the heart). We further show that SCAN model is more robust when applied to a new, unseen dataset, outperfoming the vanilla segmentation model by 4.3%.

2. Related Work

Our review focuses on two lines of literature most relevant to our problem: lung field segmentation and semantic segmentation with convolutional neural networks.

Lung Field Segmentation. Existing work on lung field segmentation broadly falls into three categories [30]. (1) Rule-based systems apply pre-defined set of thresholding and morphological operations that are derived from heuristics [12]. (2) Pixel classification methods classify the pixels as inside or outside of the lung fields based on pixel intensities [37, 15, 16, 2]. (3) More recent methods are based on deformable models such as Active Shape Model (ASM) and Active Appearance Model [7, 6, 28, 29, 24, 32, 23, 33]. Their performance can be highly variable due to the tuning parameters and whether shape model is initialized to the actual boundaries. Also, the high contrast between rib cage and lung fields can cause the model to be trapped in local minima. Our approach uses convolutional networks to perform end-to-end training from images to pixel masks without using ad hoc features. The proposed adversarial training further incorporates prior structural knowledge in a unified framework.

The current state-of-the-art method for lung field segmentation uses registration-based approach [3]. To build a lung model for a test patient, [3] finds patients in an existing database that are most similar to the test patient and perform linear deformation of their lung profiles based on key point matching. This approach relies on the test patients being well modeled by the existing lung profiles and
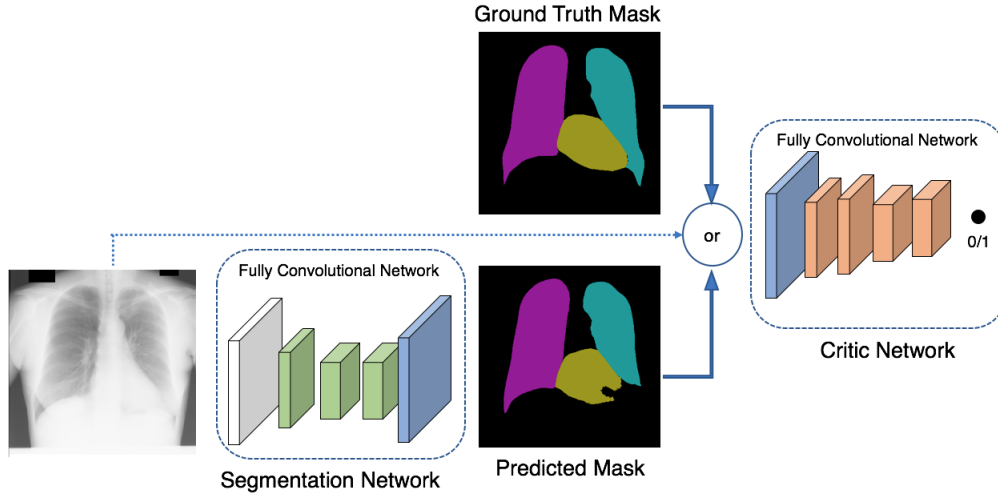
Figure 3: Overview of the proposed SCAN framework that jointly trains a segmentation network and a critic network with an adversarial mechanism. The segmentation network produces per-pixel class prediction. The critic takes either the ground truth label or the prediction by the segmentation network, optionally with the CXR image, and output the probability estimate of whether the input is the ground truth (with training target 1) or the segmentation network prediction (with training target 0).

correctly matched key points, both of which can be brittle on a different population.

Semantic Segmentation with Convolutional Networks. Semantic segmentation aims to assign a pre-defined class to each pixel, which requires a high level visual understanding. Current state-of-the-art methods for semantic segmentation use fully convolutional network (FCN) [14, 35, 5, 13]. Recently [17] applies adversarial training to semantic segmentation and observe some improvement. These works address the natural images with color input, and are pretrained with models such as VGG network [27] incorporating the learning from large-scale image classification [22]. We adapt FCNs to gray-scale CXR images under the stringent constraint of a very limited training dataset of 247 images. Our FCN departs from the usual VGG architecture and can be trained without transfer learning from existing models or dataset.

Separately, U-net [21] and similar architectures are popular convolutional networks for biomedical segmentation and have been applied to neuronal structure [21] and histology images [4]. In this work we propose to use adversarial training on existing segmentation networks to enhance the global consistency of the segmentation outcomes.

We note that there is a growing body of recent works that apply neural networks end-to-end on CXR images [25, 34]. These models directly output clinical targets such as disease labels without well-defined intermediate outputs to aid interpretability. Furthermore, they generally require a large number of CXR images for training, which is not readily available for many clinical tasks involving CXR images.

3. Problem Definition

We address the problem of segmenting the left lung field, the right lung field, and the heart on chest X-rays (CXRs) in

the posteroanterior (PA) view, in which the radiation passes through the patient from the back to the front. Due to the fact that CXR is a 2D projection of a 3D structure, organs overlap significantly and one has to be careful in defining the lung fields. We adopt the definition from [31]: lung fields consist of all the pixels for which radiation passes through the lung but not through the following structures: the heart, the mediastinum (the opaque region between the two lungs), below the diaphragm, the aorta, and, if visible, the superior vena cava (Fig. 2). The heart boundary is generally visible on two sides, while the top and bottom borders of the heart have to be inferred due to occlusion by the mediastinum. As can be seen in Fig. 1, this definition captures the common notion of lung fields and the heart, and include regions pertinent to CXR reading in the clinical settings.

4. Structure Correcting Adversarial Network

We detail our approach to semantic segmentation of lung fields and the heart using the proposed Structure Correcting Adversarial Network (SCAN) framework. To tailor to the special problem setting of CXR images, we develop our network architecture from ground up following the best practices and extensive experimentation. Using a dataset over an order of magnitude smaller than common semantic segmentation datasets for natural images, our model can be trained end-to-end from scratch to an excellent generalization capability without relying on existing models or datasets.

4.1. Adversarial Training for Semantic Segmentation

Adversarial training was first proposed in Generative Adversarial Network (GAN) [9] in the context of genera-

tive modeling[1]. The GAN framework consists of a generator network and a critic network that engage in an adversarial two-player game, in which the generator aims to learn the data distribution and the critic estimates the probability that a sample comes from the training data instead of synthesized by the generator. The generator's objective is to maximize the probability that the critic makes a mistake, while the critic is optimized to minimize the chance of mistake. It has been demonstrated that the generator produces samples (e.g., images) that are highly realistic [20].

A key insight in this adversarial process is that the critic, which itself can be a complex neural network, can learn to exploit higher order inconsistencies in the samples synthesized by the generator. Through the interplay of the generator and the critic, the critic can guide the generator to produce samples more consistent with higher order structures in the training samples, resulting in a more "realistic" data generation process.

The higher order consistency enforced by the critic is particularly desirable for CXR segmentation. Human anatomy, though exhibiting substantial variations across individuals, generally maintains stable relationship between physiological structures (Fig. 2). CXRs also pose consistent views of these structures thanks to the standardized imaging procedures. We can, therefore, expect the critic to learn these higher order structures and guide the segmentation network to generate masks more consistent with the learned global structures.

We propose to use adversarial training for segmenting CXR images. Fig. 3 shows the overall SCAN framework in incorporating adversarial process to the semantic segmentation. The framework consists of a segmentation network and a critic network that are jointly trained. The segmentation network makes pixel-level predictions of the target classes, playing the role of the generator in GAN but conditioned on an input image. On the other hand, the critic network takes as input the segmentation masks and outputs the probability that the input mask is the ground truth annotations instead of the prediction by the segmentation network. The network can be trained jointly through a minimax scheme that alternates between optimizing the segmentation network and the critic network.

4.2. Training Objectives

Let S, D be the segmentation network and the critic network, respectively. The data consist of the input images $\boldsymbol{x}_i$ and the associated mask labels $\boldsymbol{y}_i$, where $\boldsymbol{x}_i$ is of shape $[H, W, 1]$ for a single-channel gray-scale image with height H and width W, and $\boldsymbol{y}_i$ is of shape $[H, W, C]$ where C is the number of classes including the background. Note that for each pixel location (j, k), $y_i^{jkc} = 1$ for the labeled class

channel c while the rest of the channels are zero ($y_i^{jkc'} = 0$ for $c' \neq c$). We use $S(\boldsymbol{x}) \in [0, 1]^{[H, W, C]}$ to denote the class probabilities predicted by S at each pixel location such that the class probabilities normalize to 1 at each pixel. Let $D(\boldsymbol{x}_i, \boldsymbol{y})$ be the scalar probability estimate of $\boldsymbol{y}$ coming from the training data (ground truth) $\boldsymbol{y}_i$ instead of the predicted mask $S(\boldsymbol{x}_i)$. We define the optimization problem as

$$
\min_S \max_D \Big\{ J(S, D) := \sum_{i=1}^{N} J_s(S(\boldsymbol{x}_i), \boldsymbol{y}_i) \\
- \lambda \Big[J_d(D(\boldsymbol{x}_i, \boldsymbol{y}_i), 1) + J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 0) \Big] \Big\}
\tag{1}
$$

, where $J_s(\hat{\boldsymbol{y}}, \boldsymbol{y}) := \frac{1}{HW} \sum_{j,k} \sum_{c=1}^{C} -y^{jkc} \ln y^{jkc}$ is the multi-class cross-entropy loss for predicted mask $\hat{\boldsymbol{y}}$ averaged over all pixels. $J_d(\hat{t}, t) := -t \ln \hat{t} + (1 - t) \ln(1 - \hat{t})$ is the binary logistic loss for the critic's prediction. λ is a tuning parameter balancing pixel-wise loss and the adversarial loss. We can solve Eq. (1) by alternate between optimizing S and optimizing D using their respective loss functions.
Training the Critic: Since the first term in Eq. (1) does not depend on D, we can train our critic network by *minimizing* the following objective with respect to D for a fixed S:

$$
\sum_{i=1}^{N} J_d(D(\boldsymbol{x}_i, \boldsymbol{y}_i), 1) + J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 0)
$$

Training the Segmentation Network: Given a fixed D, we train the segmentation network by minimizing the following objective with respect to S:

$$
\sum_{i=1}^{N} J_s(S(\boldsymbol{x}_i), \boldsymbol{y}_i) + \lambda J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 1)
$$

Note that we use $J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 1)$ in place of $-J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 0)$, following the recommendation in [9]. This is valid as they share the same set of critical points. The reason for this substitution is that $J_d(D(\boldsymbol{x}_i, S(\boldsymbol{x}_i)), 0)$ leads to weaker gradient signals when D makes accurate predictions, such as during the early stage of training.

4.3. Segmentation Network

Our segmentation network is a Fully Convolutional Network (FCN), which is also the core component in many state-of-the-art semantic segmentation models [14, 35, 5]. The success of FCN can be largely attributed to convolutional neural network's excellent ability to extract high level representations suitable for dense classification. FCN can be divided into two modules: the down-sampling path and the up-sampling path. The down-sampling path consists of convolutional layers and max or average pooling layers, with architecture similar to those used in image classification [27]. The down-sampling path can extract the high level semantic information, usually at a lower spatial resolution. The up-sampling path consists of convolutional and

---

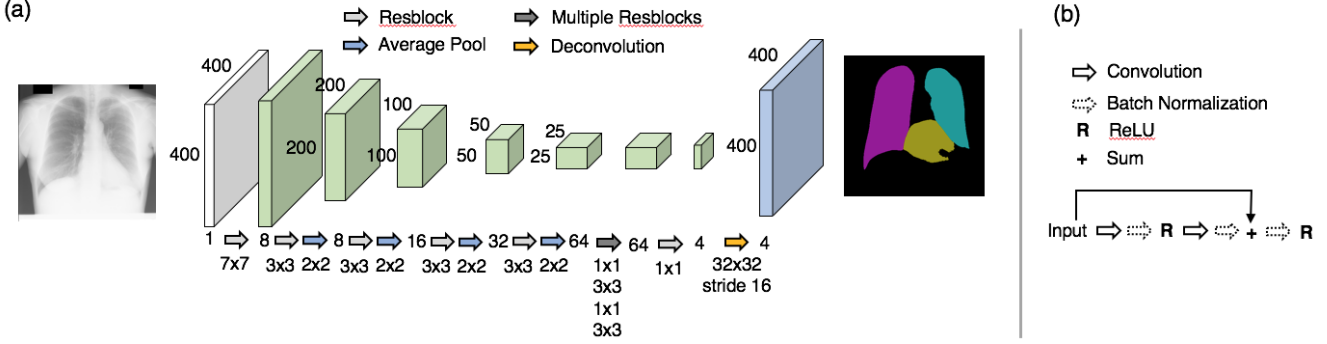[1]We point out that GAN bears resemblance to the actor-critic model in the existing reinforcement learning paradigm.

Figure 4: The segmentation network architecture. (a) Fully convolutional network for dense prediction. The feature map resolutions (e.g., 400×400) are denoted only for layers with resolutions different from the previous layer. The arrows denotes the forward pass, and the integer sequence (1, 8, 16,...) is the number of feature maps. $k \times k$ below Resblock (residual block), average pool, and deconvolution arrows indicates the receptive field sizes. The dark gray arrow denotes 5 resblocks. All convolutional layers have stride 1×1, while all average pooling layers have stride 2×2. The output is the class distribution for 4 classes (3 foreground + 1 background). The total number of parameters is 271k, $\sim$ 500x smaller than the VGG-based down-sampling path in [14]. (b) The residual block architecture is based on [11]. The residual block maintains the same number of feature maps and spatial resolution which we omit here. Best viewed in color.

deconvolutional layers (also called transposed convolution) to predict scores for each classes at the pixel level using the output of the down-sampling path.

Most FCNs are applied to color images with RGB channels, and their down-sampling paths are initialized with parameters trained in large-scale image classification [14]. However, CXR is gray-scale and thus the large model capacity used in image classification networks that leverages the richer RGB input is likely to be counter-productive for our purpose. Furthermore, our FCN architecture has to be highly parsimonious to take into account that our training dataset of 247 CXR images is orders of magnitude smaller than those in the natural image domains. Lastly, in our task we focus on segmenting three classes (the left lung, the right lung, and the heart), which is a smaller classification space compared with dataset such as PASCAL VOC which has 20 class objects. A more parsimonious model configuration is therefore highly favorable in this setting.

Figure 4 shows our FCN architecture. We find that it is advantageous to use much fewer feature maps than the conventional VGG-based down-sampling path. Specifically, we start with just 8 feature maps in the first layers, compared with 64 feature maps in the first layer of VGG [27]. To obtain sufficient model capacity, we instead go deep with 20 convolutional layers. We also interleave 1×1 convolution with 3×3 in the final layers to emulate the bottleneck design [10]. All in all the segmentation network contains 271k parameters, 500x smaller than VGG-based FCN [14]. We employ residual blocks [10] (Fig. 4(b)) to aid optimization. The parsimonious network construction allows us to optimize it efficiently without relying on any existing trained model, which is not readily available for gray-scale images.
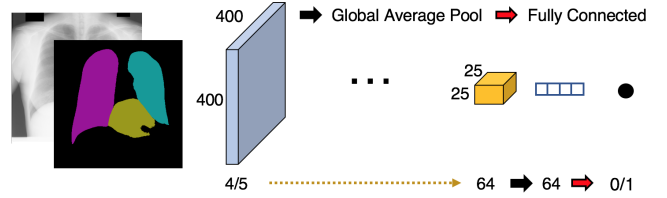

Figure 5: The critic network architecture. Our critic FCN mirrors the segmentation network (Fig. 4). The input to the critic network has 4 channels or, when the input image is included, 5 channels. The subsequent layers are the same as the segmentation network up to the yellow box with 64 channel, which corresponds to the last green box in Fig. 4 with 64 channels. We omit the intermediate layers and refer readers to Fig. 4. The training target is 0 if the input is from segmentation network, and 1 for ground truth labels.

4.4. Critic Network

Our critic network mirrors the construction of the segmentation network, and is also a fully convolutional network. Fig. 5 shows the architecture, omitting the intermediate layers that are identical to the segmentation networks. This way the critic network contains similar model capacity as the segmentation network with a similar field of view, which is important due to the large object size in the CXR images. We can optionally include the original CXR image as input to the critic as an additional channel, which is a more economic approach to incorporate the image in critic network than [17]. Preliminary experiment shows that including the original CXR image does not improve performance, and thus for simplicity we feed only the mask prediction to the critic network. Overall our critic network has 258k parameters, comparable to the segmentation network.

5. Experiments

We perform extensive evaluation of the proposed SCAN framework and demonstrate that our approach produces

highly accurate and realistic segmentation of the lung fields and the heart in CXR images.

5.1. Dataset and Processing

We use two publicly available dataset to evaluate our proposed SCAN network for the segmentation of lung fields and the heart on CXR images. To the best of our knowledge, these are the only two publicly available dataset with at least lung field annotations. We point out that the dataset come from two different countries with different lung diseases, representing diverse CXR samples.

JSRT. The JSRT dataset was released by Japanese Society of Radiological Technology (JSRT) [26] and the lung fields and the heart masks labels were made available by [31] (Fig. 1). The dataset contains 247 CXRs, among which 154 have lung nodules and 93 have no lung nodule. All images have resolution 2048×2048 in gray-scale with color depth of 12 bits. We point out that this dataset represents mostly normal lung and heart masks despite the fact that the majority of patients have lung nodule. The reason is that lung nodules in most cases do not alter the contour of the lungs and the heart, especially when the lung nodules are small.

Montgomery. The Montgomery dataset contains images from the Department of Health and Human Services, Montgomery County, Maryland, USA. The dataset consists of 138 CXRs, including 80 normal patients and 58 patients with manifested tuberculosis (TB). The CXR images are 12-bit gray-scale images of dimension 4020×4892 or 4892×4020. Only the two lung masks annotations are available (Fig. 1).

We scale all images to 400×400 pixels, which retains sufficient visual details for vascular structures in the lung fields and the boundaries. Preliminary experiments suggests that increasing the resolution to 800×800 pixels does not improve the segmentation performance, consistent with the observation in [3]. Due to the high variation in image contrast between dataset (Fig. 1), we perform per-image normalization. Given an image x we normalize it with $\tilde{x}^{jk} := \frac{x^{jk} - \bar{x}}{\sqrt{\mathrm{var}(x)}}$, where $\bar{x}$ and $\mathrm{var}(x)$ are the mean and variance of pixels in x, respectively. Note we do not use statistics from the whole dataset. Data augmentation by rotating and zooming images, did not improve results in our preliminary experiments. We thus did not apply any data augmentation.

In post-processing, we fill in any hole in the predicted mask, and remove the small patches disjoint from the largest mask. We observe that in practice this is important for the prediction output of the segmentation network (FCN alone), but does not affect the evaluation results for FCN with adversarial training.

5.2. Training Protocols

GANs are known to be unstable during the training process and can "collapse" when the generator produces outcomes that lies in a much smaller subspace than the data distribution. To mitigate this problem, we pre-train the segmentation network using only the pixel-wise loss J_s (Eq. (1)), which also gives faster training than the full adversarial training, as training the segmentation network using pixel losses involves forward and backward propagation through the segmentation network only but not the critic network. We use Adam optimizer with learning rate 0.0002 to train all models for 350 epochs, defined as a pass over the training set. We use mini-batch size 10. When training involves critic network, for each mini-batch we perform 5 optimization steps on the segmentation network for each optimization step on the critic network. Our training takes place on machines equipped with a Titan X GPU.

We use the following two metrics for evaluation: **Intersection-over-Union (IoU)** is the agreement between the ground truth and the estimated segmentation mask. Formally, let P be the set of pixels in the predicted segmentation mask for a class and G the set of pixels in the ground truth mask for the same class. We can define IoU as $\frac{|P \cap G|}{|P \cup G|} = \frac{|TP|}{|TP| + |FP| + |FN|}$, where TP, FP, FN denotes for the set of pixels that are true positive, false positive, and false negative, respectively. **Dice Coefficient** is a popular metric for segmentation in the medical domain. Using the notation defined above, Dice coefficient can be calculated as $\frac{2|P \cap G|}{|P| + |G|} = \frac{2|TP|}{2|TP| + |FP| + |FN|}$.

5.3. Experiment Design and Results

We randomly divide the JSRT dataset into the development set (209 images) and the evaluation set (38 images). We tune our architecture and hyperparameters (such as λ in Eq. (1)) using a validation set within the development set. Similarly, we randomly divide the Montgomery dataset into the development set (117 images) and the evaluation set (21 images). We tune our hyperparameters on the JSRT development set and use the same for the Montgomery experiments. We use FCN for the segmentation network only architecture, and SCAN for the full framework.

Quantitative Comparison. In our first experiment we compare FCN with SCAN when trained on the JSRT development set and tested on the JSRT evaluation set. Table 1 shows the IoU and Dice scores. We observe that the adversarial training significantly improves the performance. In particular, IoU for the two lungs improves from 92.9% to 94.7%. We also find that the performance of adversarial training is robust across a range of λ: the IoU for both lungs with $\lambda = 0.1, 0.01, 0.001$ is $94.4\% \pm 0.4\%$, $94.5\% \pm 0.4\%$, and $94.7\% \pm 0.4\%$, respectively.

In Table 2 we compare our approach to several existing methods on the JSRT dataset, as well as human performance. Our model surpasses the current state-of-the-art method which is a registration-based model [3] by a significant margin. Furthermore, our method is competitive with human performance for both lung fields and the heart.

Ground Truth

FCN

IoU (lungs): 88.4%
IoU (heart): 86.4%

IoU (lungs): 86.0%
IoU (heart): 91.7%

IoU (lungs): 87.2%

IoU (lungs): 89.7%

SCAN

IoU (lungs): 91.4%
IoU (heart): 88.9%

IoU (lungs): 95.7%
IoU (heart): 91.7%

IoU (lungs): 93.0%
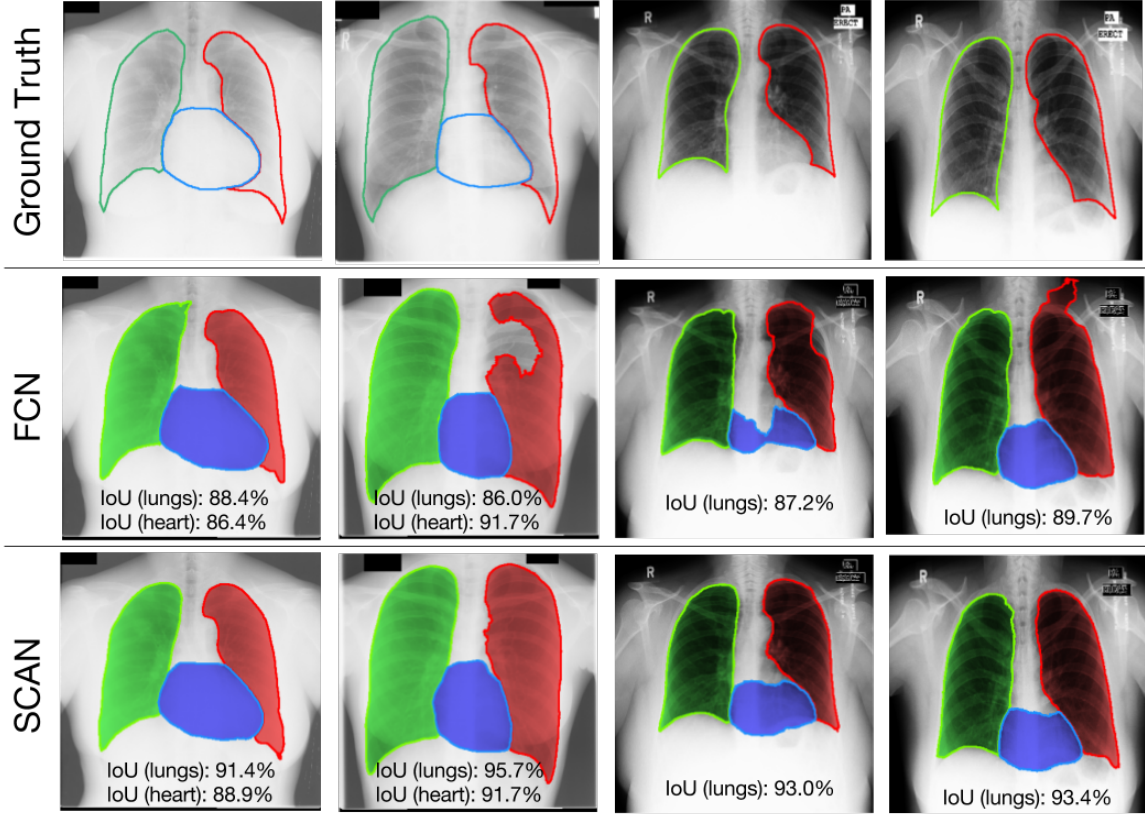
IoU (lungs): 93.4%

Figure 6: Visualization of segmentation results. From top to bottom: ground truth, FCN prediction without adversarial training, FCN prediction with adversarial training. The contours of the predicted mask are added for visual clarity. Each column is a patient. The left two columns are patients from the JSRT evaluation set with models trained on JSRT development set. The right two columns are from the Montgomery dataset using a model trained on the full JSRT dataset but not Montgomery, which is a much more challenging scenario. Note that only the two patients from JSRT dataset (left two columns) have heart annotations for evaluation of heart area IoU.

		FCN	SCAN
IoU	Left Lung	91.3% ± 0.9%	**93.8**% ± 0.8%
	Right Lung	94.2% ± 0.2%	**95.5**% ± 0.2%
	Both Lungs	92.9% ± 0.5%	**94.7**% ± 0.4%
	Heart	86.5% ± 0.9%	86.6% ± 1.2%
Dice	Left Lung	95.4% ± 0.5%	**96.8**% ± 0.5%
	Right Lungs	97.0% ± 0.1%	**97.7**% ± 0.1%
	Both Lungs	96.3% ± 0.3%	**97.3**% ± 0.2%
	Heart	92.7% ± 0.6%	92.7% ± 0.2%

Table 1: IoU and Dice scores on JSRT evaluation set for left lung (on the right side of the PA view CXR), right lung (on the left side of the image), both lungs, and the heart. The model is trained on the JSRT development set.

	IoU (Lungs)	IoU (Heart)
Human Observer [31]	**94.6**% ± 1.8%	**87.8**% ± 5.4%
Ours (SCAN)	**94.7**% ± 0.4%	**86.6**% ± 1.2%
Registration [3]	92.5% ± 0.4%	–
ShRAC [36]	90.7% ± 3.3%	
ASM [31]	90.3% ± 5.7%	79.3% ± 11.9%
AAM [31]	84.7% ± 9.5%	77.5% ± 13.5%
Mean Shape [31]	71.3% ± 7.5%	64.3% ± 14.7%

Table 2: Comparison with existing single-model approaches to lung field segmentation using JSRT dataset. Note that except the registration method, all other methods have slightly different evaluation schemes (e.g., data splits) than our evaluation. Human performance is included as a reference. IoU for heart is omitted for methods addressing only the segmentation of lung field.

For clinical deployment it is important for the segmentation model to generalize to a different population with different imaging quality, such as when deployed in another country or a specialty hospital with very different disease distribution among the patients. In our next experiment we therefore train our model on the full JSRT dataset, which is collected in Japan from population with lung nodules, and test the model on the full Montgomery dataset, which is collected in the U.S. from patients potentially with TB. As

can be seen in Fig. 1, the two dataset present very different contrast and background diseases. Table 3 shows that FCN alone does not generalize well to a new dataset, as IoU for both lungs degrades to 87.1%. However, SCAN substantially improves the performance, surpassing state-of-the-art method based on registration [3].

We further investigate the scenario when training on the two development sets from JSRT and Montgomery *com-*
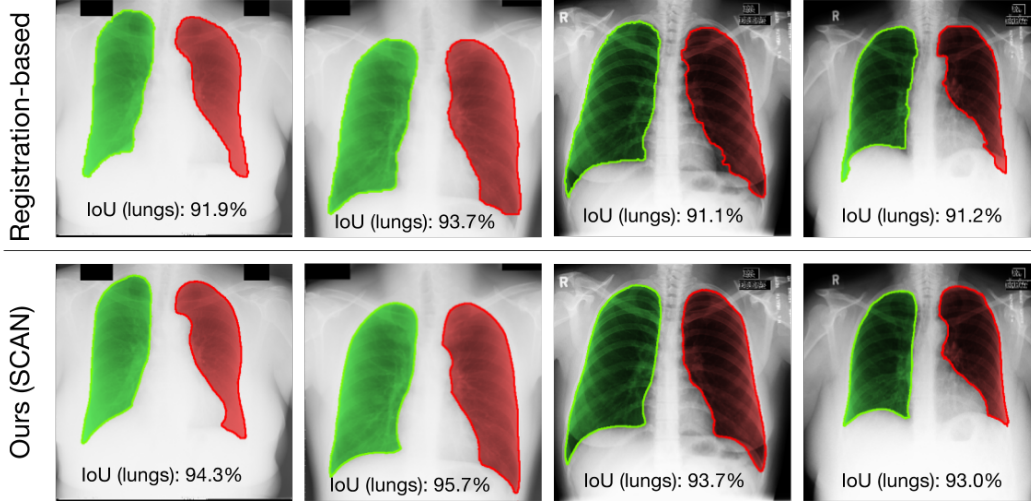
Figure 7: Comparison with the registration-based method that is the current state-of-the-art [3]. The left two columns are from the JSRT evaluation set (using model trained on JSRT development set), and the right two columns are from the Montgomery set (using model trained on the full JSRT dataset). For the left two columns our method produces more realistic contours around the sharp costophrenic anglos. For the right two columns [3] struggles due to the mismatch between test patient lung profile and the existing lung models in JSRT.

bined. Without any further hyperparameter tuning, SCAN improves the IoU on two lungs to $95.1\% \pm 0.43\%$ on the JSRT evaluation set, and $93.0\% \pm 1.4\%$ on the Montgomery evaluation set, a significant improvement compared with when training on JSRT development set alone.

Qualitative Comparison. Fig. 6 shows the qualitative results from these two experiments. The failure cases in the middle row by our FCN reveals the difficulties arising from CXR images' varying contrast across samples. For example, the apex of the ribcage of the rightmost patient's is mistaken as an internal rib bone, resulting in the mask "bleeding out" to the black background, which has similar intensity as the lung field. Vascular structures near mediastinum and anterior rib bones (which appears very faintly in the PA view CXR) within the lung field can also have similar intensity and texture as exterior boundary, leading the FCN to make the drastic mistakes seen in the middle two columns. SCAN significantly improves all of the failure cases and produces much more natural outlines of the organs. We also notice that adversarial training sharpens the segmentation of costophrenic angle (the sharp angle at the junction of ribcage and diaphragm). Costophrenic angles are important in diagnosing pleural effusion and lung hyperexpansion, among others.

	IoU (Both Lungs)
Ours (SCAN)	**91.4**% ± 0.6%
Ours (FCN)	87.1% ± 0.8%
Registration [3]	90.3% ± 0.5%

Table 3: Performance on the full Montgomery dataset using model trained on the full JSRT dataset. Compared with the JSRT dataset, the Montgomery dataset exhibits much higher degree of lung abnormalities and varying imaging quality. This setting tests the robustness of the method in generalizing to a different population and imaging setting.

Our SCAN framework is efficient at test time, as it only needs to perform a forward pass through the segmentation network but not the critic network. Table 4 shows the run time of our method compared with [3] on a laptop. [3] takes much longer due to the need to search through lung models in the training data to find similar profiles, incurring linear cost in the size of training data. In clinical setting such as TB screening [1] a fast test time result is highly desirable.

	Test time
Ours (SCAN)	0.84 seconds
Registration [3]	26 seconds
Human	~2 minutes

Table 4: Prediction time for each CXR image (resolution 400×400) from the Montgomery dataset on a laptop with Intel Core i5, along with the estimated human time.

6. Conclusion

In this work we present the Structure Correcting Adversarial Network (SCAN) framework that applies the adversarial process to develop an accurate semantic segmentation model for segmenting the lung fields and the heart in chest X-ray (CXR) images. SCAN jointly optimizes the segmentation model based on fully convolutional network (FCN) and the adversarial critic network which discriminates the ground truth annotation from the segmentation network predictions. SCAN is simple and yet effective, producing highly accurate and realistic segmentation. Our approach improves the state-of-the-art and achieves performance competitive with human experts. To our knowledge this is the first successful application of convolutional neural networks to CXR image segmentation, and our method holds the promise to integrate with many downstream tasks in computer-aided detection on CXR images

7. Acknowledgement

We thank Carol Cheng and Ellen Sun for providing the medical insights into understanding chest x-ray and the clinical practices around it. Their help has been very instrumental to this project.

References

[1] Computer-aided detection for tuberculosis. http://apps.who.int/ehealth/resources/compendium_ehealth2012_4.pdf. 1, 8

[2] P. Annangi, S. Thiruvenkadam, A. Raja, H. Xu, X. Sun, and L. Mao. A region based active contour method for x-ray lung segmentation using prior shape and low level features. In *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*, pages 892–895. IEEE, 2010. 2

[3] 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*, 33(2):577–590, 2014. 2, 6, 7, 8

[4] H. Chen, X. Qi, L. Yu, and P.-A. Heng. Dcan: Deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016. 3

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. 3, 4

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 2

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[8] N. England. Diagnostic imaging dataset annual statistical release 2015/16, 2016. 1

[9] 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*, pages 2672–2680, 2014. 3, 4

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 5

[11] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 5

[12] L. Li, Y. Zheng, M. Kallergi, and R. A. Clark. Improved method for automatic identification of lung regions on chest radiographs. *Academic Radiology*, 8(7):629–638, 2001. 2

[13] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016. 3

[14] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 3, 4, 5

[15] M. Loog and B. Ginneken. Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. *IEEE Transactions on Medical Imaging*, 25(5):602–611, 2006. 2

[16] M. Loog and B. van Ginneken. Supervised segmentation by iterated contextual pixel classification. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 925–928. IEEE, 2002. 2

[17] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016. 3, 5

[18] T. R. C. of Radiologists. Clinical radiology uk workforce census 2014 report, September 2016. 1

[19] T. R. C. of Radiologists. Clinical radiology uk workforce census 2015 report, September 2016. 1

[20] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 4

[21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[23] D. Seghers, D. Loeckx, F. Maes, D. Vandermeulen, and P. Suetens. Minimal shape and intensity cost path segmentation. *IEEE Transactions on Medical Imaging*, 26(8):1115–1129, 2007. 2

[24] Y. Shi, F. Qi, Z. Xue, L. Chen, K. Ito, H. Matsuo, and D. Shen. Segmenting lung fields in serial chest radiographs using both population-based and patient-specific shape statistics. *IEEE transactions on medical imaging*, 27(4):481–494, 2008. 2

[25] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *arXiv preprint arXiv:1603.08486*, 2016. 3

[26] 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*, 174(1):71–74, 2000. 6

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 4, 5

[28] B. Van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever. Active shape model

segmentation with optimal features. *IEEE transactions on medical imaging*, 21(8):924–933, 2002. 2

[29] B. van Ginneken, S. Katsuragawa, B. M. ter Haar Romeny, K. Doi, and M. A. Viergever. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE transactions on medical imaging*, 21(2):139–149, 2002. 2

[30] B. Van Ginneken, B. T. H. Romeny, and M. A. Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on medical imaging*, 20(12):1228–1241, 2001. 2

[31] 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*, 10(1):19–40, 2006. 3, 6, 7

[32] T. Xu, M. Mandal, R. Long, I. Cheng, and A. Basu. An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. *Computerized Medical Imaging and Graphics*, 36(6):452–463, 2012. 2

[33] T. Xu, M. Mandal, R. Long, I. Cheng, and A. Basu. An edge-region force guided active shape approach for automatic lung field detection in chest radiographs. *Computerized Medical Imaging and Graphics*, 36(6):452–463, 2012. 2

[34] P.-P. Ypsilantis and G. Montana. Learning what to look in chest x-rays with a recurrent visual attention model. *arXiv preprint arXiv:1701.06452*. 3

[35] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3, 4

[36] T. Yu, J. Luo, and N. Ahuja. Shape regularized active contour using iterative global search and local optimization. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 655–662. IEEE, 2005. 7

[37] Z. Yue, A. Goshtasby, and L. V. Ackerman. Automatic detection of rib borders in chest radiographs. *IEEE Transactions on Medical Imaging*, 14(3):525–536, 1995. 2