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

Wei Dai(✉), Nanqing Dong, Zeya Wang, Xiaodan Liang, Hao Zhang, and Eric P. Xing

Petuum Inc., Pittsburgh, USA
{wei.dai,nanqing.dong,zeya.wang,xiaodan.liang,
hao.zhang,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. These voluminous CXR scans place significant workloads on radiologists and medical practitioners. Organ segmentation is a key step towards 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 inherent in human physiology. Specifically, the critic network learns the higher order structures in the masks in order to discriminate between the ground truth organ annotations from the masks synthesized by the segmentation network. Through an adversarial process, the critic network guides the segmentation network to achieve more realistic segmentation that mimics the ground truth. Extensive evaluation shows that our method produces highly accurate and realistic segmentation. Using only very limited training data available, our model reaches human-level performance without relying on any pre-trained model. Our method surpasses the current state-of-the-art and generalizes well to CXR images from different patient populations and disease profiles.

Keywords: Chest X-ray · Medical image segmentation
Adversarial learning · Deep neural networks

1 Introduction

Chest X-ray (CXR) is one of the most common medical imaging procedures. Due to CXR's low cost and low dose of radiation, hundreds to thousands of

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-00889-5_30) contains supplementary material, which is available to authorized users.

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 263–273, 2018.
https://doi.org/10.1007/978-3-030-00889-5_30

CXRs are generated in a typical hospital daily, which create 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) [4]. Among X-ray images, 8 million are Chest X-rays, which translate to thousands of CXR readings per radiologist per year. The shortage of radiologists is well documented across the world [11,14]. It is therefore of paramount importance to develop computer-aided detection methods for CXRs to support clinical practitioners.

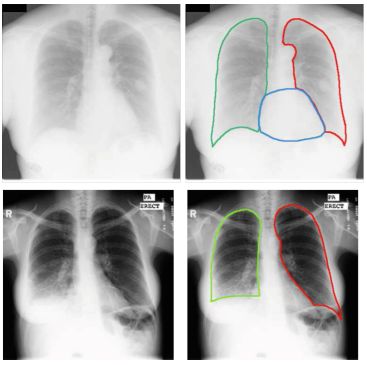

Fig. 1. Two example chest X-ray (CXR) images from two dataset: JSRT (top) and Montgomery (bottom). The left and right columns show the original CXR images and the lung field annotations by radiologists. JSRT (top) additionally has the heart annotation. Note that contrast can vary significantly between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

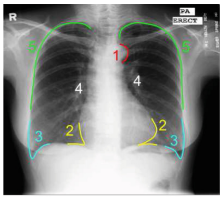

Fig. 2. Important contour landmarks around lung fields: aortic arch (1) is excluded from lung fields; costophrenic angles (3) and cardiodiaphragmatic angles (2) should be visible in healthy patients. Hila and other vascular structures (4) are part of the lung fields. The rib cage contour (5) should be clear in healthy lungs.

An important step in computer-aided detection on CXR images is organ segmentation. The segmentation of the lung fields and the heart provides rich structural information about shape irregularities and size measurements [3] 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 mask out non-lung regions to minimize the effect of imaging artifacts in computer-aided detection, which is important for the clinical use [13].

One major challenge in CXR segmentation is to incorporate the implicit medical knowledge involved in contour determination. For example, the heart and the lung contours should always be adjacent to each other due to definition of the lung boundaries (Sect. 2). 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 ambiguous boundaries 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 are very limited CXR data because of sensitive privacy issues. Even fewer training data have 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 to different patient populations, disease profiles, or image quality. Furthermore, these methods do not explicitly balance local information with global structure in a principled way, which is critical to achieving 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 organ segmentation in chest X-rays. By employing a convolutional network approach to organ segmentation, we side-step the problems faced by existing approaches based on ad hoc feature extraction. Our convolutional segmentation model alone can achieve performance competitive with existing methods. However, the segmentation model alone cannot capture sufficient global structures to produce natural contours due to the limited training data. To impose regularization based on the physiological structures, we introduce a critic network which learns the higher order structures in the masks in order to discriminate between the ground truth organ annotations from the masks synthesized by the segmentation network. Through an adversarial training process, the critic network effectively transfers this learned global information back to the segmentation network to achieve realistic segmentation outcomes that mimic the ground truth.

Without using any pre-trained models, SCAN produces highly realistic and accurate segmentation even when trained on a very small dataset. With the global structural information, our segmentation model is able to resolve difficult boundaries that require a strong prior knowledge. SCAN improves the state-of-the-art

lung segmentation methods [1,12,15] and outperforms strong baselines including U-net [9] and DeepLabV2 [2], achieving performance competitive with human experts. Furthermore, SCAN is more robust than existing methods when applied to different patient populations. To our knowledge, this is the first successful application of convolutional neural networks (CNN) to CXR image segmentation, and our CNN-based method can be readily integrated for clinical tasks such as automated cardiothoracic ratio computation [3]. We note that SCAN is similar to [8] in applying adversarial methods to segmentation. Further related work may be found in supplemental materials.

2 Structure Correcting Adversarial Network

We propose to use adversarial training for segmenting CXR images. Figure 3 shows the overall SCAN framework in incorporating the adversarial process into 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 Generative Adversarial Network (GAN) [5] but conditioned on an input image. On the other hand, the critic network takes the segmentation masks as input and outputs the probability that the input mask is the ground truth annotation instead of the prediction by the segmentation network.

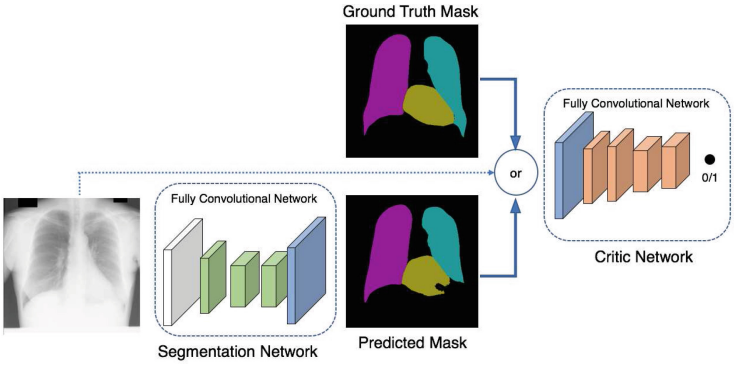

Fig. 3. Overview of the proposed SCAN framework that jointly trains a segmentation network and a critic network through an adversarial process. The segmentation network produces a mask prediction. The critic takes either the ground truth mask or the predicted mask and outputs the probability estimate of whether the input is the ground truth (with training target 1) or predicted mask (with training target 0).

The higher order consistency enforced by the critic is particularly desirable for CXR segmentation. Human anatomy, though exhibiting substantial variations across individuals, generally maintains a 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.

Training Objectives. The networks can be trained jointly through a minimax scheme that alternates between optimizing the segmentation network and the critic network. 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 sum 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 $J_s(\hat{\boldsymbol{y}}, \boldsymbol{y}) := \frac{1}{HW} \sum_{j,k} \sum_{c=1}^{C} -y^{jkc} \ln \hat{y}^{jkc}$ 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 alternating between optimizing S and optimizing D with corresponding loss function. See supplemental materials for details.

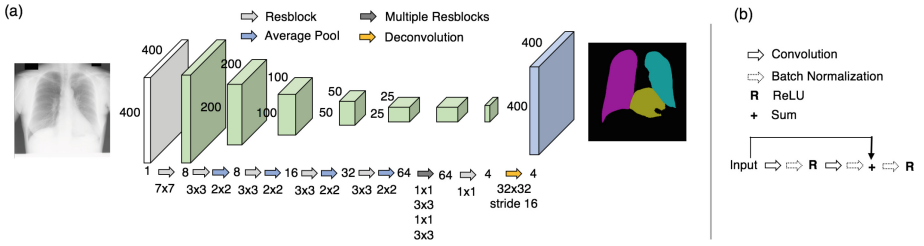

Fig. 4. The segmentation network architecture. (a) Fully convolutional network for dense prediction. (b) The residual block architecture is based on [6]. Further details are in supplementary materials.

Network Architectures. The segmentation network is a fully convolutional network (FCN) [2,7]. Figure 4 details our FCN architecture. The segmentation network contains 271k parameters, 500x smaller than VGG-based FCN [7]. Our FCN is highly parsimonious to adpat to the stringent dataset size of the medical domain: our training dataset of 247 CXR images is orders of magnitude smaller

than the dataset in the natural image domains. Furthermore, CXR is gray-scale with consistent viewpoint, which can be captured by fewer feature maps and thus fewer parameters. The parsimonious network construction allows us to optimize it efficiently without relying on any existing trained model, which is not readily available for the medical domain. Figure 5 shows the critic architecture, which has 258k parameters.

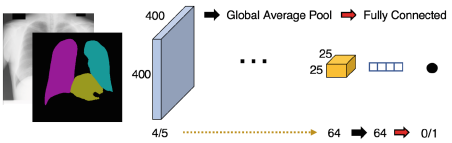

Fig. 5. The critic network architecture. Our critic FCN mirrors the segmentation network (Fig. 4). The training target is 0 for synthetic masks; 1 otherwise. Further details are in supplementary materials.

3 Experiments

We perform extensive evaluation of the proposed SCAN framework and demonstrate that our approach produces highly accurate and realistic segmentation of CXR images.

Dataset and Protocols. We use the following two publicly available datasets to evaluate our proposed SCAN framework. The datasets come from two different countries with different lung diseases, representing diverse CXR samples. **JSRT.** The dataset contains 247 CXRs, among which 154 have lung nodules and 93 have no lung nodule [10,12] (Fig. 1). **Montgomery.** The Montgomery dataset, collected in Montgomery County, Maryland, USA, consists of 138 CXRs, including 80 normal patients and 58 patients with manifested tuberculosis (TB) [1]. The CXR images are 12-bit gray-scale images of dimension 4020×4892 or 4892×4020. Only the lung masks annotations are available (Fig. 1). We scale all images to 400×400 pixels, which retains visual details for vascular structures in the lung fields and the boundaries. The evaluation metrics are Intersection-over-Union (IoU) and Dice Coefficient. We present the details of data processing and evaluation metrics in Supplementary Materials.

Quantitative Comparisons. We randomly split the JSRT dataset into the development set (209 images) and the evaluation set (38 images). We tune our architecture and hyperparameter λ (Eq. (1)) using a validation set within the development set and fix $\lambda = 0.01$. We use FCN to denote the segmentation network only architecture, and SCAN to denote the full framework with the critic.

We investigate how SCAN improves upon FCN. Table 1 shows the IoU and Dice scores using JSRT dataset. We observe that the adversarial training significantly improves the performance. In particular, IoU for the two lungs improves from 92.9% to 94.7%.

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. ± represents one standard deviation estimated from bootstrap.

		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%

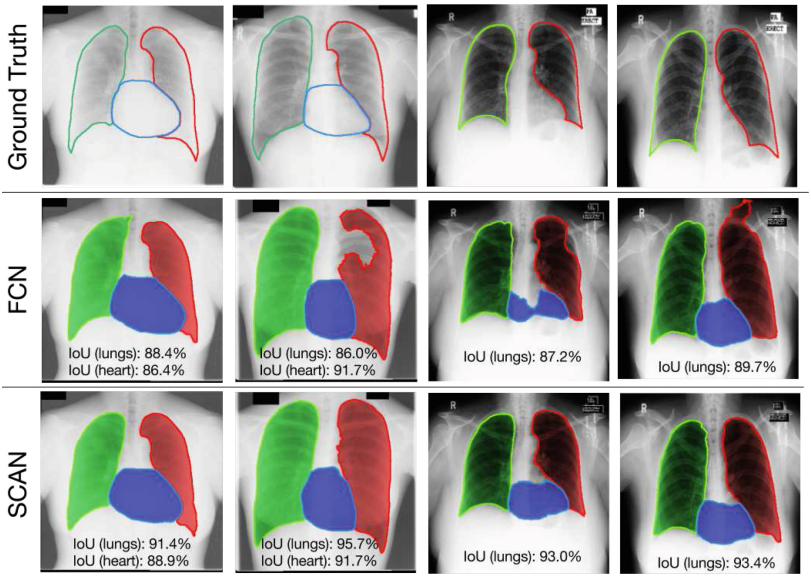

Fig. 6. Visualization of segmentation results on 4 patients, one per column. The left two columns are patients from the JSRT evaluation set with models trained on JSRT development set. The right two columns are from the Montgomery dataset using a model trained on the full JSRT dataset but not Montgomery, 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. The contours of the predicted masks are added for visual clarity.

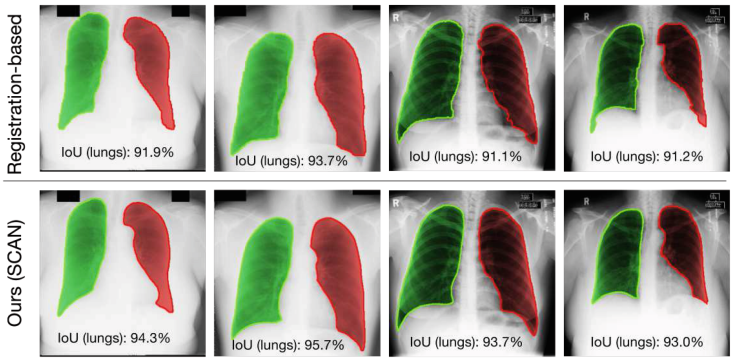

Fig. 7. Comparison with the current state-of-the-art [1]. SCAN produces sharp contours at the costophrenic angles for the left two columns (from the JSRT evaluation set). Furthermore, our model generalizes well to different patient populations and imaging setup, as shown in the Montgomery CXR in the right two columns. [1] struggles on Montgomery data due to the mismatch between train and test patient lung profiles (JSRT and Montgomery dataset, respective).

Table 2 compares 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 based on registration-based model [1] by a significant margin. Additionally, we compare with other standard CNN approaches for semantic segmentation: DeepLabV2 with ResNet101 [2] and U-Net [9] and demonstrate the advantage of our parsimonious architecture and adversarial training. Importantly, our method is competitive with the human performance for both lung fields and the heart.

For clinical deployment, it is important for the segmentation model to generalize to a different population with different patient population and image qualities, such as when deployed in another country or a specialty hospital with very different disease distributions. We therefore train our model on the full JSRT dataset, which is collected in Japan from a population with lung nodules, and test the trained model on the full Montgomery dataset, which is collected in the U.S. from patients potentially with TB. The two datasets present very different contrast and diseases (Fig. 1). Table 3 shows that FCN alone does not generalize well to a new dataset, but SCAN substantially improves the performance, surpassing [1].

We further investigate the scenario when training on the two development sets from JSRT and Montgomery *combined* to increase variation in the training data. 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.

Table 2. Comparison with existing single-model approaches to lung field segmentation on JSRT dataset. Note that [12,15] use different data splits than our evaluation.

	IoU (Lungs)	IoU (Heart)
Human Observer [12]	**94.6**% ± 1.8%	**87.8**% ± 5.4%
Ours (SCAN)	**94.7**% ± 0.4%	**86.6**% ± 1.2%
Registration-based [1]	92.5% ± 0.4%	–
DeepLabV2 101 [2]	85.7% ± 0.9%	–
U-net [9]	84.4% ± 1.3%	–
ShRAC [15]	90.7% ± 3.3%	–
ASM [12]	90.3% ± 5.7%	79.3% ± 11.9%
AAM [12]	84.7% ± 9.5%	77.5% ± 13.5%
Mean Shape [12]	71.3% ± 7.5%	64.3% ± 14.7%

Qualitative Comparison. Figure 6 shows the qualitative results from these two experiments. The failure cases in the middle row by our FCN reveal 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 a 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 the exterior boundary, causing prediction errors in the middle two columns for FCN. SCAN significantly improves all of the failure cases and produces much more realistic outlines of the organs. SCAN also sharpens the segmentation of costophrenic angle (the sharp angle at the junction of ribcage and diaphragm), which are important in diagnosing pleural effusion and lung hyperexpansion, among others.

Figure 7 compares SCAN with the current state-of-the-art [1] qualitatively. We restrict the comparison to lung fields, as [1] only supports lung field segmentation. SCAN generates more accurate lung masks especially around costophrenic angles when tested on the same patient population (left two columns of Fig. 7). SCAN also generalizes better to a different population in the Montgomery dataset (right two columns of Fig. 7) whereas [1] struggles with domain shift.

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 [1] on a laptop with Intel Core i5. [1] 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 [14] a fast test time result is highly desirable.

Table 3. Performance on the full Montgomery dataset using models trained on the full JSRT dataset. Compared with the JSRT dataset, the Montgomery dataset exhibits a much higher degree of lung abnormalities and varying imaging quality, testing the transferrability of the models.

	IoU (Both Lungs)
Ours (SCAN)	**91.4**% ± 0.6%
Ours (FCN)	87.1% ± 0.8%
Registration [1]	90.3% ± 0.5%

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.

	Test time
Ours (SCAN)	0.84 s
Registration [1]	26 s
Human	∼2 min

References

1. Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE Trans. Med. Imaging **33**(2), 577–590 (2014)
2. Chen, L.C., et al.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE TPAMI **40**, 834–848 (2018)
3. Dallal, A.H., et al.: Automatic estimation of heart boundaries and cardiothoracic ratio from chest x-ray images. In: In: SPIE Medical Imaging (2017)
4. NHS England: Diagnostic imaging dataset annual statistical release 2015/16 (2016)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
7. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
8. Luc, P., Couprie, C., Chintala, S., Verbeek, J.: Semantic segmentation using adversarial networks. In: NIPS Workshop on Adversarial Training (2016)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
10. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. **174**(1), 71–74 (2000)

11. The Royal College of Radiologists: Clinical radiology UK workforce census 2015 report, September 2016
12. Van Ginneken, B., et al.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal. **10**, 19–40 (2006)
13. Wang, X., et al.: Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. arXiv:1705.02315 (2017)
14. World Health Organization: Computer-aided Detection for Tuberculosis (2012)
15. Yu, T., Luo, J., Ahuja, N.: Shape regularized active contour using iterative global search and local optimization. In: CVPR. IEEE (2005)