Simultaneous Lung Field Detection and Segmentation for Pediatric Chest Radiographs

Wei Zhang[1], Guanbin Li[1], Fuyu Wang[1], Longjiang E[2], Yizhou Yu[3], Liang Lin[1], and Huiying Liang[⋆2]

[1] Sun Yat-sen University, Guangzhou, Guangdong, China
[2] Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou, Guangdong, China.
[3] Deepwise AI Lab, Beijing, China

Abstract. Accurate lung field segmentation (LFS) method is highly demanded in computer-aid diagnosis (CAD) system. However, LFS in pediatric CXR images has received few attention due to the lack of publicly available dataset and the challenges caused by their unique characteristics, such as great variations of the size, location and orientation of lungs. To fill this gap, this paper for the first time presents a simultaneous lung field detection and segmentation framework for pediatric CXR images. Our framework, called SDSLung Net, is a multi-tasking convolutional neural network architecture tailor-made for X-ray images with relatively weak appearance feature but abundant spatial rules and structural information. It is adapted from a Mask R-CNN framework [1] by incorporating a newly designed Organ Structure-Aware Encoding layer in the backbone network for more accurate spatial variation and structural representation, in parallel with a deeply supervised fully convolutional network based segmentation branch for precise lung field segmentation inside detected bounding box. Moreover, we also constructed a new and so far the largest pediatric CXR dataset with pixelwise lung field annotations. Experimental results demonstrate that our proposed SDSLung is capable of achieving significantly superior performance over state-of-the-art LFS methods on our large-scale pediatric CXR dataset and also achieving extremely competitive results on adults' CXR dataset.

Keywords: pediatric CXR images · lung field segmentation · segmentation · detection.

1 Introduction

With the potential to reduce the workloads of radiologists and facilite appropriate treatments, Computer-Aid Diagnosis (CAD) for Chest X-ray (CXR) images has long been desired, especially in pediatrics where the increasing shortage of radiologists has been widely noticed. In CAD systems, lung field segmentation (LFS) which precisely recognizes lung fields acts as a crucial role for intelligent diagnosis, e.g. pediatric atelectasis analysis. Unfortunately, the exist-

⋆ Corresponding author.

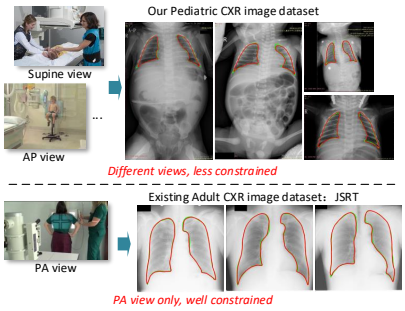

Fig. 1: An illustration of the difference between our pediatric CXR image dataset and existing adults' CXR image dataset. The lung field segmentation results of our framework and the ground-truth are shown in red and green curves, respectively. As can be observed, our framework produces high quality lung field segmentation results on both datasets.

ing LFS study did not pay sufficint attetion to pediatric CXR images. Besides the inherent difficulties in conventional CXR images, pediatric CXR images also features significant variations in terms of sizes, positions and orientations of lung fields, making the LFS significantly more challenging than in conventional adults' CXR images. Examples are shown in Figure 1. Although numbers of LFS methods have been proposed based on publicly avaliable JSRT dataset [2] and MC dataset [3], they are limited in handling pediatric CXR images due to the lack of precise lung field positioning. Moreover, existing LFS datasets only contain adults' CXR images which are filmed under standard posterior-antero (PA) view such that accurate lung field localization is less discussed in existing LFS studies. Therefore, to deal with aforementioned difficulties of LFS in pediatric CXR images, novel LFS methods that can simultaneously perform lung field detection and mask segmentation inside detected regions are particularly demanded.

To fill this gap, in this paper, we present a novel simultaneous lung field detection and segmentation framework for pediatric CXR images. Simultaneous object detection and segmentation (SDS) [4], also known as instance level object segmentation [1, 5, 6], is an important research topic in the field of computer vision, which aims at accurately detecting and segmenting generic objects of some specific categories in natural scenes. Although the goal of LFS appears to be consistent with SDS, borrowing the techniques of SDS to LFS is non-trivial due to the following reasons: first, compared with SDS models developed for natural scenes, SDS models for CXR images is expected to be simple in classification part (only left and right lung are considered) but demand more precise segmentation of lung fields to serve as important cues for diagnosis. Second, there exist inherent differences between natural images and CXR images – natural images are rich in color and texture information while CXR images are in gray-scale with lower contrast and fewer textures. Most importantly, SDS models developed for natural images solely rely on Deep Convolutional Neural Networks (DCNNs) to extract feature representations, in which the anatomical structure of lung fields in CXR

images is not well utilized. Inspired by Mask R-CNN [1], an elegant end-to-end neural network framework for object instance segmentation, we propose to adapt it for lung field segmentation in CXR images by addressing the aforementioned issues. We conduct comprehensive experiments on a newly constructed and so-far the largest pediatric LFS dataset, demonstrating state-of-the-art performance. Moreover, our proposed framework achieves extremely competitive results on adults' CXR dataset.

Pediatric CXR Dataset We collaborate with one Women and Children's Hospital and collect 733 pediatric CXR images from their daily clinical practice. According to pediatric clinical requirements, the collected CXR images are filmed under different views, such as supine antero-posterior (AP) and posterior-antero (PA). The collected CXR images are densely sampled from the age range of 1 day to over 10 years old, in which the appearance of lungs manifests more variations compared with other age ranges. We invite experienced pediatric radiologists to manually segment the lung field for each CXR image.

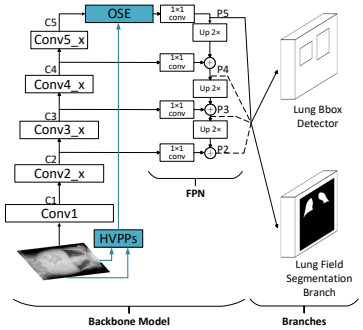

Fig. 2: The overall framework of our proposed *SDSLung Net*.

Table 1: A comparison between our newly collected Pediatric CXR Dataset and existing publicly available LFS datasets. Our dataset is significantly larger than existing datasets.

Dataset	# CXR	View	Age	gt Mask
JSRT	247	PA	4 ∼ 89 years old	✓
MC	138	PA	16 ∼ 84 years old	✓
Ours	**733**	**PA,Supine,etc.**	1 day ∼ 14 years old	✓

A comparison between our newly collected Pediatric CXR Dataset and existing publicly available LFS datasets is shown in Table 1.

2 SDSLung Net

As illustrated in Figure 2, the framework of *SDSLung Net*, which is adapted from Mask-RCNN, contains two branches: one bounding box detector that localizes two lungs in CXR images and a segmentation branch that aims at precisely masking out the lung region inside each detected lung box. Both of the two branches share the same backbone model, in which a newly proposed *Organ Structure-Aware Encoding* (OSE) layer is integrated with vanilla ResNet [7]. Inspired by the common practice in modern object detection and instance segmentation frameworks [1, 8], we add a top-down path with lateral connections in the form of the Feature Pyramid Network (FPN) [8], connected to the top of ResNet-OSE, to take full account of multi-scale deep features (P5 to P2) while performing detection and segmentation inference.

Organ Structure-Aware Encoding Based on the observation of the horizontal and vertical regular structural relationships between the lung fields and the surrounding torso region, we introduce scanning operations in OSE layer. As shown in Figure 3 (a), *(i)* there exists certain column-wise and row-wise regularity in the gray-scale patterns across lung fields; *(ii)* the ups and downs in horizontal and vertical projections contain useful cues for distinguishing body and background as well as the localization of lung fields. We rely on Recurrent Neural Networks (RNN) to capture anatomical structure around lung fields across the torso region, since fully convolutional layers are not adept at modeling such long-range dependencies. Specifically, we adopt ReNet [9, 10] which is composed of 4 Gated Recurrent Units (GRUs) that perform horizontal and vertical sweeps respectively. The architecture of OSE layer is depicted in Figure 4. Next, we further introduce global image information in the scanning process to enhance the feature representation of the backbone network. Traditionally, horizontal and vertical projection profiles (HVPPs) [11] are wildly adopted to probing the distribution of patterns in simple gray-scale images, such as text lines in handwriting samples and lung fields in high-contrast PA CXR images. We propose to incorporate HVPPs into the scanning process to serve as image-level cues. Formally, the horizontal projection profile l^h and vertical projection profile l^v are defined as:

$$l_j^h = \frac{1}{H}\sum_{i=1}^{H} s_{i,j}, \quad l_i^v = \frac{1}{W}\sum_{j=1}^{W} s_{i,j} \tag{1}$$

Where H, W denotes the height, width of the input gray-scale CXR image. $s_{i,j}$ denotes the gray-scale value of pixel located at i, j. One example of HVPPs is shown in Figure 3. To make the projection profiles compact, l^h and l^v are downsampled to an uniform dimension of 128×1. We subsequently make use of the downsampled HVPP vectors by feeding them to GRUs.

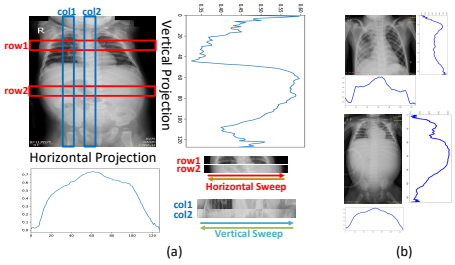

Fig. 3: (a) is an illustration of our observations regarding the anatomical structure of torso in CXR images. (b) provides two more examples of the horizontal and vertical projections.

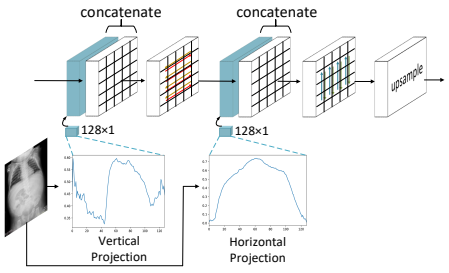

Fig. 4: The architecture of OSE layer.

Each OSE layer takes as input the raw input image or the feature activation of one intermediate layer. Formally, we denote the input tensor by $X = x_{i,j} \in$

$\mathbb{R}^{H^{'} \times W^{'} \times C}$ where $H^{'}$, $W^{'}$ and C denotes the height, width and the number of channels, respectively. In order to perform the sweeps in orthogonal directions efficiently, we divide the input tensor X into a grid, which consists of $I \times J$ non-overlapping patches $P = \{p_{i,j}\}$, where each patch $p_{i,j} \in \mathbb{R}^{\in H_p \times W_p \times C}$. We first sweep horizontally from left to right and right to left with two GRUs $f^{\rightarrow}$ and $f^{\leftarrow}$. For the sweeps working alone row j in the grid, every time step each GRUs takes as input the next non-overlapping patch $p_{i,j}$ as well as the downsampled vertical projection profile l^v, producing horizontally encoded feature $o^{\rightarrow}_{i,j}$ (or $o^{\leftarrow}_{i,j}$) and updates its hidden states $z^{\rightarrow}_{i-1,j}$ (or $z^{\leftarrow}_{i+1,j}$):

$$\begin{aligned} o^{\rightarrow}_{i,j} &= f^{\rightarrow}(z^{\rightarrow}_{i-1,j}, \{p_{i,j}, l^v\}), \quad for\ i = 1, ...I \\ o^{\leftarrow}_{i,j} &= f^{\leftarrow}(z^{\leftarrow}_{i+1,j}, \{p_{i,j}, l^v\}), \quad for\ i = I, ...1 \end{aligned} \tag{2}$$

Once the two GRUs have processed all rows in the grid of X, the horizontally encoded features $o^{\rightarrow}_{i,j}$ and $o^{\leftarrow}_{i,j}$ are concatenated to obtain a horizontal organ structure-aware feature representation $O^{\leftrightarrow}$, where $O^{\leftrightarrow} \in \mathbb{R}^{\in H_p \times W_p \times 2U}$. U denotes the dimension of encoded feature $o^{\rightarrow}_{i,j}$ (or $o^{\leftarrow}_{i,j}$) emitted by one directional GRUs at each step. Similarly, column-wise bidirectional sweeps are then performed vertically on $O^{\leftrightarrow}$ as well as the downsampled horizontal projection profile l^h using another pair of GRUs $f^{\downarrow}$ and $f^{\uparrow}$. We concatenate $o^{\downarrow}_{i,j}$ and $o^{\uparrow}_{i,j}$ to obtain a composite feature map. Finally, we upsample the obtained feature representation to the same size as the input of OSE layer, allowing the OSE layer can be integrated at any stage of CNN architecture. We found that with one OSE layer inserted between the encoding and decoding stages, the whole pipeline reaches the best performance. The framework of our proposed *SDSLung Net* is illustrated in Figure 2.

Lung Field Detector To reduce the redundancy in CNN architecture, we cut out the classification part in Mask-rcnn which is originally designed for multiple object categories in natural scenes. Moveover, instead of using the anchor boxes in $[1, 12]$ with hand-picked aspect ratios, we consider the prior knowledge of aspect ratio of lungs to generate better anchor boxes for lung field detection. Specifically, we run k-means clustering on the training set lung field bounding boxes with the following distance metric:

$$d(box, centroid) = 1 - IOU(box, centroid). \tag{3}$$

Where the *box* and *centroid* denote the bounding box and current clustering centroid box, respectively.

Segmentation With Spatial Refinement As the boundary precision is one of the key requirements in LFS, we propose a two-step architecture in lung field segmentation branch. In the first step, we obtain initial mask predictions using fully convolutional layers that work on low-resolution and fixed-size feature maps, then in the second step a multi-scale refinement sub-network is adopted to obtain high resolution segmentation results. In this refinement sub-network, we cast the architecture of HED [13] which features multi-scale representation and deep supervisions for spatial refinement. Besides, to include context information,

we extend the size of each detected bounding box by a factor of 1.5. Then, the region inside the extended bounding box is cropped from the input image and concatenated with the initial mask prediction to form the input of the refinement sub-network. We apply deep supervisions on both of the initial mask and refined result.

Multi-task Loss During the training phase, we define a multi-task loss that account for the lung detection and segmentation branches.

$$L = L_{obj} + L_{reg} + L_{seg}. \tag{4}$$

Following the objective functions in Faster-RCNN [12], we define an objectness loss L_{obj} over two classes (lung fields or not).

$$L_{obj} = -(\frac{1}{N_P}\sum_{i\in P} log(b_i) + \frac{1}{N_N}\sum_{j\in N} log(1-b_j)), \tag{5}$$

where N_p and N_N denote the number of positive and negative anchors respectively with a ratio of 1:3. The regression loss accounts for predicting the offsets of lung field bounding boxes and is activated only for positive anchors $b_i^* = 1$:

$$L_{reg} = \frac{1}{N_{reg}}\sum_i b_i^* R(t_i - t_i^*), \tag{6}$$

Where $t_i = (t_x^i, t_y^i, t_w^i, t_h^i)$ is the vector of offsets for bounding box i. R is the robust loss function (smooth L_1) defined in [14].

$$L_{seg} = -\frac{1}{N_{seg}}\sum_{s\in 1,2}(\sum_i (w_i log p_i^s + (1-w_i) log(1-p_i^s))),$$

We adopt a cross-entropy loss for the lung field segmentation branch, in which p_i^s is the predicted probability of pixel i belonging to lung field in the initial mask prediction or in the output of the refinement sub-network. The binary value $w_i = 1$ indicates that the groundtruth of pixel i belongs to lung field, otherwise $w_i = 0$. N_{seg} is the total number of pixels inside the detected bounding box.

3 Experiments

Implementation Details We randomly select 489 CXR images from aforementioned pediatric CXR dataset as the training data and use the remaining 244 CXR images for testing. Our implementation is based on Tensorflow and each experiment runs on a Nvidia GTX1080 GPU. The input of refinement sub-network are scaled to $512 \times 320 \times 4$, in which the first 3 channels represent the cropped image and the rest 1 channel being the initial mask prediction. The architecture of the refinement sub-network is the same as in [13] except the input layer is modified to tolerate 4 channel's input. During the inference phase, we run the lung field detection branch to obtain 1000 box proposals. Subsequently, the lung

field segmentation branch is performed on the 100 highest scoring boxes.Our training is conducted for 120 epochs in total with a learning rate of 0.001. We set the weight decay to 0.0001 and the momentum to 0.9.

Evaluation Metrics We conduct quantitative evaluations using three commonly used metrics in previous LFS studies: *The Jaccard Similarity Coefficient*(JSC), *Dice's Coefficient*(DC) and *Average Contour Distance*(ACD). The detailed definitions of these metrics can be found in [15].

Table 2: Comparison with state-of-the-art LFS methods on our Pediatric CXR Dataset.

Method	JSC	DC
[15]	0.677±0.171	0.794±0.137
[16]	0.864±0.092	0.924±0.063
SDSLung Net	**0.901±0.062**	**0.947±0.038**

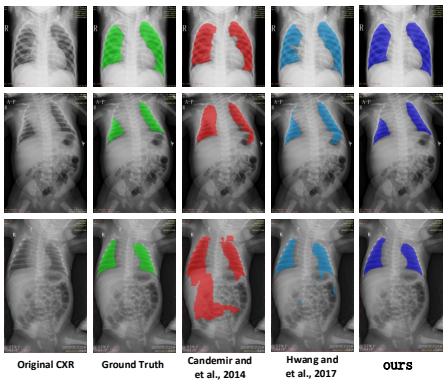

Fig. 5: A visual comparison between results of the proposed *SDSLung Net* and state-of-the-art LFS methods. Note our proposed SDS framework is also able to predict the bounding box of each lung field. Here, we only show the segmented mask for a better visualization.

Table 3: Comparison with state-of-the-art LFS methods on JSRT Dataset. Without the need of cumbersome network-wise training [16] *stage*3, the performance of the proposed *SDSLung Net* is almost equal to the state-of-the-art.

Method	JSC	DC	ACD
[17]	0.949 ± 0.020	-	1.62 ± 0.66
[15]	0.945 ± 0.015	0.967 ± 0.08	1.321 ± 0.316
[18]	0.947 ± 0.004	-	-
[19]	0.950	0.973	-
[16] *stage*1	0.950 ± 0.023	0.974 ± 0.012	1.347 ± 0.919
[16] *stage*3	**0.961 ± 0.015**	**0.980 ± 0.008**	**1.237 ± 0.702**
SDSLung Net	**0.961± 0.019**	**0.980±0.010**	**1.261±0.242**

Table 4: Performances of the proposed *SDSLung Net* by choosing different architectures for the backbone model. Note in all kinds of above settings, FPN is adopted while the refinement sub-network is not included. Note ResNet101 here denotes exactly the same backbone as in Mask-RCNN.

Backbone model	JSC	DC
ResNet101(Mask-rcnn)	0.856±0.020	0.923±0.002
ResNet18	0.868±0.015	0.905±0.030
ResNet18+OSE (w/o HVPPs)	0.880±0.007	0.936±0.014
ResNet18+OSE (w/HVPPs)	**0.884±0.061**	**0.937±0.037**

Comparisons against leading methods We compare the proposed *SDSLung Net* against two state-of-the-art LFS methods [15,16] on our newly collected pediatric CXR dataset. The quantitative comparison results are listed in Table 2. With the ability of simultaneous detection and segmentation, our method significantly outperform previous state-of-the-arts on the challenging pediatric CXR dataset. Besides, to test the adaptability of the proposed *SDSLung Net* to adults' CXR images, we also compare against state-of-the-art LFS methods on JSRT dataset. Results are listed in Table 3. As can be seen, our results significantly

outperforms most of the leading methods. Moreover, we note *SDSLung Net* is able to achieve almost equal performance with [16] *stage*3 which is obtained by repeating 3 stages of the network-wise training. Next, we provide a visual comparison in Figure 5. As can be seen, our framework is able to predict satisfying results for difficult pediatric CXR images, surpassing previous state-of-the-art LFS methods.

The Effectiveness of Organ Structure-Aware Encoding From the results listed in Table 4, we can observe 1.2% improvement in JSC gained by the RNN-based organ structure-aware encoding. By introducing HVPPs, another 0.4% improvement in JSC can be obtained.

4 Conclusions

In this paper, we present *SDSLung Net*, which is the first simultaneous lung field detection and segmentation framework for CXR images. We also introduce a novel Organ Structure-Aware Encoding layer as well as a multi-scale refinement sub-network, which allow the proposed *SDSLung Net* to achieve the state-of-the-art performance on both pediatric and adults' CXR dataset.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
2. 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. American Journal of Roentgenology **174**(1), 71–74 (2000)
3. Jaeger, S., et al.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quantitative imaging in medicine and surgery **4**(6), 475 (2014)
4. Hariharan, B., et al.: Simultaneous detection and segmentation. In: ECCV. pp. 297–312. Springer (2014)
5. Pinheiro, P.O., Dollár, P.: Learning to segment object candidates. In: NIPS. pp. 1990–1998 (2015)
6. Dai, J., et al.: Instance-sensitive fully convolutional networks. In: ECCV. pp. 534–549. Springer (2016)
7. He, K., et al.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
8. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
9. Visin, F., et al.: Renet: A recurrent neural network based alternative to convolutional networks. arXiv preprint arXiv:1505.00393 (2015)
10. Visin, F., et al.: Reseg: A recurrent neural network-based model for semantic segmentation. In: CVPR Workshops. pp. 41–48 (2016)
11. Semmlow, J.L., et al.: Biosignal and medical image processing. CRC press (2014)

12. Ren, S., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015)
13. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV. pp. 1395–1403 (2015)
14. Girshick, R.: Fast R-CNN. In: ICCV. pp. 1440–1448 (2015)
15. Candemir, S., et al.: Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. IEEE transactions on medical imaging **33**(2), 577–590 (2014)
16. Hwang, S., et al.: Accurate lung segmentation via network-wise training of convolutional networks. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 92–99. Springer (2017)
17. Van Ginneken, B., et al.: 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)
18. Dai, W., et al.: Scan: Structure correcting adversarial network for chest x-rays organ segmentation. arXiv preprint arXiv:1703.08770 (2017)
19. Novikov, A.A., Lenis, D., Major, D., Hladvka, J., Wimmer, M., Bühler, K.: Fully convolutional architectures for multiclass segmentation in chest radiographs. IEEE transactions on medical imaging **37**(8), 1865–1876 (2018)