LOCATE-THEN-DELINEATE: A FREE-TEXT REPORT GUIDED APPROACH FOR PNEUMOTHORAX SEGMENTATION IN CHEST RADIOGRAPHS

Samruddhi Shastri [†], *Naren Akash R J* [†], *Lokesh Gautham, Jayanthi Sivaswamy*

Center for Visual Information Technology,
International Institute of Information Technology Hyderabad, India
`https://cvit.iiit.ac.in/mip/projects/ptxseg`

ABSTRACT

We present a novel solution for accurate segmentation of pneumothorax from chest radiographs utilizing free-text radiology reports. Our solution employs text-guided attention to leverage the findings in the report to initially produce a low-dimensional region-localization map. These prior region maps are integrated at multiple scales in an encoder-decoder segmentation framework via dynamic affine feature map transform (DAFT). Extensive experiments on a public dataset CANDID-PTX, show that the integration of free-text reports significantly reduces the false positive predictions, while the DAFT-based fusion of localization maps improves the positive cases. In terms of DSC, our proposed approach achieves 0.60 and 0.95 for positive and negative cases, respectively, and 0.70 to 0.85 for medium and large pneumothoraces.

1. INTRODUCTION

Pneumothorax is the presence of air between the parietal and visceral pleura. Such abnormal accumulation can progressively raise the intrapleural pressure and collapse the lung, shift the mediastinum and impair venous return to the heart, resulting in fatal cardiopulmonary failure. Hence, rapid detection and immediate intervention are critical.

Pneumothorax diagnosis relies on detecting a fine, sharply defined opaque line representing the displaced visceral pleura in the chest radiograph. This is generally very subtle, and highly variable in shape as the pleura tends to overlap with the ribs or clavicle, leading to frequent under-detection. Other conditions such as emphysematous bullae, skinfolds, folded clothing on the patient, and overlap of the stomach mimic pneumothorax. The size of pneumothorax is an important determinant of treatment and necessitates accurate segmentation of the collapsed lung region [1].

Most of the existing approaches to automate pneumothorax segmentation rely solely on imaging data. A multi-scale convolutional network [2] and simple U-Net with various backbones [3] have been explored along with utilising weak supervision to address the prohibitive cost of collecting sufficient high-quality datasets with markings for training deep neural networks [4]. More recently, a multimodal approach to the problem is being considered by the community. In particular, radiology reports which are complementary to image data have been shown to provide easy-to-access textual descriptions for each finding in detail, including its location, size, shape, density, signal characteristics, or other pertinent features. Effective fusion of heterogeneous data to align the different modalities is challenging. LViT [5] incorporated a vision transformer branch to process cross-modal information and merge image and text information, whereas cross-attention and cross-position attention were utilized in ContextualNet [6] and CPAM [7], respectively, to integrate textual information in the decoder. In contrast to LViT and CPAM, which employ carefully crafted structured reports, we aim to use free-text radiology reports because they are more the norm in a clinical setting.

Since a direct fusion of information from text and image may be suboptimal, due to the inherent differences in noise topologies and input representations from different modalities, we propose a two-stage approach. In the first stage, we utilize text-guided attention based on report findings to generate a low-dimensional map for region localization. Next, in the second stage, we incorporate prior region maps at multiple scales within the encoder-decoder segmentation framework through dynamic affine feature map transform (DAFT) [8]. DAFT predicts the scales and shifts, to excite and suppress image feature maps of a convolutional layer by conditioning them on both the image and the localization map.

To summarize, the key contributions of our work are:

- A novel, two-stage, multimodal localization-guided pneumothorax segmentation framework from chest radiographs and associated free-text radiology reports.
- Cross attention-based rough localization which leverages free-text radiology reports and DAFT-based fine segmentation based on rough localization information.

[†] These authors contributed equally to this work.

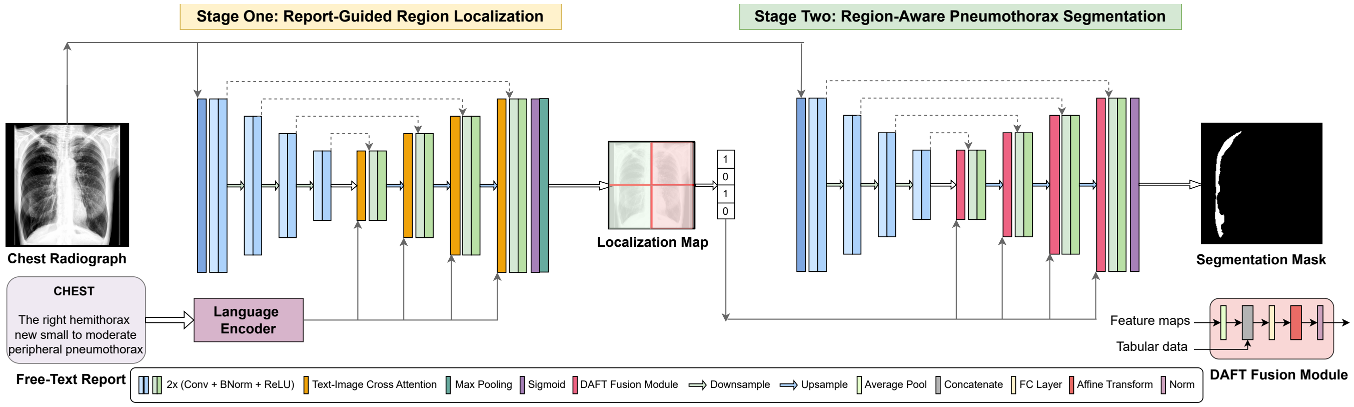


Fig. 1. A schematic illustration of the proposed approach. Given a chest radiograph I and its corresponding free-text radiology report R, we first obtain a region localization map L, leveraging text-guided attention. In the subsequent stage, we modulate the feature maps of the segmentation network at multiple scales using L to accurately segment pneumothorax.

2. METHOD

Interpreting chest radiographs is challenging due to the superimposition of anatomical structures along the projection direction leading to low-contrast features, which makes it difficult to distinguish between them. Consequently, an accurate segmentation of pneumothorax directly from radiographs is quite difficult. Radiological reports can be used to mitigate this challenge. For instance, a sample report text reads as: "*On the current film, there is a very small pneumothorax visible at the left apex. Both lungs are overinflated. There is no other abnormality.*" Thus, reports provide rich clues on the presence and possibly location and size of pneumothorax in addition to a description of the overall condition of the respiratory system. We aim to explicitly integrate findings in free-text with the image to roughly localize pneumothorax in the image before doing a fine segmentation and hence propose a two-stage approach: (i) report-guided region localization and (ii) region-aware pneumothorax segmentation. A schematic of the proposed pipeline is shown in Figure 1.

2.1. Report-Guided Region Localization

In the first stage, a 2D U-Net encoder-decoder architecture with cross-attention layers similar to ContextualNet [6] is used to fuse the input image and text. Specifically, a set of text features, $R_t \in \mathbb{R}^{N_R \times D_R}$, are extracted from the report by leveraging a language encoder. Here, N_R and D_R denote the number of channels and the dimension of the embeddings, respectively. Next, a fully connected layer is employed to project R_t into $R_d \in \mathbb{R}^{N_R \times D_d}$, where D_d denotes the depth of the feature map of the d^{th} block of the decoder. The text feature is fused with the upsampled decoder feature maps via cross-attention. The contextualized feature maps are concatenated with the encoder feature maps via skip connections. In order to derive a rough localization map at the decoder output, we take inspiration from the semi-quantitative visual assessment methods of chest radiographs [9] and divide the image into four quadrants with two per lung. The presence or absence of pneumothorax in each of the regions is inferred by incorporating a max-pooling layer on the final decoder feature map. The regional information is represented as a low-dimensional tabular localization map $\hat{L} \in \mathbb{R}^{N_L}$, where N_L denotes the number of regions.

2.2. Region-Aware Pneumothorax Segmentation

The goal of this stage is to finely segment the pneumothorax based on the image and quadrant-level information from the previous stage. A 2D U-Net is used for this stage. The rough localization information from the output of the first stage is fused with the feature maps of the segmentation branch at multiple scales via DAFT. DAFT layers effectively modulate high-level concepts by conditioning the feature maps on both the image and the weakly related tabular information. For each chest radiograph, let $F_d = \mathbb{R}^{D_d \times H_d \times W_d}$ be the input feature map of the d^{th} block of the decoder, where D_d, H_d and W_d denote the depth, height and width of the feature map, respectively. DAFT learns to predict the scale α_d and shift β_d: $\boldsymbol{F}^{'}_d = \alpha_d \boldsymbol{F}_d + \beta_d$; $\alpha_d = f(\boldsymbol{F}_d, \hat{L})$; and $\beta_d = g(\boldsymbol{F}_d, \hat{L})$, where f and g are arbitrary mappings from image and tabular space to a scalar. A single auxiliary fully connected network h models f, g and outputs a single α-β pair.

3. EXPERIMENTS AND RESULTS

3.1. Dataset and Experimental Settings

Dataset The CANDID-PTX dataset [10] was used for all our experiments. It includes 19,237 frontal chest radiographs from patients over 16 years of age and contains 3,561

Table 1. Comparison of pneumothorax (PTX) segmentation performance on the CANDID-PTX dataset. Fivefold average Dice similarity coefficient and standard deviation is listed for baseline and other state-of-the-art methods.

Methods	PTX-Positive	PTX-Negative	Small PTX	Medium PTX	Large PTX
U-Net	0.550 ± 0.019	0.261 ± 0.082	0.398 ± 0.015	0.635 ± 0.033	0.791 ± 0.046
LViT [5]	0.549 ± 0.010	0.545 ± 0.095	0.378 ± 0.020	0.635 ± 0.012	0.798 ± 0.033
CPAM [7]	0.507 ± 0.031	0.703 ± 0.201	0.343 ± 0.028	0.598 ± 0.038	0.751 ± 0.041
ContextualNet [6]	0.566 ± 0.008	**0.961** ± 0.012	0.403 ± 0.020	0.657 ± 0.018	0.806 ± 0.029
Proposed (Ours)	**0.601** ± 0.013	0.948 ± 0.012	**0.429** ± 0.024	**0.697** ± 0.011	**0.851** ± 0.017

Table 2. Segmentation performance in ablation studies. S-1 and S2 denote stage 1 and 2 networks, respectively. R denotes the free text report used in stage 1, whereas D-$\mathcal{B}$ and D-$\mathcal{D}$ refer to DAFT used in the bottleneck and decoder layers, respectively.

Variants	S-1	R	S-2	D-$\mathcal{B}$	D-$\mathcal{D}$	PTX-Positive	PTX-Negative
V1	✗	✓	✓	✓	✗	0.591 ± 0.011	0.185 ± 0.056
V2	✓	✗	✓	✓	✗	0.590 ± 0.010	0.313 ± 0.059
V3	✓	✓	✓	✓	✗	0.598 ± 0.009	0.946 ± 0.013
V4	✓	✓	✓	✓	✓	**0.601** ± 0.013	**0.948** ± 0.012

pneumothorax annotations on 3,196 pneumothorax-positive (positive images). The dataset consists of images with a 1:5 positive-to-negative case ratio and contains corresponding anonymized, free-text radiology reports.

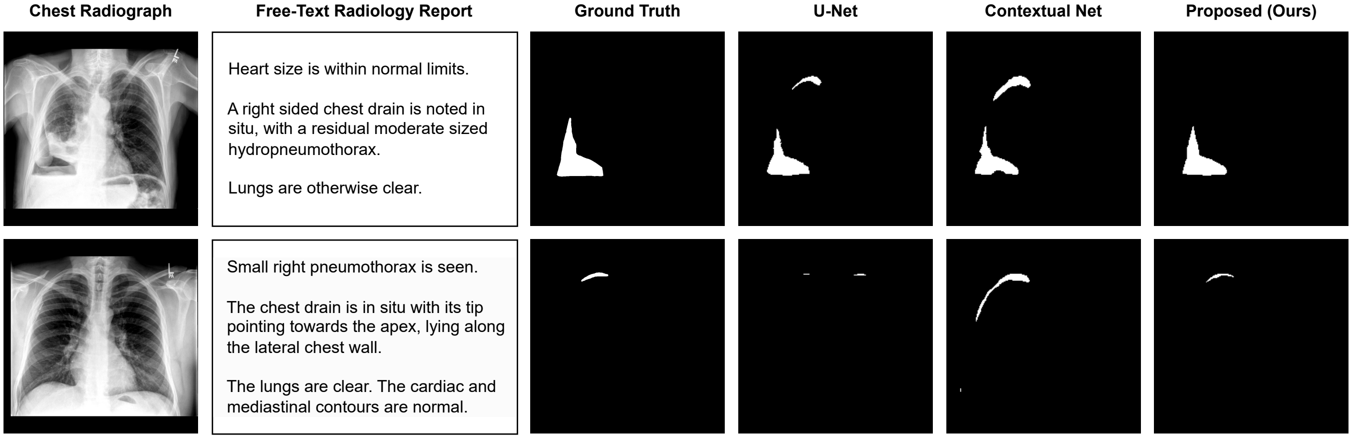
Implementation The chest radiographs were resized to 224×224. Our experimental setup involved a stratified (by size of pneumothorax) five-fold cross-validation. Each fold included a designated testing set, while the remaining data was split into 75% for training and 25% for validation. The training process utilized the AdamW optimizer with an initial learning rate and weight decay of $1\mathrm{e}{-}4$, implemented in PyTorch. The 2D U-Net with a pre-trained ResNet-50 backbone was utilized for segmentation. A frozen pre-trained T5-Large [11] model was used to extract language embeddings from free-text reports. The two stages were trained sequentially and employed the weighted combination of binary cross-entropy and Dice loss. The experiments were conducted on two NVIDIA GeForce RTX-2080 Ti GPUs. The training was limited to a maximum of 100 epochs with a batch size of 8.

Evaluation Metrics The performance of the segmentation methods was evaluated using the Dice similarity coefficient (DSC): DSC $= \frac{2\times|P_{\text{pred}}\cap P_{\text{gt}}|}{|P_{\text{pred}}|\cup|P_{\text{gt}}|}$, where P_{pred} and P_{gt} are the predicted segmentation mask and ground truth reference mask, respectively. The positive cases were subdivided into three classes based on the size of the pneumothorax in the image: small, medium and large, determined by thresholding. These thresholds were chosen based on the frequency histogram of the collapsed lung area on the chest radiograph. The performance was evaluated for each class.

3.2. Results and Discussion

Two sample images, text reports and corresponding outputs of our method, along with two other methods, are shown in Figure 2. The top row shows a large case of pneumothorax, while the bottom row shows a small case. Severe under-segmentation is notable with U-Net, while over-segmentation is seen with the ContextualNet for the small case. Both models suffer from over-segmentation for the large case.

We compare the segmentation performance of our proposed solution with baseline U-Net and the state-of-the-art (SOTA) medical vision-language frameworks, LViT [5], ContextualNet [6] and CPAM [7]. The mean DSC and the standard deviation across five folds are provided in Table 1. The U-Net, which is our baseline, considers only the chest radiograph, and yields a DSC of 0.550 and 0.261 for positive and negative cases, respectively. In positive cases, our proposed solution outperforms U-Net by 9.3%, LViT by 9.5%, ContextualNet by 6.2%, and CPAM by 18.5%. The superior performance is sustained across varying sizes of the pneumothorax, as seen from the best performance (shown in bold font) being achieved by our method for all sizes. Specifically, our two-stage approach yields a min/max boost of 7.6% to 9.8% over the U-Net; 6.6%, to 13.5% over LViT; 13.3% to 25.1% over CPAM and 5.6% to 6.5% over ContextualNet. While ContextualNet also uses free-text reports, it performs segmentation in one stage, similar to CPAM and LViT. The improvement in the performance of our method in positive cases over all these three methods is, we believe, due to the two-stage design. Since medium and large pneumothoraces are typically of interest to human experts, a boost in the segmentation of such cases can be of significant aid to them [1].



Heart size is within normal limits.

A right sided chest drain is noted in situ, with a residual moderate sized hydropneumothorax.

Lungs are otherwise clear.

Small right pneumothorax is seen.

The chest drain is in situ with its tip pointing towards the apex, lying along the lateral chest wall.

The lungs are clear. The cardiac and mediastinal contours are normal.

Fig. 2. Qualitative results of pneumothorax segmentation by different methods.

Ablation Studies In order to understand the contributions of different information and design choices, we performed ablation studies. The results are provided in Table 2. The best performance is seen to be achieved by variants V3 and V4, which employ free-text guided attention in stage 1 to generate a rough localization map. This is subsequently fused using DAFT at the bottleneck layer of Stage 2 in V3 and in the decoder layer in V4, which is seen to result in a marginal boost in performance. Ablation was done across two main dimensions: utilization of free-text reports and the two-stage design. In the following, all comparisons are done with respect to V4. First, we assess the importance of free-text radiology reports in the first stage. The variant V2 does not use text-guided attention in stage 1 and does rough localisation based only on the input radiograph. A large drop can be noted in performances in the negative cases. This implies that the information in the report serves to reduce false positive pneumothorax predictions. Second, we experimented with removing stage 1 to create V1. Here, the report is used to directly predict the rough localization map using an MLP. Once again, there is a significant drop in the performance in the negative cases. Hence, we conclude that these results underscore the benefits of a two-stage design choice with explicit localization for segmentation and inclusion of cross-modal information, such as free text reports, on segmentation.

4. CONCLUSION

In this paper, we presented a novel, two-stage, multimodal framework for accurate segmentation of pneumothorax in chest radiographs. The free-text radiology reports have been used only in [6] to aid segmentation. However, DAFT-based fusion has not been attempted and shows much promise. Our results show that incorporating free-text reports reduces false positive predictions significantly, and the DAFT-based fusion of localization maps improves positive cases. A stratified analysis of performance on different-sized pneumothorax was presented, and the proposed method was seen to give the best results regardless of the size of the abnormality, which is a major strength.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using the CANDID-PTX human sample data [10] made available in open access. Ethical approval was not required, as confirmed by the license attached with the open-access data.

6. REFERENCES

[1] Anne-Maree Kelly et al., “Comparison between two methods for estimating pneumothorax size from chest x-rays,” *Respiratory Medicine*, vol. 100, no. 8, pp. 1356–1359, 2006.

[2] Qingfeng Wang et al., “Automated segmentation and diagnosis of pneumothorax on chest x-rays with fully convolutional multi-scale scse-densenet: A retrospective study,” *BMC Medical Informatics and Decision Making*, vol. 20, no. 14, pp. 1–12, 2020.

[3] Alexey Tolkachev et al., “Deep learning for diagnosis and segmentation of pneumothorax: The results on the kaggle competition and validation against radiologists,” *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1660–1672, 2020.

[4] Xi Ouyang et al., “Weakly supervised segmentation framework with uncertainty: A study on pneumothorax segmentation in chest x-ray,” in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*. Springer, 2019, pp. 613–621.

[5] Zihan Li et al., "Lvit: Language meets vision transformer in medical image segmentation," *IEEE Transactions on Medical Imaging*, 2023.

[6] Zachary Huemann et al., "Contextual net: A multimodal vision-language model for segmentation of pneumothorax," *arXiv preprint arXiv:2303.01615*, 2023.

[7] Go-Eun Lee et al., "Text-guided cross-position attention for segmentation: Case of medical image," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2023*. Springer, 2023, pp. 537–546.

[8] Tom Nuno Wolf et al., "Daft: A universal module to interweave tabular data and 3d images in cnns," *NeuroImage*, vol. 260, pp. 119505, 2022.

[9] Stefanie E Mason et al., "Semi-quantitative visual assessment of chest radiography is associated with clinical outcomes in critically ill patients," *Respiratory Research*, vol. 20, no. 1, pp. 1–9, 2019.

[10] Sijing Feng et al., "Curation of the candid-ptx dataset with free-text reports," *Radiology: Artificial Intelligence*, vol. 3, no. 6, pp. e210136, 2021.

[11] Colin Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.