Location-Aware Parameter Fine-Tuning for Multimodal Image Segmentation

Sicong Gao[1], Maurice Pagnucco[1], and Yang Song[1⋆]

The University of New South Wales, Sydney, Australia
z5306254@ad.unsw.edu.au

Abstract. Accurate segmentation of lung infection regions is critical for early diagnosis and quantitative assessment of disease severity. However, existing segmentation methods largely depend on high-quality, manually annotated data. Although some approaches have attempted to alleviate the reliance on detailed annotations by leveraging radiology reports, their complex model architectures often hinder practical training and widespread clinical deployment. With the advent of large-scale pretrained foundation models, efficient and lightweight segmentation frameworks have become feasible. In this work, we propose a novel segmentation framework that utilizes CLIP to generate multimodal high-quality prompts, including coarse mask, point, and text prompts, which are subsequently fed into the Segment Anything Model 2 (SAM2) to produce the final segmentation results. To fully exploit the informative content of medical reports, we introduce a localization loss that extracts positional cues from the text to guide the model in localizing potential lesion regions. Experiments on the CT dataset MosMedData+ and the X-ray dataset QaTa-COV19 demonstrate that our method achieves state-of-the-art performance while requiring only minimal parameter fine-tuning. These results highlight the effectiveness and clinical potential for pulmonary infection segmentation.

Keywords: Medical Image Segmentation · Vision-language Model · Transfer Learning.

1 Introduction

Lung diseases pose a significant global health challenge, with radiological imaging such as X-ray and Computed Tomography (CT) playing a crucial role in early detection, especially during outbreaks like COVID-19 [8,19,4]. While convolutional neural network (CNN) [10,1] and Transformer [6,3] based segmentation methods have advanced the field, their dependence on extensive pixel-level annotations limits scalability. Reducing annotation needs without sacrificing segmentation accuracy remains a critical challenge [11,13,15,20,23].

To tackle this issue, several recent studies have begun exploring textual information to guide segmentation, and existing multimodal methods for medical

⋆ Corresponding author

imaging predominantly mimic multimodal approaches from the natural image domain [24]. For example, Li et al. [14] proposed LViT, the first method to integrate CNN and Transformer architectures for text-guided segmentation. Similarly, GuideDecoder [26] performs multimodal fusion at the decoder stage, and MMI-UNet [2] applies multimodal fusion in the encoder stage. TGANet [21] proposes a text-guided attention mechanism for polyp segmentation. MTPTN [9] introduces progressive text-based prior prompts to generate multimodal features for nuclei segmentation.

On the other hand, existing methods have not considered that medical reports differ from the textual descriptions associated with natural images. They usually contain various localization terms. As illustrated by the blue text in Fig. 1, this information concisely describes the location of the lesion area. However, current existing approaches heavily rely on attention mechanisms to align the text with the visual modality and extract additional information from the text. Such reliance fails to explicitly capture effective medical cues (e.g., localization information), making it challenging to guide the visual modality in lesion localization accurately. To address this limitation, we propose a location-aware method that employs a global positional feature classification (GPFC) loss to guide the visual modality in accurately localizing lesion regions. By using lesion localization classification as a loss function, we validate whether the model has successfully learned localization cues from the text. Additionally, incorporating such classification at multiple stages further ensures that the text-image encoders can extract effective global features, thereby generating high-quality prompts.

In addition, large-scale pre-trained models have shifted segmentation toward efficient fine-tuning. Foundation models like CLIP [17] and the recent release of SAM2 [18] have further advanced this direction. Although SAM2, an extension of SAM [12] with a larger training dataset and faster performance, offers promising results, it still requires high-quality prompts to avoid segmentation errors [22]. Therefore, it remains challenging to effectively adapt SAM2 to downstream tasks. In this work, we leverage CLIP to generate the prompts required by SAM2 while freezing the text and image encoder and fine-tuning only the bridging module during encoding. To fully exploit CLIP's multimodal capabilities, we introduce a Query-guided Attention Fusion (QGAF) module, which differs from previous approaches that rely on simple attention fusion. The QGAF module acts as a bridge between CLIP's text and image encoders, using learnable query tokens to extract lesion location information from text, enabling CLIP to generate high-quality prompts that meet SAM2's requirements. Our method reduces the burden of pixel-level annotation by leveraging available medical report text.

In summary: (1) We propose a text-guided, location-aware segmentation framework that uses global positional feature classification (GPFC) to precisely localize lesions from positional cues in medical reports. (2) We design a Query-guided Attention Fusion (QGAF) module for fine-tuning CLIP to generate three effective prompts (coarse mask, point, and text) for SAM2. (3) Experimental results on the MosMedData+ [16] and QaTa-COV19 [7] datasets demonstrate our model achieves state-of-the-art performance with minimal parameter fine-tuning.

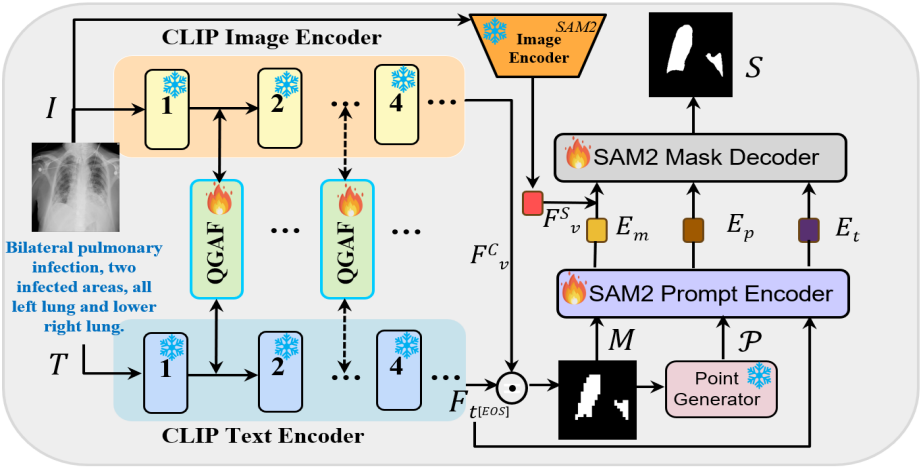

Fig. 1: An overview of our proposed framework, highlighting that only minimal parameter fine-tuning is required. The details of Query-guided Attention Fusion (QGAF) module are shown in Fig. 2.

2 Methodology

Figure 1 shows an overview of our proposed method. Given an image I and a medical report T, we employ CLIP as the encoder and use the QGAF module to fuse multimodal information for extracting image features $F^C{}_v$ (Section 2.1). After that, we generate the coarse mask prompt, point prompt, and text prompt required by SAM2. The SAM2 prompt encoder then processes these prompts to obtain the corresponding encodings E_m, E_p, and E_t (Section 2.2). The SAM2 mask decoder subsequently produces the final segmentation mask S (Section 2.3). The whole process is trained using a combination of location-aware loss and segmentation loss (Section 2.4). All encoders remain frozen during training.

2.1 Text and Image Encoder

Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we extract image features $F^C{}_v \in \mathbb{R}^{N\times C_i}$ using the CXR-CLIP image encoder E_I^C. The corresponding medical report T is processed by the CXR-CLIP text encoder E_T^C to obtain text features $F_t \in \mathbb{R}^{L\times C_t}$. CXR-CLIP [25] is a pretrained model on relevant medical datasets based on CLIP. We utilize the MedSAM2 image encoder E_I^S to extract image features $F_v^S \in \mathbb{R}^{N\times C_i}$ for final segmentation. MedSAM2 [28] is a pretrained model by fine-tuning SAM2 on several medical datasets. Here, C_i and C_t denote the dimensions of the extracted image and text features respectively, N denotes the number of image patches, L signifies the length (i.e., the number of tokens) of the text description. Throughout the process, we partition the CXR-CLIP image-text encoder into four stages based on its blocks. After each stage, the QGAF module is applied for fine-tuning while all encoders remain frozen.

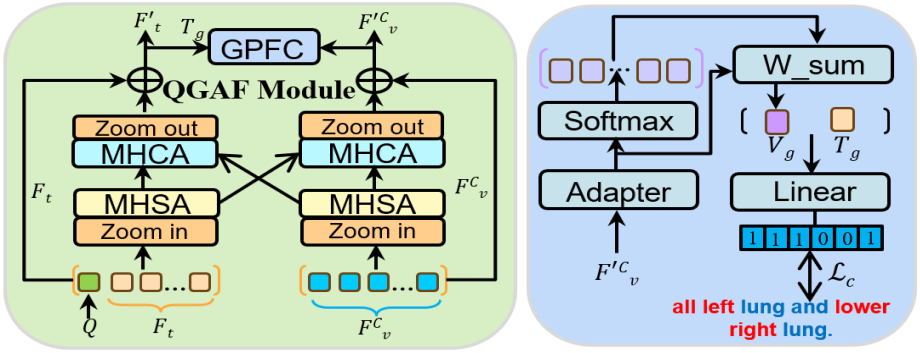

Fig. 2: Overview of the proposed module: the left side shows the Query-guided Attention Fusion (QGAF) module, while the right side presents the Global Positional Feature Classification (GPFC) component in QGAF.

Query-guided Attention Fusion Module As shown in Figure 2, during the image encoding stage, we fuse textual information using the QGAF bridging module to learn localization knowledge for lesion regions. Unlike previous methods that rely on simple attention mechanisms, we employ a learnable query token $Q \in \mathbb{R}^{T \times C_t}$ (T is the token length) and concatenate it with the text feature F_t. To save memory and improve efficiency, we apply a zoom-in operation to reduce the feature dimension to $C_f = 64$, then perform multimodal fusion through self-attention and cross-attention mechanisms, and finally restore the original dimension with a zoom-out operation. The process is formulated as follows:

$$F_t' = F_t + MHCA(MHSA([Q, F_t]), MHSA(F^C{}_v)), \tag{1}$$

$$F'^C{}_v = F^C{}_v + MHCA(MHSA(F^C{}_v), MHSA([Q, F_t])). \tag{2}$$

where, $[\cdot]$ denotes concatenation. For simplicity, the zoom-in and zoom-out operations are not shown in the equations; they are implemented via a feed-forward network (FFN) that reduces and then restores the feature dimension; MHCA is the multihead cross-attn. We describe the GPFC component in Section 2.4.

2.2 Prompt Generation

Coarse Mask Prompt After four encoder stages, we obtain the final image feature $F^C{}_v$ and text feature F_t. We use the end-of-sequence token $F_{t[EOS]} \in \mathbb{R}^{1 \times C_t}$ as the global text feature. Next, we reshape $F^C{}_v$ into $h_i \times w_i \times C_i$ and perform element-wise multiplication with the global text feature to generate the pseudo mask $M \in \mathbb{R}^{h_i \times w_i \times 1}$: $M = F_{t[EOS]} \odot F^C{}_v$. This approach, based on contrastive learning, computes the similarity score between the global text feature and each image pixel feature. A higher similarity score indicates a greater likelihood that the pixel belongs to a lesion region.

Point Prompt Our point prompts are generated from the coarse mask. From the coarse mask M, we derive probability values for each of the $h_i \times w_i$ points indicating the likelihood of belonging to a lesion area. We then select the Top-k points as point prompts. First, candidate points with a probability above a threshold ($\theta > 0.9$) are identified. The first point is chosen as the candidate with the highest likelihood and is added to the set of selected points $\mathcal{P}$. Subsequent points are determined based on the criterion $Dis_{max}(p, \mathcal{P})$, where point p is selected if it is the farthest from the already selected points. This approach ensures that the chosen points are not only highly probable but also well distributed, covering various lesion regions rather than clustering in a single area.

Text Prompt We use the end-of-sequence text token $F_{t[EOS]} \in \mathbb{R}^{1\times C_t}$ from the final text encoder layer as the global information. Then a simple feed-forward network (FFN) process is performed to generate the text prompt.

2.3 Mask Decoder

The three types of prompts are encoded by the MedSAM2 prompt encoder to obtain the coarse mask E_m, point prompt E_p, and text prompt E_t. The SAM2 image encoder extracts the image features F_v^S, which, combined with E_m and the other prompts, are fed into the MedSAM2 mask decoder to yield the final segmentation S. The overall process is expressed as:

$$S = SAM2_{Dec}(F_v^S + E_m, [E_p, E_t]). \tag{3}$$

2.4 Loss Design

Global Positional Feature Classification Component The proposed GPFC component (Fig. 2, right) classifies global positions by partitioning the lung into six lesion regions (upper, middle, and lower lobes for both lungs). To ensure multimodal consistency, the module leverages both the global image feature V_g and global text feature T_g. Firstly, an adapter (comprising two linear layers with activation) is used to refine the image feature. A Softmax function then assigns regional weights, which are aggregated via a weighted sum to produce V_g. Simultaneously, T_g is extracted from the EOS token of the text representation V_t. The global image feature V_g and text feature T_g are concatenated and passed through a classification layer to yield the prediction. Since the GPFC is performed with each encoder block. So, for encoder block i, we have $\ell_{\mathrm{c}_i} = \mathrm{BCE}\big(\mathrm{Linear}([V_g, T_g]), \mathcal{Y}\big)$, where $\mathcal{Y}$ denotes the ground-truth label and BCEWithLogitsLoss is adopted for backpropagation. The overall classification loss is defined as $\ell_{\mathrm{c}} = \sum_{i=1}^{4} \ell_{\mathrm{c}_i}$, since we have 4 encoder blocks.

Segmentation Loss We employ a loss function that combines Dice and Cross-Entropy terms on both coarse mask M and final segmentation S to evaluate segmentation performance. Let U be the total number of pixels and V be the number of classes. For each pixel u and class v (positive or negative), $\hat{y}_{u,v}$

denotes the predicted probability that pixel u belongs to class v, and $y_{u,v}$ is the corresponding ground-truth label. $\ell_{\text{Dice}} = 1 - \frac{1}{UV}\sum_{u=1}^{U}\sum_{v=1}^{V}\frac{2|\hat{y}_{u,v}\cap y_{u,v}|}{|\hat{y}_{u,v}|+|y_{u,v}|}$, $\ell_{\text{CE}} = -\frac{1}{U}\sum_{u=1}^{U}\sum_{v=1}^{V}[\, y_{u,v}\log(\hat{y}_{u,v})]$. The overall segmentation loss ℓ_{s} is formed by equally weighting these two terms: $\ell_{\text{s}} = 0.5\,\ell_{\text{Dice}} + 0.5\,\ell_{\text{CE}}$. Therefore, our proposed loss ℓ can be formed as:

$$\ell = \lambda_1\ell_{\text{s}}(M) + \lambda_2\ell_{\text{s}}(S) + \lambda_3\ell_{\text{c}}. \tag{4}$$

3 Experiments

Datasets We evaluated our method on two publicly available COVID-19 medical image datasets: QaTa-COV19 [7] and MosMedData+ [16]. QaTa-COV19, compiled jointly by Qatar University and Tampere University, comprises 9258 chest X-ray images of COVID-19 cases, partitioned into training (5716 samples), validation (1429 samples), and testing (2113 samples) sets. Each image is accompanied by a lesion mask that delineates the infected regions. MosMedData+ is a large-scale lung CT dataset containing 2729 CT scan slices from multiple sources, with 2183 training, 273 validation, and 273 testing samples—all annotated with ground-truth infection masks. Moreover,we use the medical reports generated in [14] to enable multimodal learning.

Implementation Details We trained our model using the PyTorch framework with a batch size of 5 and an initial learning rate of 3×10^{-4}. The model was optimized using AdamW with a weight decay of 0.001, and a cosine annealing learning rate schedule was applied with a warm-up period of 1000 iterations. The training ran for 70 epochs. We set the length of the learnable query tokens T to 10 and selected $Top3$ point prompts. The loss hyperparameters λ_1, λ_2, and λ_3 were set to 0.3, 0.6, and 0.1, respectively. All experiments were conducted on a system equipped with an RTX-6000 48 GB GPU and 128 GB of RAM.

Quantitative Results We conducted comparative experiments on five unimodal methods and five multimodal methods, among which LViT, GuideDecoder, and MMI-UNet were proposed explicitly for these two datasets. As shown in Table 1, our method achieves consistently the best performance on both datasets. Notably, our model requires only 16.9M trainable parameters, demonstrating its efficiency and ability to transfer pretrained large-scale foundation models, thereby saving training resources. Furthermore, when compared with standalone CXR-CLIP and MedSAM2, our approach, despite using fewer parameters, yields superior performance, proving the feasibility of fine-tuning. Additionally, as shown in Fig. 3, our proposed method enhances the localization of the lesion area, with segmentation results closely aligned with the actual lesions. This validates the effectiveness of our location-aware approach. In contrast, CXR-CLIP and MedSAM2 tends to misclassify normal regions as lesions, which further underscores the efficiency of our prompt generation strategy. Overall, our method effectively overcomes the limitations of both CLIP and SAM2.

Table 1: Performance comparison of various segmentation models on MosMedData+ and QaTa-COV19 datasets.

Method	**Text**	**Param(M)**	**MosMedData+**		**QaTa-COV19**	
			Dice	**mIoU**	**Dice**	**mIoU**
U-Net++ [27]	✗	74.5	0.7175	0.5839	0.7962	0.7025
nnUNet [10]	✗	19.1	0.7259	0.6036	0.8042	0.7081
TransUNet [5]	✗	105	0.7124	0.5844	0.7863	0.6913
SwinUNet [3]	✗	82.3	0.6329	0.5019	0.7807	0.6834
MedSAM2 [28]	✗	38.9	0.5427	0.4109	0.7536	0.6428
CXR-CLIP [25]	✓	136.6	0.7628	0.6327	0.8325	0.7409
LAVT [24]	✓	118.6	0.7329	0.6041	0.7928	0.6989
LViT [14]	✓	29.7	0.7457	0.6133	0.8366	0.7511
GuideDecoder [26]	✓	44	0.7775	0.6360	0.8978	0.8145
MMI-UNet [2]	✓	56.2	0.7842	0.6450	0.9088	0.8328
Ours	✓	16.9	**0.7981**	**0.6572**	**0.9206**	**0.8489**

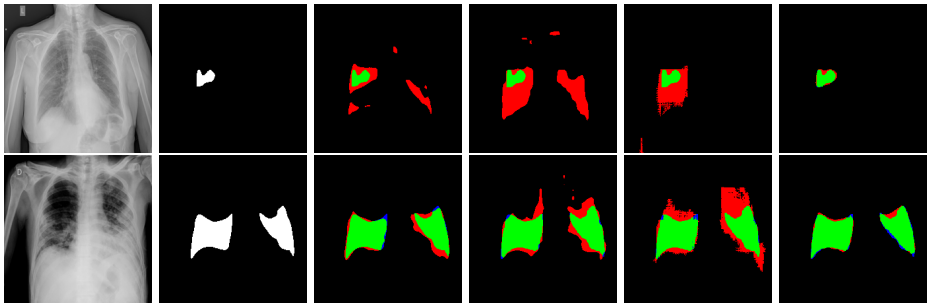

Fig. 3: Visualization of segmentation outcomes on the QaTa-COV19 dataset. In the images, green represents true positives, red denotes false negatives, and blue signifies false positives.

Table 2: Ablation study on the QGAF module and GPFC loss.

Ablation	**MosMedData+**		**QaTa-COV19**	
	Dice	**mIoU**	**Dice**	**mIoU**
Baseline (w/o QGAF)	0.6948	0.5791	0.8036	0.7301
QGAF (w/o GPFC)	0.7627	0.6478	0.9012	0.8227
QGAF (Ours)	**0.7981**	**0.6572**	**0.9206**	**0.8489**

Ablation Studies We conducted experiments to validate our components. Firstly, an ablation study on the QGAF module (Table 2) demonstrates its ability to learn multimodal information for segmentation tasks. We compared a baseline without the QGAF module and evaluated two variants—with and without the GPFC loss. The results indicate that our QGAF module effectively captures

cross-modal information, while the GPFC loss further enhances lesion localization by integrating positional data extracted from radiology reports. Specifically, on MosMedData+, our method improves the Dice score by 10.33% and mIoU by 7.81%, and on QaTa-COV19, the Dice score increases by 11.70% and mIoU by 11.88%. Furthermore, Table 3 shows the QGAF module achieves optimal performance when using the length 10 learnable query token, which effectively leverages text information. Table 4 reveals that using the Top-3 point prompts generally results in superior segmentation performance compared to the Top-4, despite the latter showing a slight increase of 0.38% in mIoU on MosMedData+. Table 5 highlights MedSAM2's dependence on high-quality prompts, as their absence significantly degrades performance. Our analysis of three prompts reveals that the coarse mask prompt has the most significant impact on segmentation results—improving the Dice score by 23.20% and mIoU by 23.45% on MosMedData+, and by 18.65% and 21.03% on QaTa-COV19. This is followed by point and text prompts. Overall, the results demonstrate the critical importance of high-quality prompts for effectively transferring SAM2 to downstream tasks.

Table 3: Ablation study on the length (T) of learnable query token Q.

T	**MosMedData+**		**QaTa-COV19**	
	Dice	**mIoU**	**Dice**	**mIoU**
5	0.7742	0.6273	0.8847	0.7911
10	**0.7981**	**0.6572**	**0.9206**	**0.8489**
15	0.7829	0.6501	0.9182	0.8339
20	0.7641	0.6176	0.8922	0.8001

Table 4: Ablation Study on Top-k point prompts selection for MedSAM2

Top-k	**MosMedData+**		**QaTa-COV19**	
	Dice	**mIoU**	**Dice**	**mIoU**
1	0.7650	0.6220	0.8620	0.7940
2	0.7820	0.6400	0.9170	0.8481
3	**0.7981**	0.6572	**0.9206**	**0.8489**
4	0.7935	**0.6610**	0.9071	0.8215

Table 5: Ablation study on three different prompts.

Coarse Mask	**Points**	**Text**	**MosMedData+**		**QaTa-COV19**	
			Dice	**mIoU**	**Dice**	**mIoU**
✗	✗	✗	0.4816	0.3527	0.6914	0.5829
✓	✗	✗	0.7136	0.5872	0.8779	0.7932
✓	✓	✗	0.7731	0.6392	0.8930	0.8154
✓	✓	✓	**0.7981**	**0.6572**	**0.9206**	**0.8489**

4 Conclusions

In this study, we propose a location-aware segmentation framework that leverages CLIP to generate three high-quality prompts for SAM2. By fine-tuning the

QGAF bridging module with minimal training, our framework achieves high accuracy. In addition, the GPFC localization loss, derived from medical reports, enables our method to precisely localize lesion regions. This work offers a new approach for future lightweight multimodal medical segmentation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of U-Net. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
2. Bui, P.N., Le, D.T., Choo, H.: Visual-textual matching attention for lesion segmentation in chest images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 702–711. Springer (2024)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: European Conference on Computer Vision. pp. 205–218. Springer (2022)
4. Chaunzwa, T.L., Hosny, A., Xu, Y., Shafer, A., Diao, N., Lanuti, M., Christiani, D.C., Mak, R.H., Aerts, H.J.: Deep learning classification of lung cancer histology using CT images. Scientific Reports **11**(1), 1–12 (2021)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUnet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
6. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., et al.: TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
7. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for COIVD-19 detection using chest X-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
8. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. IEEE Transactions on Medical Imaging **39**(8), 2626–2637 (2020)
9. Han, X., Chen, Q., Xie, Z., Li, X., Yang, H.: Multiscale progressive text prompt network for medical image segmentation. Computers & Graphics **116**, 262–274 (2023)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
11. Kim, K., Lee, Y., Park, D., Eo, T., Youn, D., Lee, H., Hwang, D.: Llm-guided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 239–249. Springer (2024)

12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
13. Lei, Y., Li, Z., Shen, Y., Zhang, J., Shan, H.: CLIP-lung: Textual knowledge-guided lung nodule malignancy prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 403–412. Springer (2023)
14. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: LViT: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging (2023)
15. Luo, Y., Liu, W., Fang, T., Song, Q., Min, X., Wang, M., Li, A.: CARL: cross-aligned representation learning for multi-view lung cancer histology classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 358–367. Springer (2023)
16. Morozov, S.P., Andreychenko, A.E., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I.A., Gelezhe, P., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest CT scans with COVID-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
18. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
19. Singh, T., Mishra, S., Kalra, R., Satakshi, Kumar, M., Kim, T.: COVID-19 severity detection using chest X-ray segmentation and deep learning. Scientific Reports **14**(1), 19846 (2024)
20. Su, H., Lei, H., Guoliang, C., Lei, B.: Cross-graph interaction and diffusion probability models for lung nodule segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 482–492. Springer (2024)
21. Tomar, N.K., Jha, D., Bagci, U., Ali, S.: TGANet: Text-guided attention for improved polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 151–160. Springer (2022)
22. Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J., Li, G.: SAM2-Unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. arXiv preprint arXiv:2408.08870 (2024)
23. Yang, H., Shen, L., Zhang, M., Wang, Q.: Uncertainty-guided lung nodule segmentation with feature-aware attention. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 44–54. Springer (2022)
24. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: LAVT: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)
25. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: CXR-CLIP: Toward large scale chest X-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)

26. Zhong, Y., Xu, M., Liang, K., Chen, K., Wu, M.: Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest X-ray images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 724–733. Springer (2023)
27. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Transactions on Medical Imaging **39**(6), 1856–1867 (2019)
28. Zhu, J., Qi, Y., Wu, J.: Medical SAM 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)