MedTSC-Net: Bridging Semantic and Scale Gaps in Text-Guided Medical Image Segmentation

Zhaomin Chen[1], Danyang Cheng[1], Yisu Ge[1(✉)], Guodao Zhang[2], Hui Huang[1], and Huiling Chen[1(✉)]

College of Computer Science and Artificial Intelligence, Wenzhou University, WenZhou, China
ysg@wzu.edu.cn
chenhuiling.jlu@gmail.com
Institute of Intelligent Media Computing, Hangzhou Dianzi University, Hangzhou, China

Abstract. Leveraging diagnostic reports as auxiliary guidance has emerged as a promising solution to data scarcity in medical referring image segmentation. However, existing methods often suffer from modality heterogeneity due to shallow interaction strategies and scale mismatch between pre-trained encoders and task-specific decoders. In this paper, we propose MedTSC-Net, a systematic and optimized framework designed to bridge these semantic and scale gaps through a harmonized encoding-decoding paradigm. Specifically, we introduce a Visual-Linguistic Harmonizer (VLH) that projects image and text representations into a shared latent space, mitigating semantic shifts at the early encoding stage. To ensure precise lesion delineation, a Channel-Spatial Attention (CASA) module is devised to model complementary text-guided dependencies across multiple dimensions. Furthermore, we address the receptive field inconsistency via a Receptive Field Alignment (RFA) mechanism, which recalibrates feature scales to enhance boundary localization. Extensive experiments on the QaTa-COV19 and MosMedData+ datasets demonstrate that MedTSC-Net significantly outperforms state-of-the-art methods, achieving Dice scores of 91.89% and 80.64%, respectively. Notably, our framework exhibits superior robustness in few-shot scenarios, highlighting its potential for clinical applications with limited annotations. Our code is publicly available at https://github.com/DYcheng-tech/MedTSCNet.

Keywords: Vision Language Model · Medical Image Segmentation · Text Guided · Multimodal Learning · Cross-Modal Alignment

1 Introduction

Automated medical image segmentation is pivotal for diagnosis and therapy monitoring [1, 2]. Despite the success of standard architectures like U-Net [3] and Transformers [4, 5], their performance relies heavily on large-scale annotated datasets. In clinical practice, strict privacy regulations and annotation

costs result in data scarcity [6], limiting the generalization of unimodal models, especially for ambiguous boundaries and complex lesion morphologies [7].

To mitigate this, leveraging diagnostic reports as auxiliary guidance has emerged as a promising direction. In medical referring segmentation, these descriptions are typically derived from radiological findings or physician referral notes and contain clinically relevant information such as lesion location, distribution, and quantity. Unlike general referring image segmentation [8, 9] focusing on holistic scene descriptions, medical text is structured and spatially coupled with anatomy. While pioneering works like LViT [10] and Ariadne's Thread [11] incorporate textual priors, and others explore bidirectional interaction [12, 13], existing methods remain constrained by three limitations. First, restricted interaction often arises from late-fusion strategies [11] or shallow early alignment [10], neglecting low-level visual-linguistic synergies. Second, suboptimal attention mechanisms (e.g., standard dot-product attention) primarily model spatial dependencies [14], overlooking channel-wise semantics crucial for tissue distinction. Third, receptive field mismatch between pre-trained encoders and task-specific decoders creates semantic gaps that hinder precise boundary delineation [15].

In this work, we propose MedTSC-Net, a systematic and clinically motivated framework for COVID-19 referring segmentation designed to bridge the semantic and scale gaps. Specifically, we introduce a Visual-Linguistic Harmonizer (VLH) to align cross-modal representations within the encoder, effectively mitigating early semantic shifts. To refine the decoding stage, a Channel-Spatial Attention (CASA) module is devised to model complementary multi-dimensional dependencies, while a Receptive Field Alignment (RFA) mechanism recalibrates feature scales for precise boundary localization. Experiments on MosMedData+ and QaTa-COV19 demonstrate that MedTSC-Net significantly outperforms state-of-the-art methods, demonstrating consistent robustness across two COVID-19 imaging modalities and few-shot settings.

2 Methods

2.1 Overview

The overall framework of MedTSC-Net is illustrated in Fig 1. Following a U-Net-like encoder-decoder paradigm, it integrates a CNN-based Visual Encoder (ConvNeXt-Tiny) and a Transformer-based Text Encoder (BERT) to achieve text-guided segmentation.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the visual encoder extracts multi-scale hierarchical features $F^E = \{F_1^E, \ldots, F_4^E\}$, where F_i^E represents the feature map at the i-th stage with a resolution of $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$. Simultaneously, the input text T is processed by BERT to yield contextual embeddings $H \in \mathbb{R}^{L\times d}$. Unlike standard separate encoding, we introduce the Visual-Linguistic Harmonizer (VLH) directly into the intermediate layers of both backbones. VLH projects modality-specific features into a unified latent space, enabling implicit alignment and domain adaptation at the early encoding stage.

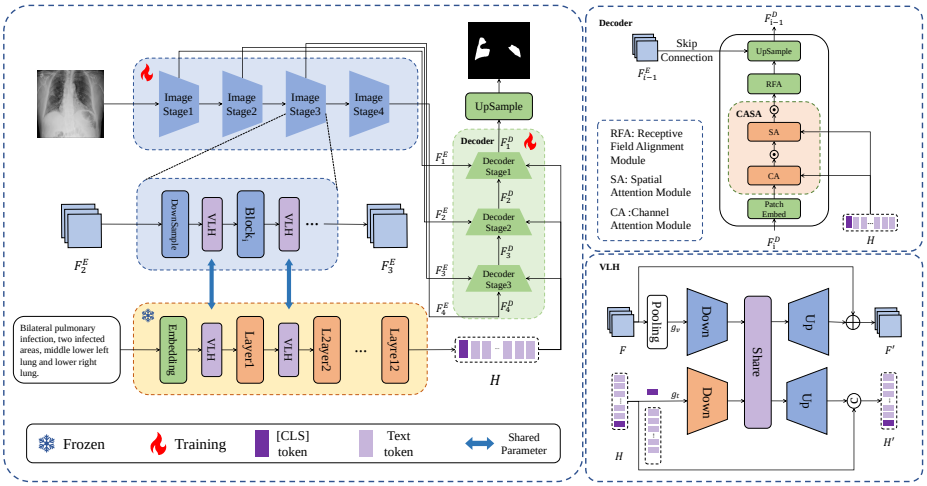

Fig. 1: Overview of MedTSC-Net. A VLH module enables early cross-modal alignment in the hybrid encoder, while the CASA module for multi-dimensional feature fusion, and RFA module for receptive field alignment.

The decoder aims to reconstruct the segmentation mask M by progressively upsampling features and fusing them with encoder features via skip connections. To simultaneously bridge the semantic gap and resolve the receptive field mismatch, we integrate the Channel-Spatial Attention (CASA) and Receptive Field Alignment (RFA) modules as the core decoding units.

At each decoding stage i, CASA first modulates the decoder feature F_i^D using the textual embedding H as a semantic guide:

$$\tilde{F}_i^D = \mathrm{CASA}(F_i^D, H). \tag{1}$$

Subsequently, to align the feature scale with the encoder's granularity, the refined feature $\tilde{F}_i^D$ is processed by the RFA module:

$$\hat{F}_i^D = \mathrm{RFA}(\tilde{F}_i^D), \tag{2}$$

where $\hat{F}_i^D$ denotes the semantically and spatially aligned feature. Finally, we upsample $\hat{F}_i^D$ and concatenate it with the corresponding encoder feature F_{i-1}^E to generate the input for the next stage:

$$F_{i-1}^D = \mathrm{Conv}_{3\times 3}\left(\left[\mathrm{Upsample}(\hat{F}_i^D), F_{i-1}^E\right]\right), \tag{3}$$

where $[\cdot,\cdot]$ denotes channel-wise concatenation. This cascaded design ensures that the recovery of spatial details is guided by both rich semantic context and precise receptive field calibration.

2.2 Visual-Linguistic Harmonizer (VLH)

To resolve structural asymmetry between the hierarchical visual backbone (N_v blocks) and columnar text encoder (N_t layers), we introduce VLH with a Top-down Symmetric Alignment strategy. Positing that cross-modal affinity peaks at deeper abstraction levels, we align the last K units of both encoders in a paired manner as $(\mathcal{B}^v_{N_v-k}, \mathcal{L}^t_{N_t-k})$ for $k \in \{0, \ldots, K-1\}$, where K determines the alignment depth. This ensures that high-level abstract visual concepts are directly calibrated against the richest textual context.

Mechanically, VLH functions as a parameter-efficient adapter. We first extract global descriptors g_m ($m \in \{v, t\}$) via GAP (for vision) or the [CLS] token (for text). These descriptors are projected into a unified latent manifold via a shared interaction matrix W_{share}:

$$\hat{g}_m = W^m_{up}\sigma\left(W_{share}\sigma\left(W^m_{down}g_m\right)\right), \tag{4}$$

where $W^m_{down/up}$ are modality-specific projectors. By forcing heterogeneous semantics through the unique W_{share}, VLH implicitly learns a modality-invariant representation. Finally, the aligned semantics $\hat{g}_m$ are reinjected to update the original features:

$$F' = F + \text{Broadcast}(\hat{g}_v), \quad H' = [\hat{g}_t, H_1, \ldots, H_L]. \tag{5}$$

This lightweight design achieves deep semantic alignment with negligible overhead.

2.3 Channel-Spatial Attention (CASA)

The CASA module sequentially modulates decoder features F^D_i using global text embeddings H to enforce semantic consistency and spatial precision.

Semantic-Aware Channel Attention (CA). To inject lesion-specific semantics (e.g., "Bilateral pulmonary infection") into visual channels, we propose an efficient local interaction strategy replacing heavy MLPs. We fuse visual descriptors (from GAP (v_{avg}) and GMP ($v_{\max}$)) with the projected text token h'_{cls}. The channel saliency map W_i is computed as:

$$W_i = \sigma\left(\mathcal{K}_k(v_{avg} + h'_{cls}) + \mathcal{K}_k(v_{\max} + h'_{cls})\right), \tag{6}$$

where $\mathcal{K}_k$ is a 1D convolution with adaptive kernel size k. The feature is then modulated via $F'_i = W_i \odot F^D_i$.

Text-Guided Spatial Attention (SA). Subsequently, to refine localization, we employ a parameter-efficient cross-attention mechanism. We compute the affinity between the channel-refined feature F'_i and text sequence H, simplifying standard attention to focus solely on spatial saliency:

$$S_i = \text{Softmax}\left(\text{Pool}_L\left(\frac{(F'_iW_q)(HW_k)^\top}{\sqrt{C}}\right)\right) \in \mathbb{R}^{HW}, \tag{7}$$

Table 1: Comparative results on QaTa-COV19 and MosMedData+ versus state-of-the-art (SOTA) methods. The best results are highlighted in bold, while the second-best are underlined. An asterisk (*) indicates that the model's source code is not publicly available, and we have adopted the results from the original literature.

	Method	QaTa-COV19				MosMedData+			
		Dice	IoU	P	R	Dice	IoU	P	R
Uni-modal	Unet [3]	84.57	73.26	83.68	85.47	69.60	53.38	73.76	65.89
	SwinUnetr [5]	84.90	73.77	82.93	86.97	75.40	60.52	81.64	70.05
	Unet++ [16]	87.25	77.39	96.90	85.61	74.55	59.43	79.50	70.19
	UCTransNet [17]	87.97	78.53	86.71	89.27	76.62	62.10	78.90	74.47
	nnUnet [18]	87.59	77.92	85.43	89.86	78.73	64.92	82.04	75.68
Multi-modal	LAVT [8]	89.78	81.46	89.19	90.38	75.80	61.03	78.53	73.26
	CRIS [9]	89.85	81.58	89.24	90.48	77.74	63.59	78.06	77.42
	TGCAM* [19]	90.60	82.81	-	-	77.82	63.69	-	-
	ABP* [13]	91.03	<u>83.53</u>	-	-	-	-	-	-
	LViT [10]	90.70	82.99	90.35	<u>91.06</u>	78.96	65.24	<u>83.15</u>	75.18
	Ariadne [11]	<u>91.10</u>	83.48	<u>91.23</u>	90.76	<u>79.87</u>	<u>66.49</u>	79.79	**79.96**
	MedTSC-Net (Ours)	**91.89**	**84.99**	**91.43**	**92.35**	**80.64**	**67.56**	**83.85**	<u>77.66</u>

where W_q and W_k are learnable projection matrices, C denotes the channel dimension of the feature F_i', and $\sqrt{C}$ serves as the scaling factor to prevent gradient vanishing in the Softmax function. Pool_L aggregates scores across the text length L via sum-pooling of average and max values. The final spatially-refined feature is obtained by $F_i'' = S_i \odot F_i'$.

2.4 Receptive Field Alignment (RFA)

Standard upsampling operations often dilute the Effective Receptive Field (ERF), leading to feature misalignment where textual priors lose their contextual anchor points. To bridge the intrinsic scale gap between the deep encoder's global semantic scope and the decoder's limited local focus, we introduce the RFA module immediately after spatial attention.

Inspired by InceptionNeXt [20], RFA recalibrates the ERF to match the encoder's granularity before fusion. Specifically, the input feature F_i'' is split along the channel dimension into four groups $x_{1\ldots4}$. We explicitly model multi-scale contexts via parallel large-kernel convolutions:

$$F_i^{RFA} = \text{Conv}_{1\times1}\left([\mathcal{D}_{3\times3}(x_1), \mathcal{D}_{1\times R}(x_2), \mathcal{D}_{R\times1}(x_3), x_4]\right) + F_i'', \tag{8}$$

where $[\cdot]$ denotes channel-wise concatenation, $\mathcal{D}_{k_1\times k_2}$ refers to a depth-wise convolution with a kernel size of $k_1 \times k_2$. By setting an adaptive kernel size $R \approx H/2$, the orthogonal branches capture long-range dependencies, expanding the effective receptive field to encompass large-scale anatomical structures.

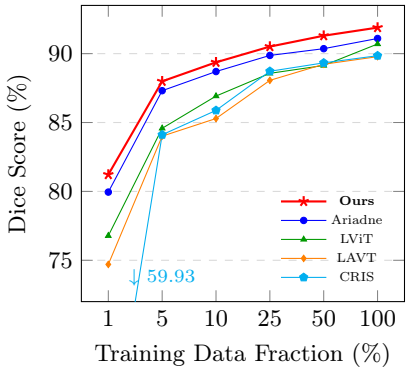

Fig. 2: **Few-shot Analysis.** Comparison on QaTa-COV19 shows MedTSC-Net (red) maintains superior robustness. Note that CRIS yields 59.93% at 1% data (marked below axis) due to unstable convergence.

Table 2: **Comprehensive Ablation Study.** Quantitative evaluation of different module combinations on the QaTa-COV19 dataset. The results demonstrate the synergistic effect of the VLH, CASA, and RFA components, with the full MedTSC-Net achieving the peak performance.

VLH	CASA	RFA	Dice
			87.40
✓			91.08
	✓		91.30
		✓	89.08
✓		✓	91.48
✓	✓		91.43
	✓	✓	91.50
✓	✓	✓	**91.89**

3 Experiments

3.1 Dataset

We evaluate MedTSC-Net on two public multimodal COVID-19 benchmarks covering both chest X-ray and CT modalities. Following LViT [10], we strictly adopt the same data partition protocols to ensure fair comparison. The text annotations are directly adopted from LViT and consist of expert-curated clinical descriptions summarizing lesion location, spatial distribution, and infection quantity, mimicking real-world radiological referral notes. Specifically, QaTa-COV19 [21] contains 9,258 chest X-ray images (224×224) with pixel-level annotations, split into 6,429 training, 716 validation, and 2,113 testing samples, while MosMedData+ [22] comprises 2,729 CT slices of varying infection severities, with 2,182 images for training, 274 for validation, and 273 for testing.

3.2 Implementation Details

Experiments were implemented in PyTorch on dual NVIDIA RTX 3090 GPUs. We trained using AdamW (weight decay 10^{-4}, batch size 64) with initial learning rates of 6×10^{-4} (QaTa-COV19) and 1×10^{-3} (MosMedData+), managed by ReduceLROnPlateau. Images were resized to 224^2 following the preprocessing protocol of LViT [10]. For VLH, the alignment depth was set to $K = 12$, enabling all layers of the 12-layer BERT encoder to participate in cross-modal alignment. For CASA, the adaptive kernel size was computed as $k = \left|\frac{\log_2(C)+1}{2}\right|_{odd}$. Performance was evaluated using Dice, IoU, Precision, and Recall.

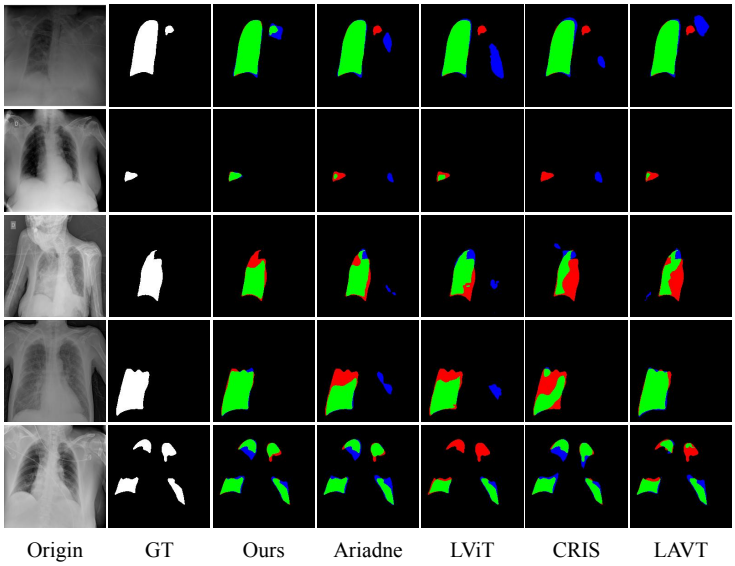

Fig. 3: Qualitative analysis results on the Qata-COV19 dataset. Green: TP, Red: FP, Blue: FN.

3.3 Quantitative Evaluation

Comparison with State-of-the-Art Methods We compare MedTSC-Net against comprehensive baselines, including uni-modal networks (U-Net family [3, 5, 16–18]) and multi-modal approaches (LAVT [8], CRIS [9], TAGCAM [19], ABP [13], LViT [10], Ariadne [11]). Following the standard referring segmentation protocol established by LViT [10] and Ariadne [11], all multimodal methods are evaluated using identical image-text pairs. For fair comparison, we re-trained all open-source models under identical settings. Uni-modal methods are additionally included to quantify the contribution of textual priors. As shown in Table 1, MedTSC-Net consistently achieves state-of-the-art performance on both QaTa-COV19 (X-ray) and MosMedData+ (CT), surpassing the strongest competitor Ariadne (Table 1). Notably, our method achieves the best balance between Precision and Recall, indicating that the proposed alignment strategies effectively reduce both false positives and false negatives.

Data Efficiency and Few-Shot Analysis. To assess robustness under data scarcity, we trained models using subsets (1%–50%) of the QaTa-COV19 training data. As shown in Fig 2, MedTSC-Net consistently dominates across all data regimes. Most notably, with only 5% labeled data, MedTSC-Net achieves a Dice score of 87.99%, rivaling the performance of some fully supervised uni-modal methods (e.g., U-Net++ at 87.25%). This confirms that leveraging textual priors effectively compensates for the lack of visual supervision, significantly reducing the annotation burden for clinical deployment.

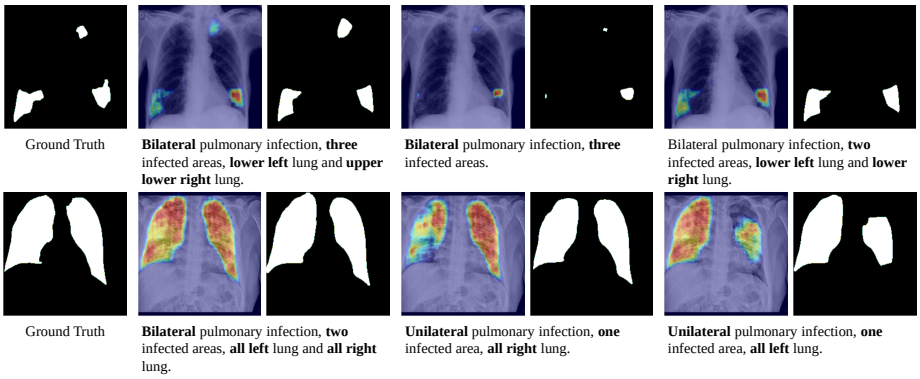

Fig. 4: Visualization of attention guidance via text pairs: We manually adjust the text descriptions of samples to observe their impact on the attention maps and final segmentation results.

Efficiency Analysis. Compared with Ariadne, MedTSC-Net introduces only a modest increase in trainable parameters (45M vs. 43M, +4.7%). The complete framework requires 15.7 GFLOPs per image and achieves an average inference latency of 45 ms on a single RTX 3090 GPU, indicating that the proposed modules introduce limited computational overhead.

3.4 Ablation study

We validate the necessity of each proposed component by evaluating all possible combinations on the QaTa-COV19 dataset. As shown in Table 2, the baseline uni-modal network yields a Dice score of merely 87.40%. The individual insertion of VLH, CASA, and RFA independently boosts the performance, confirming their separate values in semantic alignment and spatial refinement. Notably, combining any two modules yields further improvements over single-module additions. Ultimately, the integration of all three modules (the full MedTSC-Net) achieves the highest Dice score of 91.89%, demonstrating a strong synergistic effect where early representation harmonization (VLH) perfectly complements decoder-stage multi-dimensional attention (CASA) and scale calibration (RFA).

3.5 Qualitative Evaluation

Fig 3 shows that while competitors (e.g., LViT) struggle with small or dispersed lesions due to limited alignment, MedTSC-Net achieves superior structural coherence via VLH and CASA, precisely delineating subtle boundaries. We further validate interpretability in Fig 4 by manipulating textual prompts. The model demonstrates accurate attention shifts in response to attribute changes (Case 1) while retaining sensitivity to visually salient lesions despite biased text (Case 2). This confirms that MedTSC-Net utilizes text as a flexible semantic prior, ensuring robustness against noisy clinical reports.

4 Conclusion and Discussion

In this paper, we presented MedTSC-Net, a text-guided framework for medical referring segmentation. By integrating VLH, CASA, and RFA, the proposed framework enhances multimodal interaction and boundary localization. Experiments on CT and chest X-ray COVID-19 benchmarks demonstrate consistent improvements over strong multimodal baselines, particularly under few-shot settings. These results suggest that clinically relevant textual descriptions can provide valuable semantic priors for medical referring segmentation. Future work will investigate broader disease scenarios and anatomical structures, as well as extensions to 3D volumetric medical image analysis.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. U25A20450, 62571374, 62401398), and the Zhejiang Provincial Natural Science Foundation of China (Grant No. LMS26F020041, LQ24F020016).

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

References

1. Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., Liu, F.: Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. Bioengineering **11**(1034) (2024)
2. Shi, T., Kujawa, A., Linares, C., Vercauteren, T., Booth, T.C.: Automated longitudinal treatment response assessment of brain tumors: A systematic review. Neuro Oncology **27**(8), 1946–1971 (2025)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 234–241 (2015)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis **97**, 103280 (2024)
5. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. pp. 272–284 (2022)
6. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence **2**(6), 305–311 (2020)
7. Tohka, J.: Partial volume effect modeling for segmentation and tissue classification of brain magnetic resonance images: A review. World Journal of Radiology **6**(11), 855–864 (2014)
8. 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. 18134–18144 (2022)

9. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11686–11695 (2022)
10. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D.: Lvit: Language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging **43**(1), 96–107 (2024)
11. 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: Medical Image Computing and Computer Assisted Intervention. pp. 724–733 (2023)
12. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: Medical Image Computing and Computer Assisted Intervention. pp. 192–201 (2024)
13. Zeng, X., Zeng, P., Cui, J., Li, A., Liu, B.: Abp: Asymmetric bilateral prompting for text-guided medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 54–64 (2024)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 1–11 (2017)
15. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 29 (2016)
16. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. pp. 3–11 (2018)
17. Wang, H., Cao, P., Wang, J., Zaiane, O.R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. Proceedings of the AAAI Conference on Artificial Intelligence **36**(3), 2441–2449 (2022)
18. Isensee, F., Jaeger, P.F., Kohl, S.A.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)
19. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L.: Tgcam: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: Medical Image Computing and Computer Assisted Intervention. pp. 192–201 (2024)
20. Yu, W., Zhou, P., Yan, S., Wang, X.: Inceptionnext: When inception meets convnext. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5672–5683 (2024)
21. Degerli, A., Kiranyaz, S., Chowdhury, M.E.H., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing. pp. 2306–2310 (2022)
22. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A.V., Ledikhova, N.V.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)