ViTexNet: Vision-Text Guided Dynamic Convolution Network for Medical Image Segmentation

Rahul Bhardwaj[1⋆], Utkarsh Yashwant Tambe[2], and Debanga Raj Neog[1]

[1] Mehta Family School of Data Science & Artificial Intelligence, Indian Institute of Technology (IIT) Guwahati, Guwahati, India
{r.bhardwaj, dneog}@iitg.ac.in
[2] Department of Data Science & Business Systems, SRM Institute of Science & Technology, Kattankulathur, India
ty7171@srmist.edu.in

Abstract. Recent advancements in medical image segmentation have leveraged multi-modal learning, incorporating textual descriptions to enhance segmentation accuracy. However, existing approaches suffer from high computational costs and inefficient text-vision fusion mechanisms, necessitating a more accurate yet computationally efficient solution. To address this, we propose ViTexNet, a novel vision-language segmentation model that introduces Text-Guided Dynamic Convolution (TGDC) for effective and lightweight fusion of medical visual features and textual cues. Unlike standard cross-attention mechanisms, which impose high parameter complexity, TGDC dynamically refines image features by leveraging relevant textual semantics at each decoder stage, ensuring efficient feature modulation without excessive overhead. By adaptively emphasizing clinically significant regions based on textual descriptions, TGDC enhances segmentation performance while maintaining computational efficiency. Extensive evaluations on QaTa-COV19 and MosMedData+ datasets demonstrate ViTexNet's state-of-the-art performance, achieving 90.76% Dice and 83.25% mIoU on QaTa-COV19, and 78.19% Dice and 64.04% mIoU on MosMedData+, while operating at just 11.5G FLOPs, substantially lower than competing models. Ablation studies confirm TGDC's superiority over cross-attention-based methods, highlighting its effectiveness in optimizing segmentation accuracy without computational trade-offs. The source code is made publicly available at: https://github.com/bhardwaj-rahul-rb/vitexnet

Keywords: Multi-modal learning · Language-guided segmentation · Medical image segmentation.

1 Introduction

Image segmentation is a fundamental task in medical image analysis, focusing on identifying key regions of interest (ROIs) such as tumors, which critically influ-

⋆ Corresponding author

ence disease monitoring and treatment effectiveness [3]. Although deep learning-based approaches have shown promising results for automatic segmentation, they rely heavily on large, expert-labeled datasets. Such annotations are costly and time-consuming to acquire, and even self-supervised techniques lack the explicit supervision needed to reach high accuracy.

Fortunately, medical text reports, typically generated by clinicians, can serve as auxiliary semantic guidance and are often readily available alongside imaging data, thereby mitigating the need for additional data collection. Recent language-guided methods [15, 11, 8] leverage these text features to compensate for visual data limitations, enriching segmentation models with complementary clinical insights and reducing reliance on scarce labeled datasets. For instance, TGANet [15] introduced a text-guided attention network for polyp segmentation, while LViT [11] employed a hybrid CNN–Transformer to fuse image and text features for chest X-ray segmentation. More recently, RecLMIS [8] proposed a cross-modal alignment mechanism using reconstruction-based techniques, demonstrating robust performance across multiple tasks. HCFNet [21] introduced a hybrid decoder integrating multi-head cross-attention, a learnable feature-modulation block (LCFM), and a multi-stage contrastive loss, resulting in a resource-intensive segmentation network. Despite these advances, existing approaches still rely on relatively heavy attention-based fusion strategies, which increase computational overhead and highlight the need for more lightweight yet effective methods.

This work presents ViTexNet, an efficient multimodal segmentation approach that integrates visual and textual information at a notably low computational cost (11.5G FLOPs). Its core is a novel Text-Guided Dynamic Convolution (TGDC) module that globally pools text into a gating vector, which modulates depthwise convolutions for image–text fusion. Experiments on QaTa-COV19 and MosMedData+ show that ViTexNet outperforms both uni-modal and multi-modal state-of-the-art methods. Ablation studies confirm TGDC's effectiveness compared to cross-attention and a combined self+cross attention mechanism. Additional experiments on text prompt granularity reveal that more detailed positional information in the textual description leads to improved segmentation results, underscoring the value of fine-grained linguistic guidance.

2 Method

Figure 1 provides an overview of the ViTexNet architecture, which is composed of three main components: (i) an Image Encoder and Text Encoder that extract feature representations for the visual and textual data, (ii) a Text-Guided Dynamic Convolution (TGDC) module for efficient fusion of these features, and (iii) a Multi-modal Decoder that refines and upsamples the fused tokens for segmentation. The following sections describe each of these components in detail.

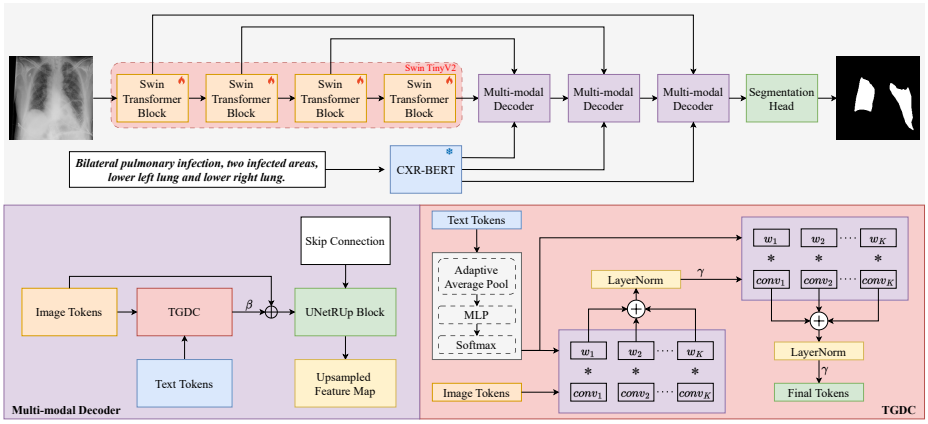

Fig. 1. Overview of ViTexNet, comprising three components: Image and Text Encoders, TGDC Fusion Module, and Multi-Modal Decoder.

2.1 Image Encoder and Text Encoder

Image Encoder: We adopt the Swin-V2 Tiny Transformer [12] as our image encoder to extract multi-scale visual features from an input image of size $I \in \mathbb{R}^{H\times W\times 3}$. The model processes the image in four hierarchical stages, progressively downsampling the spatial resolution while increasing the channel dimensionality. We denote the resulting feature maps as: $I_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 96}$, $I_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 192}$, $I_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 384}$, $I_4 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 768}$. These multi-level feature maps capture both global contextual information and fine-grained spatial details, which are subsequently leveraged in our decoder for accurate segmentation.

Text Encoder: We employ the domain-specific CXR-BERT [1], which has been pretrained on large-scale chest X-ray reports and demonstrates strong capability in extracting clinical semantics. For a tokenized text input $T \in \mathbb{R}^{L}$, with L denoting the length of the tokenized text sequence, CXR-BERT maps T to a final embedding in $\mathbb{R}^{L\times C}$, where C is the dimensionality of each token embedding. Following the standard BERT-base configuration, we set $C = 768$ and freeze all parameters to preserve its pretrained domain knowledge. These text embeddings are reused at each stage of the multi-modal decoder, providing consistent textual guidance throughout the segmentation process.

2.2 Text-Guided Dynamic Convolution (TGDC)

The TGDC module fuses image tokens $V \in \mathbb{R}^{B\times N\times C}$ (from the image encoder) with text tokens $T \in \mathbb{R}^{B\times L\times C}$ (from the text encoder). Here, B is the batch size, N the number of flattened spatial tokens, L the text sequence length, and C the feature dimension. The process unfolds in three main stages.

Global Text Gating: The text tokens T are first pooled along the token dimension to obtain a single vector per sample, as given in Eq. (1):

$$t_{\text{pool}} = \text{AdaptiveAvgPool1d}(T) \in \mathbb{R}^{B\times C} \tag{1}$$

A two-layer MLP with ReLU activation then maps each t_{pool} to K scalar weights, as defined in Eq. (2), which are then normalized by a softmax function:

$$[w_1, w_2, \ldots, w_K] = \text{Softmax}\big(\text{MLP}(t_{\text{pool}})\big) \tag{2}$$

where $w_i \geq 0$ and $\sum_{i=1}^{K} w_i = 1$. These weights specify how strongly each filter should contribute, based on the global textual description (e.g., "Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung").

Parallel Depthwise Convolutions and Weighted Fusion: Next, K parallel depthwise 1D convolutions $\{\text{conv}_1, \ldots, \text{conv}_K\}$ are applied to V. Each convolution has a 3×1 kernel, sliding along the token dimension N. A depthwise 1D convolution is applied so that each of the C channels is convolved independently, substantially reducing parameters and compute compared to a standard convolution. Since the Swin encoder already captures 2D structure, flattening V from (B, N, C) to (B, C, N) for a 1D kernel remains effective.

For the i-th filter, the activation is computed as in Eq. (3):

$$\mathbf{O}_i = \text{conv}_i\big(V^\top\big) \in \mathbb{R}^{B\times C\times N}. \tag{3}$$

This is then rearranged back to (B, N, C) for consistency. Each output $\mathbf{O}_i$ is scaled by its corresponding weight w_i and summed, as shown in Eq. (4):

$$\mathbf{F}^{(1)} = \sum_{i=1}^{K} w_i\, \mathbf{O}_i. \tag{4}$$

A LayerNorm is applied, followed by multiplication by a learnable scale γ, as defined in Eq. (5):

$$\mathbf{F}^{(1)}_{\text{scaled}} = \gamma\, \text{LayerNorm}\big(\mathbf{F}^{(1)}\big). \tag{5}$$

Choosing $K = 4$ follows prior dynamic convolution studies [4], balancing representation power and generalization, while a 3×1 kernel captures short-range token interactions efficiently.

Iterative Refinement: To refine the features further, the same depthwise convolutions and weighted fusion step is repeated within the TGDC module. The partially fused tokens $\mathbf{F}^{(1)}_{\text{scaled}}$ feed back into the same filters and text-derived weights, as defined in Eq. (6):

$$\mathbf{F}^{(2)} = \sum_{i=1}^{K} w_i\, \text{conv}_i\big(\mathbf{F}^{(1)\top}_{\text{scaled}}\big) \tag{6}$$

$$\mathbf{F}^{(2)}_{\text{scaled}} = \gamma \, \text{LayerNorm}\big(\mathbf{F}^{(2)}\big) \tag{7}$$

The final fused tokens $\mathbf{F}^{(2)}_{\text{scaled}}$ (in Eq. (7)) are passed on to subsequent stages in the decoder. Over training, each depthwise filter can specialize in distinct local patterns, while the global gating from t_{pool} ensures the text context highlights the most relevant features for accurate segmentation.

2.3 Multi-modal Decoder

The original image tokens V and the final TGDC output $\mathbf{F}^{(2)}_{\text{scaled}}$ are combined with a learnable scale parameter β, as given in Eq. (8):

$$\mathbf{X}_{\text{fused}} = V + \beta \, \mathbf{F}^{(2)}_{\text{scaled}} \tag{8}$$

This ensures the text-driven enhancements from TGDC are integrated with the raw image features while preserving essential spatial information.

The fused tokens $\mathbf{X}_{\text{fused}}$ are reshaped from (B, N, C) to a 2D feature map (B, C, H, W), where $H \times W = N$. The UNETRUp block increases the spatial resolution (Eq. (9)), aligning the tokens with the corresponding skip feature map S from the encoder:

$$\mathbf{U} = \text{UnetrUpBlock}\big(\text{Reshape}(\mathbf{X}_{\text{fused}})\big) \tag{9}$$

Concatenating $\mathbf{U}$ and $\mathbf{S}$ along the channel dimension provides a higher-resolution feature map $\mathbf{M}$ that combines encoder and text-refined information, as shown in Eq. (10):

$$\mathbf{M} = [\mathbf{U}, \mathbf{S}]_{\text{channel}}. \tag{10}$$

Finally, a 1×1 convolution projects the channels to the desired number of classes, and a Sigmoid activation produces the segmentation mask, as defined in Eq. (11):

$$\mathbf{O} = \sigma\Big(\text{Conv}_{1\times1}(\mathbf{M})\Big) \;\in\; \mathbb{R}^{B\times1\times H\times W} \tag{11}$$

3 Experiments and Results

3.1 Dataset

To assess the effectiveness of the proposed method, two medical datasets were employed: QaTa-COV19 [5] and MosMedData+ [13]. QaTa-COV19 contains 9258 chest X-ray images of COVID-19 cases, accompanied by extended medical notes from [11]. Following the data split used in LViT [11], 5716 images were allocated for training, 1429 for validation, and 2113 for testing. MosMedData+ consists of 2729 CT slices depicting lung infections; again, adhering to the LViT [11] split, the dataset was divided into 2183 training images, 273 validation images, and 273 test images.

3.2 Implementation Details

Implementation was carried out in PyTorch on an NVIDIA A100 GPU (40GB). A cosine annealing schedule reduced the learning rate from 3e4 to 1e6. Input images were 224×224, with a 10% probability of random zoom. The batch size was 32, and DiceCE loss was used alongside AdamW optimization. Training ran for 200 epochs, with early stopping after 50 epochs of no improvement. Dice and mIoU served as primary metrics, with Dice offering a more precise measure for smaller targets.

Table 1. Quantitative comparison of segmentation results on QaTa-COV19 and MosMedData+, encompassing both uni-modal and multi-modal learning methods (separated by a dashed line). Each approach is evaluated in terms of parameters (M), FLOPs (G), and Dice/IoU scores (in %). Methods are annotated according to backbone architectures: CNN-based ($^\diamond$), SAM-based (¶), and Hybrid CNN-Transformer (†). The best and second-best performances are indicated in **bold** and underline, respectively.

Method	Venue	Params ↓ (M)	FLOPs ↓ (G)	QaTa-COV19		MosMedData+	
				Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
U-Net$^\diamond$ [14]	MICCAI'15	14.8	50.3	79.02	69.46	64.60	50.73
U-Net++$^\diamond$ [22]	MICCAI'18	74.5	94.6	79.62	70.25	71.75	58.39
nnUNet$^\diamond$ [9]	Nature'21	19.1	412.7	80.42	70.81	72.59	60.36
Swin-UNet† [2]	ECCV'22	82.3	67.3	78.07	68.34	63.29	50.19
UCTransNet† [16]	AAAI'22	65.6	63.2	79.15	69.60	65.90	52.69
MedSA¶ [18]	ArXiv'23	104.3	55.2	82.77	70.60	77.65	63.47
GLoRIA† [7]	ICCV'21	45.6	60.8	79.94	70.68	72.42	60.18
ViLT† [10]	ICML'21	87.4	55.9	79.63	70.12	72.36	60.15
LAVT† [19]	CVPR'22	118.6	83.8	79.28	69.89	73.29	60.41
TGANet$^\diamond$ [15]	MICCAI'22	19.8	41.9	79.87	70.75	71.81	59.28
Ariadne's Thread† [20]	MICCAI'23	44.0	22.4	89.78	81.45	77.75	63.60
LViT† [11]	IEEE TMI'23	29.7	54.1	83.66	75.11	74.57	61.33
LGA¶ [6]	MICCAI'24	**8.24**	381.1	84.65	76.23	75.63	62.52
RefSegformer† [17]	IEEE TIP'24	195.0	103.6	84.09	75.48	74.98	61.70
RecLMIS† [8]	IEEE TMI'24	23.7	24.1	85.22	77.00	77.48	**65.07**
ViTexNet (Ours)†		37.7	**11.5**	**90.76**	**83.25**	**78.19**	64.04

3.3 Comparison with State-of-the-Art Models

ViTexNet is evaluated against widely used uni-modal approaches and the latest multi-modal segmentation methods, ensuring fair comparison by employing publicly available source codes (or re-implementations) with consistent hyperparameters and preprocessing. As shown in Table 1, ViTexNet surpasses all baselines in FLOPs. Qualitative results in Figure 2 further illustrate ViTexNet's robustness, demonstrating superior segmentation performance across both chest X-ray and CT datasets.

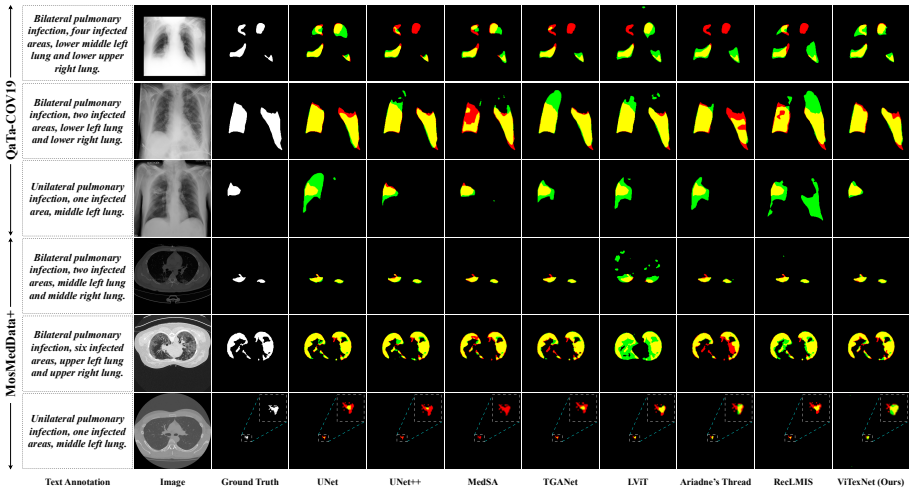

Fig. 2. Segmentation visualizations on QaTa-COV19 (top three examples) and MosMedData+ (bottom three examples) dataset. The overlays use yellow for true positives, red for false negatives, and green for false positives. The final row includes a dashed box highlighting a zoomed-in lesion region, offering a closer look at the model's segmentation accuracy.

3.4 Ablation Study

We explore the impact of replacing TGDC with cross attention or a self+cross attention pipeline (self-attention on image tokens, followed by cross-attention with text). Both alternatives increase the parameter count, yet TGDC achieves better metrics overall. Table 2 details these findings, while Figure 3 provides a qualitative comparison, further highlighting the advantages of our proposed module.

Table 2. Impact of Attention.

Method	Params ↓ (M)	QaTa-COV19		MosMedData+	
		Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
Self & Cross Attention	44.4	90.40	82.49	77.73	63.44
Cross Attention	39.8	89.24	81.22	77.25	62.41
TGDC	**37.7**	**90.76**	**83.25**	**78.19**	**64.04**

We evaluate the impact of text prompt granularity by splitting each annotation into Part A and Part B, each providing positional details at different levels. As shown in Table 3 and Figure 4.

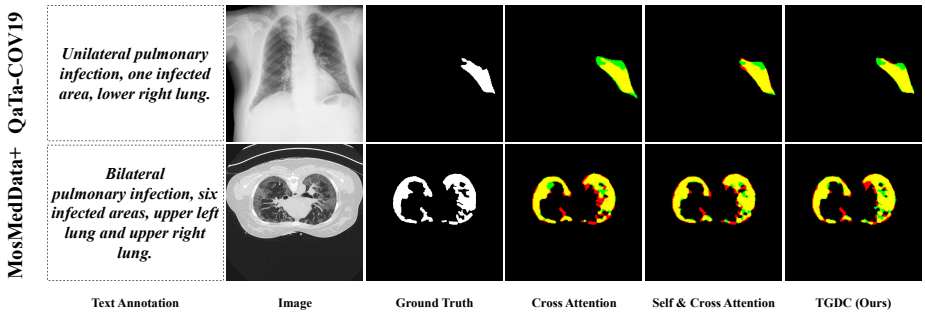

Fig. 3. Segmentation visualization comparing cross attention, self+cross attention, and TGDC on QaTa-COV19 (top row) and MosMedData+ (bottom row). Overlays use yellow for true positives, red for false negatives, and green for false positives.

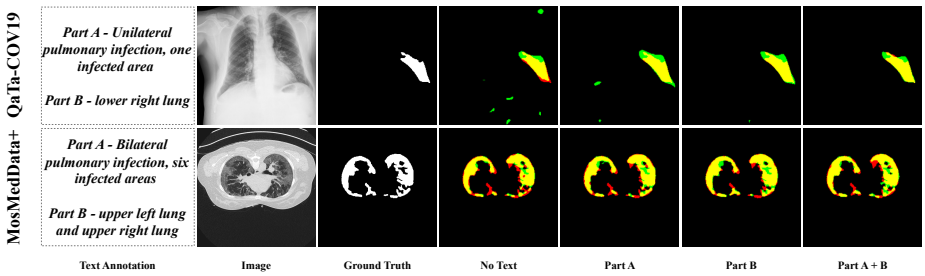

Fig. 4. Segmentation visualization with No Text, Part A, Part B, and Part A + B on QaTa-COV19 (top row) and MosMedData+ (bottom row). Overlays use yellow for true positives, red for false negatives, and green for false positives.

Table 3. Impact of text annotations

Text Parts	QaTa-COV19		MosMedData+	
	Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
No Text	71.98	56.23	74.54	59.42
Part A	87.82	78.47	76.21	61.49
Part B	90.25	82.40	76.78	62.21
Part A + B	**90.76**	**83.25**	**78.19**	**64.04**

4 Conclusion

In this paper, ViTexNet is proposed as an efficient multimodal segmentation approach that integrates textual and visual features through a Text-Guided Dynamic Convolution (TGDC) mechanism, rather than standard cross-attention.

Validation on MosMedData+ (CT) and QaTa-COV19 (X-Ray) datasets demonstrates its applicability across different imaging modalities. Experimental results show that ViTexNet outperforms state-of-the-art uni-modal and multi-modal methods in terms of Dice and IoU, while operating at a notably low FLOP count (11.5G), reflecting a strong accuracy–efficiency trade-off. Further ablation studies confirm TGDC's advantage over both cross-attention and combined self- and cross-attention blocks, and highlight the benefits of detailed text prompts. Future work will extend ViTexNet to additional imaging modalities, explore 3D segmentation, and evaluate it on a broader range of medical datasets to enhance generalizability.

Disclosure of Interests

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

References

1. Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *17th European Conference on Computer Vision (ECCV)*, pages 1–21. Springer, 2022.
2. Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *17th European Conference on Computer Vision (ECCV)*, pages 205–218. Springer, 2022.
3. Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis (MedIA)*, 79:102444, 2022.
4. Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *37th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11030–11039, 2020.
5. Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *29th IEEE International Conference on Image Processing (ICIP)*, pages 2306–2310. IEEE, 2022.
6. Jihong Hu, Yinhao Li, Hao Sun, Yu Song, Chujie Zhang, Lanfen Lin, and Yen-Wei Chen. Lga: A language guide adapter for advancing the sam model's capabilities in medical image segmentation. In *27th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 610–620. Springer, 2024.

7. Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *18th IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3942–3951, 2021.
8. Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 2024.
9. Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.
10. Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *38th International Conference on Machine Learning (ICML)*, pages 5583–5594. PMLR, 2021.
11. Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging (IEEE TMI)*, 43(1):96–107, 2023.
12. Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *35th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.
13. Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.
14. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.
15. Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *25th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 151–160. Springer, 2022.
16. Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *36th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)*, volume 36, pages 2441–2449, 2022.
17. Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *IEEE Transactions on Image Processing (IEEE TIP)*, 2024.
18. Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023.

19. Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *35th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18155–18165, 2022.
20. Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *26th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 724–733. Springer, 2023.
21. Xichuan Zhou, Qianqian Song, Jing Nie, Yujie Feng, Haijun Liu, Fu Liang, Lihui Chen, and Jin Xie. Hybrid cross-modality fusion network for medical image segmentation with contrastive learning. *Engineering Applications of Artificial Intelligence (EAAI)*, 144:110073, 2025.
22. Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *21st International Conference on Medical Image Computing and Computer-Assisted Intervention Workshops (MICCAIw)*, pages 3–11. Springer, 2018.