Visual-Textual Matching Attention for Lesion Segmentation in Chest Images

Phuoc-Nguyen Bui[1], Duc-Tai Le[2], and Hyunseung Choo[2*(✉)]

[1] Department of AI Systems Engineering
[2] Department of Electrical and Computer Engineering
Sungkyunkwan University, Korea
{phuocnguyen,ldtai,choo}@skku.edu

Abstract. Lesion segmentation in chest images is crucial for AI-assisted diagnostic systems of pulmonary conditions. The multi-modal approach, which combines image and text description, has achieved notable performance in medical image segmentation. However, the existing methods mainly focus on improving the decoder using the text information while the encoder remains unexplored. In this study, we introduce a Multi-Modal Input UNet model, namely MMI-UNet, which utilizes visual-textual matching (VTM) features for infected areas segmentation in chest X-ray images. These VTM features, which contain visual features that are relevant to the text description, are created by a combination of self-attention and cross-attention mechanisms in a novel Image-Text Matching (ITM) module integrated into the encoder. Empirically, extensive evaluations on the QaTa-Cov19 and MosMedData+ datasets demonstrate MMI-UNet's state-of-the-art performance over both uni-modal and previous multi-modal methods. Furthermore, our method also outperforms the best uni-modal method even with 15% of the training data. These findings highlight the interpretability of our vision-language model, advancing the explainable diagnosis of pulmonary diseases and reducing the labeling cost for segmentation tasks in the medical field. The source code is made publicly available at https://github.com/nguyenpbui/MMI-UNet.git.

Keywords: Chest X-ray · Chest CT · Cross-attention mechanism · Medical image segmentation · Multi-modal learning

1 Introduction

Chest imaging modalities, such as X-ray and CT scans, play a pivotal role in diagnosing and monitoring a diverse range of lung conditions, encompassing infectious diseases and neoplastic disorders. The advent of deep learning has spurred the utilization of deep neural networks for analyzing radiological images in support of various tasks related to assisted diagnosis, including disease classification, lesion detection, and segmentation. Among these tasks, lesion segmentation is crucial, as it facilitates the accurate identification and delineation of pathological regions within the thorax. Existing medical image segmentation

methods [2, 4, 23], built upon the UNet architecture [14], have achieved significant results. However, these methods often require a large amount of annotated data for training, which remains a significant obstacle in the medical field due to the high labeling cost and time associated with expert annotation.

Therefore, leveraging accompanying textual data from medical notes offers a compelling opportunity. Unlike acquiring and annotating entirely new data sources, text data from medical records is often readily available alongside the corresponding images, eliminating the need for extra costs associated with data collection, which can be a major bottleneck in medical research and development. Furthermore, the inherent complementary nature of the medical text and image data allows the textual information to enrich and potentially compensate for limitations in the visual data, leading to improved performance for image analysis tasks such as segmentation, classification, and disease diagnosis.

Following the groundbreaking work of CLIP [13] in 2021, which utilized 4 million image-text pairs for contrastive learning, the field of multi-modal learning has drawn extensive attention. This approach extends beyond computer vision and into the medical field, where researchers are applying vision-language pre-training and processing to downstream tasks such as classification and segmentation. For instance, Tomar et al. [15] proposed a Text-Guided Attention method that enables the model to learn additional, case-specific feature representations for polyp segmentation. Similarly, Li et al. [10] introduced LViT, a hybrid CNN-Transformer architecture that fuses image and text features to segment infected regions in chest X-ray images. Recently, GuideDecoder [22] achieved state-of-the-art (SOTA) performance on the QaTa-COV19 dataset [5] by implementing a novel approach that focuses on improving the decoder using both image and text features.

This work presents Multi-Modal Input UNet or MMI-UNet, a novel multi-modal learning approach for lesion segmentation in chest images. MMI-UNet integrates visual and linguistic features through a newly developed Image-Text Matching (ITM) module. This module leverages self-attention and cross-attention mechanisms to generate visual-textual matching (VTM) features, capturing visual elements relevant to the accompanying text descriptions. Experimental results on the QaTa-COV19 and MosMedData+ datasets demonstrate that MMI-UNet surpasses SOTA uni-modal and multi-modal methods. Notably, MMI-UNet also outperforms the best uni-modal method even with limited training data, highlighting its potential to reduce the need for extensive and expensive data labeling in medical image segmentation.

2 Method

As illustrated in Fig. 1, the proposed method leverages a UNet architecture [14] with a multi-modal encoder and a segmentation decoder. The encoder progressively extracts features from both the image and the accompanying text description and then combines them via a novel Image-Text Matching (ITM) module from levels 1 to 4. The visual-textual matching features from the ITM module are

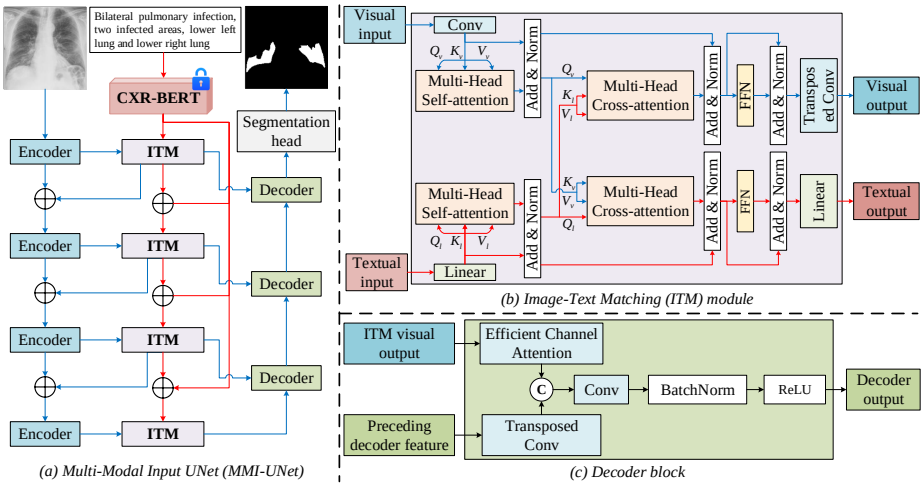

Fig. 1: The overview of the proposed Multi-Modal Input UNet (a) with Image-Text Matching module (b) and decoder block (c).

then passed through the decoder block at each level to produce the segmentation output. We elaborate on these components in the following sections.

2.1 Multi-modal encoder

Given an input image in the size of $H \times W \times 3$, we employ the tiny version of ConvNeXt [11] as the image encoder to extract multiple visual features from the last four levels, which are denoted as F_4, F_3, F_2, F_1 in the size of $\frac{H}{4} \times \frac{W}{4} \times 96$, $\frac{H}{8} \times \frac{W}{8} \times 192$, $\frac{H}{16} \times \frac{W}{16} \times 384$, and $\frac{H}{32} \times \frac{W}{32} \times 768$, respectively. For the corresponding text description, we adopt the pre-trained CXR-BERT [1] as the text encoder to extract the textual features, denoted as T. These textual features have a dimension of $L \times C$, where C represents the dimensionality of the extracted features and L signifies the length (number of tokens) of the text description. We set the value of C to 768 and freeze the text encoder during training.

Image-Text Matching (ITM) module: Leveraging both the image and its accompanying text description, the encoder aims to extract features that are optimally suited for segmentation tasks. This is accomplished by jointly interpreting the visual and textual information through the Image-Text Matching (ITM) module, as shown in Fig. 1 (b). To ensure computational efficiency, we first reduce the dimensionality of both the visual and textual features. This is achieved via a 2×2 convolutional layer ($Conv$) with the stride of 2 for the visual features and a linear layer ($Linear_1$) for the textual features. The process is described as follows:

$$F_i^c = LayerNorm(Conv(F_i))$$
$$T_i = LayerNorm(Linear_1(T_{i-1}^{ITM} + T))$$

Inspired by [16], we incorporate both multi-head self-attention ($MHSA$) for each modality and multi-head cross-attention ($MHCA$) within the ITM module. $MHSA$ allows each modality, i.e. image and text, to attend to its features, capturing internal relationships and dependencies. $MHCA$, on the other hand, enables the interaction between features from different modalities. In the $MHCA$ operation, for each feature from either modality, we treat it as the query (Q) while the keys (K) and values (V) used for attention are obtained from the other modality. The process is described as:

$$
\begin{aligned}
F_i^{sa} &= LayerNorm(MHSA(F_i^c) + F_i^c) \\
T_i^{sa} &= LayerNorm(MHSA(T_i) + T_i) \\
F_i^{ca} &= LayerNorm(MHCA(F_i^{sa}, T_i^{sa}) + F_i^{sa}) \\
T_i^{ca} &= LayerNorm(MHCA(T_i^{sa}, F_i^{sa}) + T_i^{sa})
\end{aligned}
$$

Afterward, the visual-textual matching features from each modality are then passed through a feed-forward network (FFN), followed by a layer normalization ($LayerNorm$). We use a transposed convolution ($ConvT$) and linear layer ($Linear_2$) to upsample the size of visual and textual features to match the size of the original features, respectively, described as follows:

$$
\begin{aligned}
F_i^{ITM} &= ConvT(LayerNorm(FFN(F_i^{ca}) + F_i^{ca})) \\
T_i^{ITM} &= Linear_2(LayerNorm(FFN(T_i^{ca}) + T_i^{ca}))
\end{aligned}
$$

2.2 Segmentation decoder

Within each decoder level, the visual output from the ITM module and the preceding decoder feature are combined. The former is fed into an efficient channel attention (ECA) module [18] while the latter is upsampled via a transposed convolution layer. These features are then channel-wise concatenated before being processed by a 3×3 convolutional layer. This is followed by batch normalization and a ReLU activation function, as shown in Fig. 1 (c). Finally, in the segmentation head, the preceding decoder feature is upsampled to match the original input image resolution. A 1×1 convolution and a sigmoid activation function are then applied to generate the segmentation output.

3 Experiments and Results

3.1 Datasets, evaluation metrics and implementation

Datasets: To evaluate the performance of our proposed MMI-UNet model, we conduct experiments on two publicly available datasets: QaTa-COV19 and MosMedData+. The first dataset, compiled by a collaborative effort between researchers from Qatar University and Tampere University, comprises 9258 chest X-ray images featuring COVID-19 cases. The second dataset, MosMedData+, consists of 2729 CT scan slices specifically capturing lung infections. Notably,

both datasets share similar text annotations, focusing on key clinical aspects like the presence of infection in both lungs, the number of lesions, and their approximate locations. These annotations are shown in Fig. 1(a) for further reference.

Evaluation metrics: We employ the Dice coefficient and Jaccard coefficient to quantitatively evaluate the segmentation performance. Both metrics assess the overlap between the predicted segmentation mask and the ground truth mask by calculating the ratio of the intersection area to the union area. Notably, the Dice coefficient is sensitive to the accurate segmentation of small objects.

$$DSC = \frac{2\left|GT \cap PR\right|}{\left|GT\right| + \left|PR\right|} = \frac{2 \times TP}{2 \times TP + FP + FN}$$
$$IoU = \frac{\left|GT \cap PR\right|}{\left|GT \cup PR\right|} = \frac{TP}{TP + FP + FN}$$

Implementation: Following [10], we split the QaTa-COV19 dataset into training, validation, and testing sets with 5716, 1429, and 2113 samples, respectively. the MosMedData+ dataset is divided into a training set with 2183 images, a validation set with 273 images and a testing set with 273 samples. All images are cropped to the size of 224 × 224. Data augmentation is employed using a random zoom technique with a probability of 10%. PyTorch [7], PyTorch Lightning, MONAI [3], and Transformers [19] are utilized for implementation. The entire training and testing procedure is conducted on a single NVIDIA A100 80 GB VRAM GPU. To train the model, we employ a combined loss function consisting of Dice loss and cross-entropy loss, and the network is optimized using the AdamW optimizer with a batch size of 32. A cosine annealing learning rate schedule is employed, starting at 3e-4 and decreasing to 1e-6.

3.2 Performance comparison with existing methods

We compare MMI-UNet against commonly used mono-modal and the latest multi-modal medical image segmentation methods. For a fair comparison, we employ publicly available source codes or re-implement based on the corresponding papers and then apply the same hyperparameters and preprocessing techniques. As presented in Table 1, MMI-UNet surpasses all evaluated methods on both the QaTa-COV19 and MosMedData+ datasets. Specifically, compared to nnUNet, our method achieves significant improvements of 10.46% and 5.83% DSC, respectively, across both datasets. In comparison with other multi-modal methods such as LViT [10] and GuideDecoder [22], the proposed method demonstrates superior performance, with improvements of 7.22% and 1.1% DSC on the QaTa-COV19 dataset and 3.85% and 0.67% DSC on the MosMedData+ dataset.

We qualitatively demonstrate the results of the proposed MMI-UNet and other methods on the QaTa-COV19 and MosMedData+ in figures 2 and 3, respectively. MMI-UNet exhibits a significant reduction in mis-segmentation compared to other SOTA methods. This improvement is attributed to the effective integration of textual information with image data in the encoder via the ITM

Table 1: Quantitative comparison on segmentation results with uni-modal and previous multi-modal learning methods. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Type	Param ↓ (M)	FLOPs ↓ (G)	QaTa-COV19		MosMedData+	
				DSC ↑	IoU ↑	DSC ↑	IoU ↑
UNet [14]	CNN	**14.8**	50.3	79.02	69.46	64.60	50.73
UNet++ [23]	CNN	74.5	94.6	79.62	70.25	71.75	58.39
AttUNet [12]	CNN	34.9	101.9	79.31	70.04	66.34	52.82
nnUNet [12]	CNN	19.1	412.7	80.42	70.81	72.59	60.36
TransUNet [4]	Hybrid	105	56.7	78.63	69.13	71.24	58.44
UCTransNet [17]	Hybrid	65.6	63.2	79.15	69.60	65.90	52.69
Swin-UNet [2]	Hybrid	82.3	67.3	78.07	68.34	63.29	50.19
CLIP [13]	Hybrid	87.0	105.3	79.81	70.66	71.97	59.64
GLoRIA [6]	Hybrid	45.6	60.8	79.94	70.68	72.42	60.18
ConVIRT [21]	CNN	35.2	44.6	79.72	70.58	72.06	59.73
TGANet [15]	CNN	19.8	41.9	79.87	70.75	71.81	59.28
ViLT [9]	Hybrid	87.4	55.9	79.63	70.12	72.36	60.15
LAVT [20]	Hybrid	118.6	83.8	79.28	69.89	73.29	60.41
LViT [10]	Hybrid	29.7	54.1	83.66	75.11	74.57	61.33
GuideDecoder [22]	Hybrid	44.0	22.4	89.78	81.45	77.75	63.60
MMI-UNet	Hybrid	56.2	**22.1**	**90.88**	**83.28**	**78.42**	**64.50**

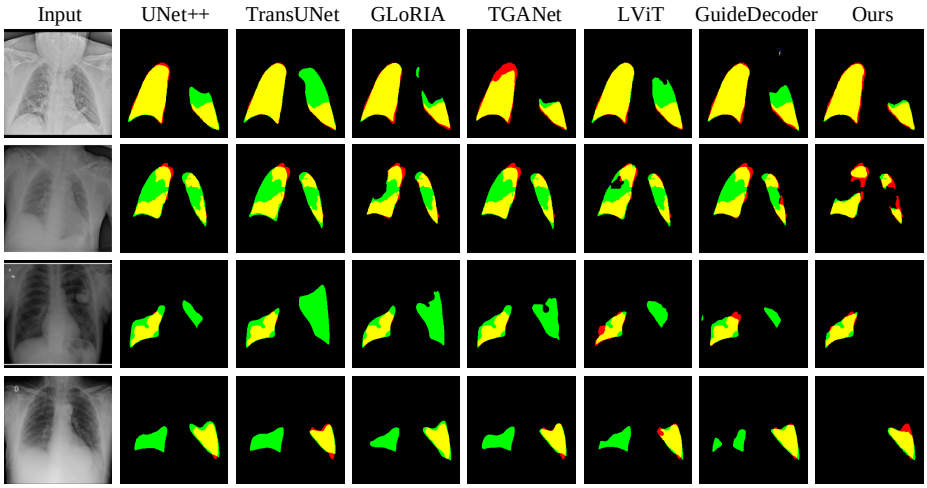

Fig. 2: Segmentation visualization on the QaTa-COV19 dataset. Yellow, red, and green indicate true positive, false negative, and false positive, respectively.

module, enabling our model to accurately segment infected areas while filtering out irrelevant regions, as shown in the last two rows of Fig. 2. In particular, the segmentation output of MMI-UNet in the third row precisely aligns with the text description: "**Unilateral pulmonary infection, one infected area, middle lower left lung**." while other methods over-segment the right side of the image, misidentifying it as the infected area mentioned in the text description.

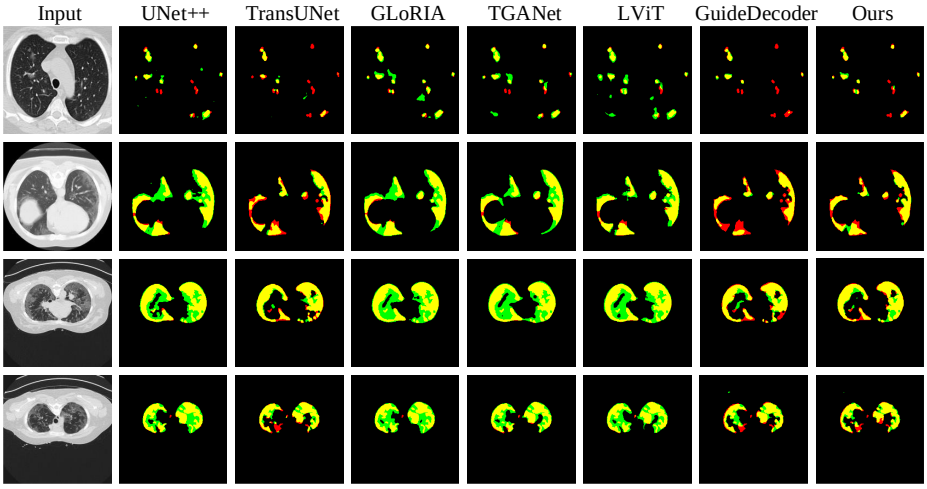

Fig. 3: Segmentation visualization on the MosMedData+ dataset. Yellow, red, and green indicate true positive, false negative, and false positive, respectively.

3.3 Ablation study

We further explore the impact of incorporating the existing GuideDecoder [22] into MMI-UNet, resulting in a variant termed MMI-UNet with GD, as detailed in Table 2. We strictly follow the implementations of GuideDecoder by replacing the encoder features with enhanced visual ones from the Image-Text Matching (ITM) module. However, despite this significant rise in model complexity, an increase of 9.6M parameters, the MMI-UNet with GD does not surpass the performance of the original MMI-UNet with the conventional CNN decoder. This finding suggests that the original MMI-UNet effectively balances performance and model efficiency, achieving SOTA results without requiring additional complexity.

We also perform extensive experiments to analyze the impact of varying training data sizes on the model's segmentation performance. As illustrated in Table 3, the proposed MMI-UNet achieves comparable performance with nnUNet even with limited training data. In particular, MMI-UNet achieves 8.31% and 1.41% higher DSC compared to nnUNet, the best mono-modal model trained on

Table 2: Impact of different decoder architectures on segmentation performance.

Method	Param ↓ (M)	Covid-19		MosMedData+	
		DSC ↑	IoU ↑	DSC ↑	IoU ↑
MMI-UNet w/ GD	65.8	90.36	82.46	78.38	64.45
MMI-UNet	**56.2**	**90.88**	**83.28**	**78.42**	**64.50**

the entire dataset when using only a quarter of the data. This finding emphasizes the advantages of multi-modal approaches, where the effective integration of text prompts significantly improves segmentation performance with limited data.

Table 3: Impact of the training data size on segmentation performance.

Method	Covid-19		MosMedData+	
	DSC ↑	IoU ↑	DSC ↑	IoU ↑
nnUNet [8] (100% training data)	80.42	70.81	72.59	60.36
MMI-UNet (15% training data)	88.72	79.73	74.00	58.73
MMI-UNet (25% training data)	89.37	80.79	75.40	60.51
MMI-UNet (50% training data)	90.04	81.89	76.27	61.64
MMI-UNet (100% training data)	**90.88**	**83.28**	**78.42**	**64.50**

4 Conclusion

This paper presents MMI-UNet, a novel multi-modal learning method for infected area segmentation in chest images. By integrating visual features from the image and textual information from its accompanying description through a newly designed Image-Text Matching (ITM) module, MMI-UNet leverages the power of both modalities. This innovative module employs self-attention and cross-attention mechanisms to generate visual-textual matching (VTM) features, capturing visual elements directly related to the specifics mentioned in the text descriptions. Extensive evaluations on the QaTa-COV19 and MosMedData+ datasets demonstrate that MMI-UNet achieves superior performance compared to both best-performing uni-modal and multi-modal methods. Notably, MMI-UNet even surpasses the best uni-modal method even with limited training data, signifying its potential to significantly reduce the need for extensive and expensive data labeling, a critical challenge in medical image segmentation. These findings not only showcase the effectiveness of MMI-UNet but also hint at its

potential interpretability, which is essential for the development of a trustworthy and explainable diagnostic system for pulmonary diseases.

Acknowledgments. This work was partly supported by IITP grant funded by the Korea government (MSIT) under ICT Creative Consilience program (IITP-2024-2020-0-01821, 50%), Artificial Intelligence Graduate School program (Sungkyunkwan University, 25%), and Artificial Intelligence Innovation Hub (No.RS-2021-II212068, 25%).

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

References

1. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
2. 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)
3. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)
4. 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)
5. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
6. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3942–3951 (2021)
7. Imambi, S., Prakash, K.B., Kanagachidambaresan, G.: Pytorch. Programming with TensorFlow: Solution for Edge Computing Applications pp. 87–104 (2021)
8. 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)
9. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
10. 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)
11. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

12. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arxiv 2018. arXiv preprint arXiv:1804.03999 (1804)
13. 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)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
15. 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)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
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. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2441–2449 (2022)
18. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
19. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. pp. 38–45 (2020)
20. 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)
21. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
22. 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)
23. 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)