MG-UNet: A Memory-Guided UNet for Lesion Segmentation in Chest Images

Shuaipeng Ding, Mingyong Li(✉) and Chao Wang

School of Computer and Information Science, Chongqing Normal University,
Chongqing 401331, China
limingyong@cqnu.edu.cn

Abstract. Lesion segmentation in medical images is a key task for the intelligent diagnosis of lung diseases. Although existing multimodal methods have achieved significant progress in medical image segmentation by combining image and text information, these methods still rely on textual input during the inference phase, limiting their applicability in real-world scenarios. To address this limitation, this paper proposes an innovative Memory-Guided UNet model (MG-UNet). MG-UNet introduces a learnable memory bank that automatically extracts and stores textual information during the training phase. In the decoding stage, the proposed memory-guided decoder retrieves knowledge relevant to the current image from the memory bank, thereby eliminating the need for textual input during inference. Extensive experiments were conducted on the QaTa-Cov19 and MosMedData+ datasets to validate the effectiveness of MG-UNet. The experimental results demonstrate that MG-UNet not only outperforms existing unimodal and multimodal methods in terms of segmentation performance but also excels in text-free inference scenarios using only 15% of the training data, surpassing the current best unimodal methods. This characteristic significantly reduces the reliance on annotated data for medical image segmentation, offering greater flexibility and scalability for practical clinical applications. The code will be available soon.

Keywords: Chest CT · Multi-modal · Memory Bank · Medical image segmentation

1 Introduction

Chest imaging modalities, such as X-rays and CT scans, play a crucial role in the diagnosis and monitoring of various pulmonary diseases, including infectious diseases and neoplastic disorders. With the rapid development of deep learning, deep neural networks have been widely applied to radiological image analysis, supporting auxiliary diagnostic tasks such as disease classification, lesion detection, and segmentation. Among these tasks, lesion segmentation is particularly critical, as it not only enables precise localization and delineation of pathological regions within the thorax but also provides quantitative foundations for disease staging and treatment planning, making it a key component of clinical diagnosis. Existing medical image segmentation methods [1–3], primarily based on the UNet architecture [4], have achieved significant progress. However, due to their reliance on large amounts of high-quality annotated data and the high cost and inefficiency associated with expert involvement in the annotation process, their widespread application in real clinical settings remains severely limited.

To mitigate similar issues available in natural image processing, CLIP [5] capitalizes on the complementary information provided by accompanying text, thereby reducing the dependency on high-quality annotated image data and maximizing the utilization of available information. Inspired by CLIP, approaches like MedCLIP [6] and GLoRIA [7] have extended these ideas to medical image-report pairs situations, still showcasing excellent performance. Motivated by the impressive performance gains achieved through the integration of textual information, Li et al. [8] proposed a novel CNN-Transformer structure named LViT to integrate multimodal information in the early stage. GuideDecoder [9] implemented a novel approach that focuses on improving the decoder using both image and text features. MMI-UNet [10] achieved state-of-the-art (SOTA) performance on the QaTa-COV19 [11] dataset by integrating visual and linguistic features during the encoder stage.

Although the above research has significantly improved segmentation performance by leveraging text, these multimodal methods require textual input during both training and inference, limiting their practicality in real clinical scenarios. To address this issue, we introduce a learnable Memory Bank to store textual information, serving as a bridge between textual and visual information. This allows us to enhance visual representation during inference by retrieving relevant historical information from the Memory Bank, eliminating the need for text input. Our contributions are as follows:

- We propose Memory-Guided UNet (MG-UNet), an innovative multimodal learning method for lesion segmentation in chest images, achieving excellent segmentation results during the inference phase without relying on text input.
- We introduce an intermittent memory bank updating (IMBU) strategy that allows the model to progressively transition from reliance on text input to operating without text input during the training phase, thereby achieving greater adaptability.
- Dual inference mechanism: The model dynamically updates the Memory Bank through corresponding text, utilizing the updated Memory Bank to extract knowledge relevant to the current image. In the absence of text input, the model can directly retrieve knowledge from the historical Memory Bank. This mechanism not only enhances the model's adaptability but also significantly improves the efficiency of knowledge utilization, thereby increasing its operational viability in clinical diagnosis.

2 Method

The overall architecture of our proposed Memory-Guided UNet model (MG-UNet) is illustrated in Fig. 1. The model consists of three main components: the image encoder, the Memory Bank Updating (MBU) mechanism, and the Memory-Guided Decoder (MG-Decoder). The image encoder encodes the input images into feature representations, while the MBU mechanism automatically learns and stores textual knowledge into the Memory Bank during training. Subsequently, the image features are processed by the MG-Decoder, which extracts knowledge relevant to the current image from the Memory Bank and fuses it with the encoder features. The iterative process of the MG-Decoder effectively utilizes different levels of image features and knowledge from the Memory Bank, resulting in a more efficient fused representation. Finally, a segmentation head is used to generate the final segmentation predictions.

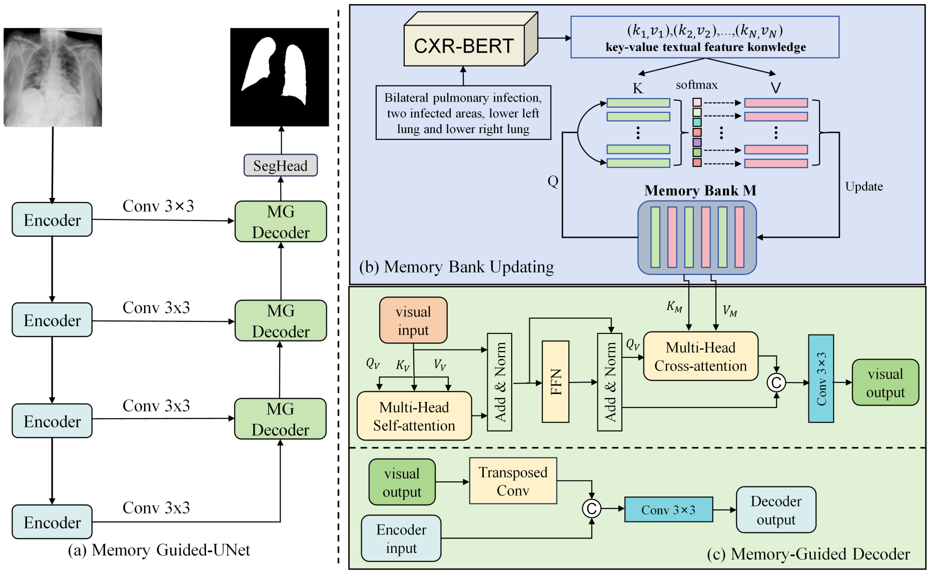

Fig. 1. The overview of the proposed (a) Memory-Guided Unet with (b) Memory Bank Updating and (c) Memory-Guided Decoder.

We will elaborate on these components in the following sections and introduce an intermittent memory bank updating strategy to enable the model to better adapt to the text-free inference process.

2.1 Image Encoder

We utilize ConvNeXt tiny [12] as the image encoder to extract multiple visual features from the input image. Given an input image of size $H \times W \times 3$, the encoder extracts feature maps with dimensions $\frac{H}{4} \times \frac{W}{4} \times 96$, $\frac{H}{8} \times \frac{W}{8} \times 192$, $\frac{H}{16} \times \frac{W}{16} \times 384$, and $\frac{H}{32} \times \frac{W}{32} \times 768$, respectively. These feature maps are then adjusted to a uniform number of channels through 3×3 convolution operations, set to 64 in this paper to balance model performance and efficiency. The resulting features are denoted as E_4, E_3, E_2, and E_1.

2.2 Memory Bank Updating

We first adopt the pre-trained CXR-BERT [13] as the text encoder to extract textual features, which are then reduced in dimensionality through a linear mapping, denoted as $\mathbf{T}$. The dimensionality of these textual features is $L \times C$, where C represents the dimensionality of the extracted features and L signifies the length (number of tokens) of the text description. We set C to 64 and freeze the text encoder during training.

Then we initialize the Memory Bank $\mathbf{M}_0 \in \mathbb{R}^{N_m \times D}$ through a matrix, where N_m is the base size and D is the dimension. To fully extract knowledge in the Memory Bank, during initialization, we set $N_m = L$ and $D = 64$, the diagonal elements of $\mathbf{M}_0$ are set to 1, while the other elements are set to 0. We update the Memory Bank using a multi-head attention (MHA) mechanism [14]:

$$Att_i(X, Y) = softmax\left(\frac{\mathrm{Q_X} \cdot \mathrm{K_Y}^T}{\sqrt{d_n}}\right) \cdot \mathrm{V_Y} \tag{1}$$

$$MHA(X, Y) = [Att_1(X, Y); \ldots ; Att_n(X, Y)] \tag{2}$$

where[;] stands for concatenation operation.

To update the Memory Bank $\mathbf{M}_{t-1}$ at the training step t-1, we use the textual features extracted from the corresponding text description to identify the knowledge missing in $\mathbf{M}_{t-1}$.

$$\varDelta\mathbf{M}_t = MHA\left(\mathbf{M}_{t-1}, \mathbf{T}\right) \tag{3}$$

where $\varDelta\mathbf{M}_t$ represents the incremental knowledge gained in the training step t, while T represents the textual features gained in this equation. By integrating incremental knowledge, we can update the Memory Bank during the training step.

$$\mathbf{M}_t = \mathbf{M}_{t-1} + Norm\left(\varDelta\mathbf{M}_t\right) \tag{4}$$

where $Norm$ refers to layer normalization to normalize the incremental knowledge.

2.3 Memory-Guided Decoder

For the decoder feature D_i, we first capture internal relationships and dependencies through multi-head self-attention and a feed-forward network (FFN). Then, we extract image-related knowledge from the Memory Bank using multi-head cross-attention. Finally, the fused features are refined through a 3×3 convolutional layer. The process can be described as follows:

$$D_i^{sa} = Norm\left(MHA\left(D_i, D_i\right) + D_i\right) \tag{5}$$

$$F_i^{sa} = Norm\left(FFN\left(D_i^{sa}\right) + D_i^{sa}\right) \tag{6}$$

$$F_i^{ca} = Norm\left(MHA\left(F_i^{sa}, \mathbf{M}_t\right)\right) \tag{7}$$

$$F_i = Conv_{3\times 3}\left[F_i^{ca}; F_i^{sa}\right] \tag{8}$$

Then, the processed features F_i are upsampled using a transposed convolution and concatenated with the feature E_{i+1} from the encoder, and finally processed using a 3×3 convolution to obtain the decoder features D_{i+1}, as shown in Fig. 1 (c).

After three iterations, the decoder feature is first upsampled to match the resolution of the original input image. Then, a 1×1 convolution and a sigmoid activation function are applied to generate the segmentation output.

2.4 Intermittent Memory Bank Updating

During the training phase, since text corresponding to the images is used to update the memory bank, directly utilizing a fixed-parameter memory bank for text-free inference may result in feature degradation, leading to inaccurate segmentation. To mitigate this potential issue, we propose a straightforward training strategy (intermittent memory bank updating). Specifically, we train MG-UNet over several rounds, each consisting of two epochs. In the first epoch, the memory bank is updated using the text associated with the images. In the second epoch, we simulate the inference phase by omitting the text update process and instead directly retrieving relevant historical information from the memory bank for the current image.

3 Experiments and Results

3.1 Datasets and implementation

Datesets: To evaluate the performance of our proposed MG-UNet model, we conducted experiments on two publicly available datasets: QaTa-COV19 [11] and MosMedData+ [15]. The first dataset, compiled through a collaborative effort between researchers at Qatar University and Tampere University, consists of 9258 chest X-ray images depicting COVID-19 cases. The second dataset, MosMedData+, includes 2729 CT scan slices specifically depicting lung infections. Notably, both datasets feature similar textual annotations that focus on key clinical aspects such as the presence of infection in both lungs, the number of lesions, and their approximate locations. These annotations are illustrated in Fig. 1(a) for reference.

Implementation: Following [8], we split the QaTa-COV19 dataset into training, validation, and testing sets, containing 5716, 1429, and 2113 samples, respectively. The MosMedData+ dataset is divided into a training set of 2183 images, a validation set of 273 images, and a testing set of 273 samples. All images are cropped to a size of 224 × 224. For data augmentation, a random zoom technique with a probability of 10% is applied. The implementation utilizes PyTorch [16], PyTorch Lightning, and MONAI [17]. The entire training and testing process is conducted on a Nvidia GeForce RTX 3090 with 24GB VRAM. During training, we use a combined loss function consisting of Dice loss and cross-entropy loss, with the network optimized using the AdamW optimizer and a batch size of 32. A cosine annealing learning rate schedule is employed, starting from 3e-4 and decaying to 1e-6.

3.2 Performance comparison with existing methods

We compared MG-UNet and the text-free reasoning model MG-UNet* with commonly used single-modal methods and the latest multi-modal medical image segmentation methods. As shown in Table 1, MG-UNet outperformed all evaluation methods on the QaTa-COV19 and MosMedData+ datasets. Notably, MG-UNet* improved the DSC (Dice Similarity Coefficient) by 7.68% and 3.80%, respectively, compared to the best single-modal method nnunet on the two datasets. It also demonstrated better performance compared to some multi-modal methods. Although there is still a gap compared to the latest multi-modal method MMI-UNet, but MG-UNet* achieved a roughly 50% reduction in FLOPs compared to MMI-UNet, and it does not rely on text input during the inference phase, providing greater flexibility and efficiency for practical clinical applications.

The qualitative experimental results compared with ground truth are shown in Fig. 2. We conducted qualitative comparisons between the proposed MG-UNet method and other approaches on the QaTa-COV19 and MosMedData+ datasets, respectively. In the visualization results, yellow indicates true positive regions (correctly identified infected areas), red denotes false negative regions (missed infections), and green represents false positive regions (normal tissues misclassified as infections). The experimental results demonstrate that due to limited feature extraction capabilities, traditional unimodal methods exhibit significant performance limitations in complex scenarios, leading to substantial false negative segmentation results (noticeable increase in red regions). Compared with other multimodal fusion methods, MG-UNet not only effectively suppresses over-segmentation phenomena (indicated by blue annotations) through its innovative feature fusion mechanism but, more importantly, maintains excellent true positive recognition capability. This advantage enables our model to accurately segment infected regions while effectively filtering out irrelevant areas.

3.3 Ablation study

To validate the effectiveness of the proposed module, we conducted an ablation study on the QaTa-COV19 dataset. As shown in Table 2, compared to the

Table 1. Quantitative comparison on segmentation results with uni-modal and previous multi-modal learning methods. optimal and suboptimal performance is highlighted, ↑(↓) denotes the higher (lower) the better. MG-UNet results are averaged over five runs.

Method	Params↑	FLOPs↓	QaTa-COV19		MosMedData+	
			DSC↑	IoU↑	DSC↑	IoU↑
UNet[4]	14.8M	50.3G	79.02	69.46	64.60	50.73
UNet++ [3]	74.5M	94.6G	79.62	70.25	71.75	58.39
AttUNet [18]	34.9M	101.9G	79.31	70.04	66.34	52.82
TransUNet [2]	105M	56.7G	78.63	69.13	71.24	58.44
UCTransNet [19]	65.6M	63.2G	79.15	69.60	65.90	52.69
Swin-UNet [1]	82.3M	67.3G	78.07	68.34	63.29	50.19
nnUNet [20]	19.1M	412.7G	80.42	70.81	72.59	60.36
CLIP [5]	87.0M	105.3G	79.81	70.66	71.97	59.64
LAVT [21]	118.6M	83.8G	79.28	69.89	73.29	60.41
LViT [8]	29.7M	54.1G	83.66	75.11	74.57	61.33
GuideDecoder [9]	44.0M	22.4G	89.78	81.45	77.75	63.60
TGCAM [22]	-	-	90.60	82.81	77.82	63.69
MMI-UNet [10]	56.2M	22.1G	90.88	83.28	78.42	64.50
MAdapter [23]	-	-	90.22	82.16	78.62	64.78
MG-UNet	30.5M	16.19G	90.90	83.30	78.65	64.80
MG-UNet*	30.5M	10.95G	88.10	77.80	76.39	61.79

baseline model that uses only images, MG-UNet achieved 6.72% improvement in Dice Similarity Coefficient (DSC), confirming the effectiveness of the Memory-Guided Decoder. MG-UNet* showed 3.72% decrease in DSC compared to MG-UNet, likely due to the model's reliance on a fixed-parameter memory bank, which resulted in knowledge confusion and affected segmentation accuracy. By incorporating our proposed training strategy, the model gradually transitioned from text-based training to text-free inference, reducing the knowledge confusion caused by the rigid memory bank, leading to 0.92% improvement in DSC scores.

Furthermore, we performed extensive experiments to assess the impact of varying training data sizes on the model's segmentation performance. As illustrated in Table 3, even with limited training data, MG-UNet demonstrated comparable performance to nnUNet. Specifically, when trained with only 15% of the dataset, MG-UNet and MG-UNet* achieved 7.60% and 4.88% higher DSC scores, respectively, compared to nnUNet, the best performing unimodal model trained on the full dataset.

Table 2. The ablation study of the QaTa-COV19 test set. "w/o text" indicates that there is no text, and the model uses only the UNet decoder. "w/ text" means that text is used during both the training and inference phases, referring to MG-UNet. "Test w/o text" indicates that text is used only during the training phase, referring to MG-UNet*. TS represents the IMBU strategy we proposed during the training phase.

Model		QaTa-COV19	
		DSC↑	IoU↑
w/o text	Baseline	84.18	72.62
w/ text	MG-UNet	90.90	83.30
Test w/o text	MG-UNet*	87.18	77.27
	MG-UNet*+TS	88.10	77.80

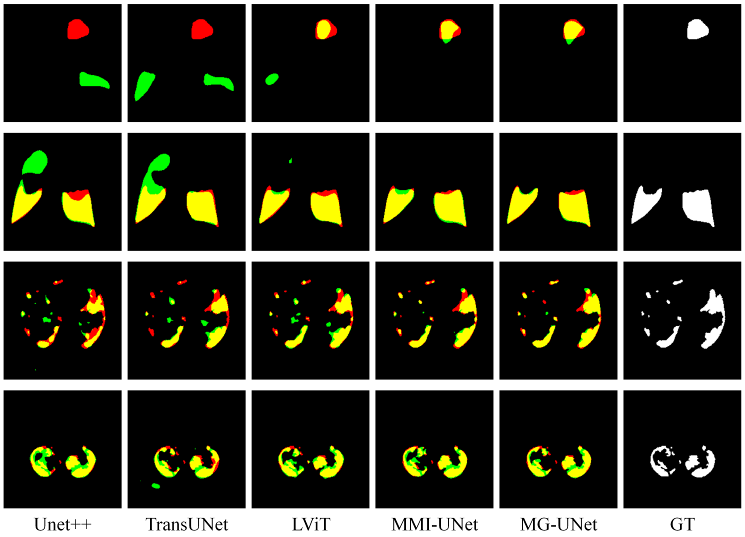

Fig. 2. Qualitative results on the QaTa-COV19 dataset and the MosMedData+ dataset. Yellow, red, and green indicate true positive, false negative, and false positive, respectively.

Table 3. Impact of the training data size on segmentation performance. The best results are shown in bold.

Method	DSC↑	IoU↑
nnUNet (100% training data)	80.42	70.81
MG-UNet (15% training data)	88.02	78.60
MG-UNet (25% training data)	88.91	80.03
MG-UNet (50% training data)	89.94	81.87
MG-UNet (100% training data)	**90.90**	**83.30**
MG-UNet* (15% training data)	85.30	74.37

4 Conclusion

This paper proposes the MG-UNet model for the segmentation of infected regions in chest images. By designing a learnable memory bank, we establish a bridge between visual and textual information, ensuring that the model can perform inference and maintain segmentation performance even in the absence of textual input. Extensive evaluations on the QaTa-COV19 and MosMedData+ datasets demonstrate that MG-UNet outperforms the best-performing unimodal and multimodal methods. Notably, even with limited training data during the text-free inference phase, MG-UNet surpasses the best unimodal method, highlighting its potential to significantly reduce the reliance on extensive data annotation while offering greater flexibility and scalability for practical applications.

Acknowledgements. This work was supported by the Nature Science Foundation Project of Chongqing Science and Technology Bureau (No. 2024NSCQ-MSX3120).

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

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)
6. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
7. 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)

8. Li Z, Li Y, Li Q, et al.: Lvit: language meets vision transformer in medical image segmentation. IEEE Transactions on Medical Imaging, vol. 43, no. 1, 96-107 (2023)
9. 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)
10. Bui, Phuoc-Nguyen, Duc-Tai Le, and Hyunseung Choo. "Visual-Textual Matching Attention for Lesion Segmentation in Chest Images." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.
11. 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)
12. 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)
13. 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)
14. Vaswani, A., Shazeer, N. , Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Pro- cessing Systems 30 (2017)
15. Morozov S P, Andreychenko A E, Pavlov N A, et al.: Mosmeddata: Chest ct scans with covid- 19 related findings dataset. arXiv preprint arXiv:2005.06465. (2020)
16. Imambi, S., Prakash, K.B., Kanagachidambaresan, G.: Pytorch. Programming with TensorFlow: Solution for Edge Computing Applications pp. 87 104 (2021)
17. 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)
18. 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)
19. 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) (2021)
20. 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)
21. 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)
22. Guo, Yunpeng, et al. "Common Vision-Language Attention for Text-Guided Medical Image Segmentation of Pneumonia." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.
23. Zhang, Xu, et al. "MAdapter: A Better Interaction Between Image and Language for Medical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.