LM-UNet: Whole-body PET-CT Lesion Segmentation with Dual-Modality-based Annotations Driven by Latent Mamba U-Net

Anglin Liu[1,3], Dengqiang Jia[4], Kaicong Sun[1], Runqi Meng[1], Meixin Zhao[5], Yongluo Jiang[6], Zhijian Dong[7], Yaozong Gao[3](✉), and Dinggang Shen[1,2,3](✉)

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
{liual2022,dgshen}@shanghaitech.edu.cn
[2] Shanghai Clinical Research and Trial Center, Shanghai, China
[3] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
yaozong.gao@uii-ai.com
[4] Hong Kong Centre for Cerebro-cardiovascular Health Engineering, Hong Kong, China
[5] Department of Nuclear Medicine, Peking University Third Hospital, Beijing, China
[6] Department of Nuclear Medicine, Sun Yat-sen University Cancer Center, Guangzhou, China
[7] Department of Nuclear Medicine, Xi'an Gaoxin Hospital, Xi'an, China

Abstract. PET-CT integrates metabolic information with anatomical structures and plays a vital role in revealing systemic metabolic abnormalities. Automatic segmentation of lesions from whole-body PET-CT could assist diagnostic workflow, support quantitative diagnosis, and increase the detection rate of microscopic lesions. However, automatic lesion segmentation from PET-CT images still faces challenges due to 1) limitations of single-modality-based annotations in public PET-CT datasets, 2) difficulty in distinguishing between pathological and physiological high metabolism, and 3) lack of effective utilization of CT's structural information. To address these challenges, we propose a threefold strategy. First, we develop an in-house dataset with dual-modality-based annotations to improve clinical applicability; Second, we introduce a model called Latent Mamba U-Net (LM-UNet), to more accurately identify lesions by modeling long-range dependencies; Third, we employ an anatomical enhancement module to better integrate tissue structural features. Experimental results show that our comprehensive framework achieves improved performance over the state-of-the-art methods on both public and in-house datasets, further advancing the development of AI-assisted clinical applications. Our code is available at https://github.com/Joey-S-Liu/LM-UNet.

Keywords: PET-CT · Lesion Segmentation · Annotations.

1 Introduction

Positron Emission Tomography-Computed Tomography (PET-CT) combines the metabolic activity provided by PET with the anatomical details provided by CT, offering desirable diagnostic capability in medical imaging [3]. The lesion regions can be distinctly marked by employing ^{18}F-FDG as the PET tracer, thereby enabling the identification of tumors, inflammation, and infections. However, given the complexity of imaging, there emerges a necessity to design an automatic framework for 3D whole-body lesion segmentation based on PET-CT.

With the advancement of deep learning, numerous approaches have emerged for segmenting whole-body lesions from PET-CT scans. U-Net [17] and its variations [1,7,8], alongside specialized models [11,18] specifically designed for PET-CT lesion segmentation, have demonstrated remarkable efficacy. Notably, Shi *et. al.* [18] introduced a Transformer-based multi-path parallel embedding module for tumor segmentation, effectively harnessing the complementary information provided by PET and CT modalities. However, despite these advancements, the performance of these approaches is still constrained, primarily due to 1) single-modality-based annotation in publicly available PET-CT datasets, 2) challenges in discerning between pathological and physiological high metabolism, and 3) under-exploitation of the advantages offered by the "PET+CT" mode over the "PET only" mode. All these challenges will be explained below.

Firstly, current public PET-CT datasets only have lesion masks annotated on one modality, which proves inadequate for accurate tumor segmentation. For instance, bone metastases and early-stage liver or spleen lesions may exhibit clear abnormalities on PET scans while showing no structural changes on CT scans. Relying solely on one modality for annotation in such scenarios can yield misleading interpretations. Thus, dual-modality-based annotation is essential to better cater to clinical requirements.

Secondly, distinguishing pathological from physiological high metabolism poses a challenge due to the complexity of human metabolic activities. Pathological metabolic regions, like tumors, often closely border normal physiological metabolic regions, complicating the accurate identification of abnormalities. Long-range modeling techniques, such as state space sequence models (SSMs) [10] and structured state space sequence models (S4) [6], have shown promise in extracting effective information from noise which offer significant implications for distinguishing pathological metabolism from surrounding physiological metabolism. However, while Mamba [5] significantly improves concentration on pertinent information compared to S4, its current applications in medical imaging [13,21] lack emphasis on crucial high-level features, such as identifying pathological characteristics. Thus, employing a hybrid CNN-Mamba structure allows us to leverage Mamba's capabilities to focus specifically on the high-level features of PET-CT in the latent space.

Thirdly, the advantages of the "PET+CT" mode remain under-exploited. Current segmentation methods often heavily rely on PET due to its high intensity towards tumor regions, while the potential contribution of CT features is often under-explored due to ineffective structural constraints in CT data.

To tackle these challenges, our paper introduces a dual-modality-based annotation strategy alongside a novel network (LM-UNet) tailored for whole-body lesion segmentation from PET-CT.

Our contributions can be outlined as follows:

- We propose a dual-modality-based annotation method for our in-house 3D whole-body PET-CT dataset, which shows promising performance for clinical application.
- We develop a novel multi-task hybrid CNN-Mamba network, called **LM-UNet**, which can model long-range dependencies on high-level features in the latent space. Besides, we introduce an anatomical enhancement module to better constrain anatomical shape of lesions.
- Experimental results show that our method obtains superior performance on both public and in-house datasets over the representative methods, especially its clinical practicality on our in-house dataset.

2 Method

We will mainly explain our method from three aspects: 1) dataset construction with dual-modality-based annotation as described in Section 2.1, 2) segmentation model using hybrid CNN-Mamba structure as detailed in Section 2.2 as well as anatomical enhancement module as introduced in Section 2.3, and 3) loss functions utilized in our work as formulated in Section 2.4.

2.1 Dual-Modality-based Annotation Strategy

As shown in Fig. 1, each modality has its own emphasis in annotation, therefore we use a dual-modality-based annotation strategy on our in-house dataset. Specifically, each set consists of a PET volume, a corresponding CT volume, and both PET and CT lesion masks, annotated by three trained annotators with the support of two senior nuclear medicine physicians.

2.2 Hybrid CNN-Mamba Encoder

Mamba has achieved outstanding performance compared to other methods in the domain of long-range modeling, which can yield a significant improvement for the high-level feature extraction. As depicted in Fig. 2, we merge CT lung window volume, CT soft tissue window volume, and PET SUV volume into a three-channel input $\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$ with $C = 3$. An 8× downsampling CNN encoder transforms the input into a latent space representation $\mathbf{x}_L \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times J}$, denoted as $\mathbf{x}_L = \mathbf{a}_0$.

The feature map from the CNN encoder is partitioned into non-overlapping patches using convolution, then flattened into $\mathbf{x}_M \in \mathbb{R}^{N \times K}$ where $N = (\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8})$ is the length of the sequence and K is the hidden dimension. We then add a 1D learnable positional embedding $\mathbf{x}_{pos} \in \mathbb{R}^{N \times K}$ to $\mathbf{x}_M$ for forming $\mathbf{z}_0$ as

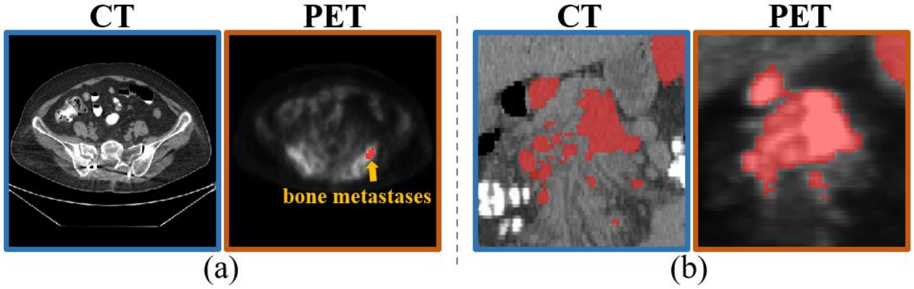

Fig. 1. (a) Illustration of an annotation condition where bone metastases (as shown in PET scan) have no corresponding structural change in CT scan, highlighting discrepancies between two modalities. (b) Demonstration of how the same lesion's annotated boundaries vary between CT and PET due to distinct characteristics of each modality.

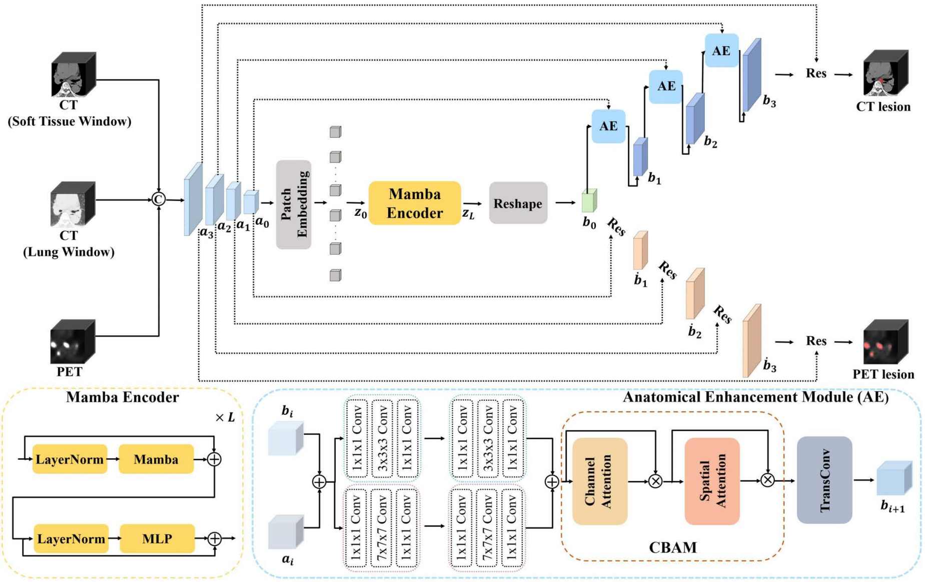

Fig. 2. Overview of the proposed LM-UNet architecture. Multi-channel volume is fed into a Hybrid CNN-Mamba encoder for feature extraction. Then these extracted features are passed into two parallel decoders for lesion segmentation of CT and PET, respectively. Particularly, the CT lesion segmentation decoder employs the proposed anatomical enhancement module (AE).

the input of Mamba encoder. The input and output dimensions of Mamba block are both $\mathbb{R}^{N \times K}$.

The Mamba encoder consists of multiple Mamba blocks, each with similar block structure as Vision Transformer (ViT) [2]. The only difference is that the Mamba layer replaces the multi-head self-attention in the ViT block. The detailed formulation is given below:

$$\begin{aligned} \widetilde{\mathbf{z}}_l &= Mamba(LN(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, & l = 1 \cdots L, \\ \mathbf{z}_l &= MLP(LN(\widetilde{\mathbf{z}}_l)) + \widetilde{\mathbf{z}}_l, & l = 1 \cdots L, \end{aligned} \tag{1}$$

where $\mathbf{z}_0$ is the input of the Mamba encoder, and $Mamba$ denotes the Mamba layer. MLP denotes Multi-Layer Perceptron, and LN is the layer normalization operator. As the output of Mamba encoding, $\mathbf{z}_L$ is the final representation of the whole encoding process.

2.3 Anatomical Enhancement Module

The anatomical enhancement module (AE) is introduced to provide detailed anatomical constraints. We reshape $\mathbf{z}_L$ to $\mathbf{b}_0$ with a dimension of $\mathbf{b}_0 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times \frac{D}{8} \times J}$ for compatibility with the CNN upsampling process. The LM-UNet decoder has two paths, i.e., one for PET lesions, (by using a ResNet architecture with skip connections), and another for CT lesions, (by incorporating an AE strategy for skip connections). The proposed AE module is defined as:

$$\begin{aligned} \widetilde{\mathbf{b}}_i &= Conv_1(\mathbf{a}_i + \mathbf{b}_i) + Conv_2(\mathbf{a}_i + \mathbf{b}_i), & i = 0, 1, 2 \\ \mathbf{b}_{i+1} &= TransConv(CBAM(\widetilde{\mathbf{b}}_i)), & i = 0, 1, 2 \end{aligned} \tag{2}$$

where $Conv_1$ represents a larger kernel (7×7×7) and $Conv_2$ represents a smaller kernel (3×3×3) to focus on narrow edge information. $CBAM$ [20] improves network performance by focusing on important channels and spatial locations. We find that multi-scale convolutions combined with the $CBAM$ module can effectively enhance segmentation performance, particularly for fine structural details.

2.4 Loss Function

Our model employs a dual loss strategy, combining Dice Loss [14] and Cross Entropy Loss. For each modality, the loss is defined as:

$$\mathcal{L} = \mathcal{L}_{dice} + \mathcal{L}_{ce}, \tag{3}$$

Therefore, combining the loss for CT $\mathcal{L}_{CT}$ and the loss for PET $\mathcal{L}_{PET}$, we have the overall loss function as below:

$$\mathcal{L}_{all} = \mathcal{L}_{CT} + \mathcal{L}_{PET}. \tag{4}$$

This balanced loss between CT and PET ensures effective learning for both modalities.

3 Experiments

3.1 Datasets and Annotations

To assess the performance of our proposed architecture, we utilize the public dataset autoPET [4] and our in-house dataset from three medical centers.

autoPET. The public autoPET dataset consists of 1014 3D whole-body FDG PET-CT sets from 900 patients, annotated by two experts. Each set includes a PET volume, a corresponding CT volume, and a binary mask for tumor lesions based on PET. Approximately half of these cases are cancer-free, and the rest are with histologically-confirmed malignant melanoma, lymphoma, or lung cancer. To align two image modalities in spatial resolution, we resample the CT scans to the same resolution as PET images.

In-house dataset. The in-house dataset collected from three medical centers consists of 344 3D whole-body PET-CT sets from patients with cancer-positive, covering various types of cancer such as lymphoma, lung cancer, and intestinal cancer.

For both datasets, 90% of the cases are kept for training and cross-validation, and the rest 10% of the cases are used as test set.

3.2 Implementation Details

We implement our framework in PyTorch[8] on one NVIDIA A100 GPU equipped with 80G RAM. For fair comparison, all the investigated models are trained according to the nnU-Net [9] scheme. Namely, the default settings of nnU-Net are used for pre-processing, data augmentation, and training strategy. We crop the input volume into patches with a size of 128×128×128. All the models are trained using the AdamW [12] optimizer from the scratch for 1500 epochs. We set the learning rate as 10^{-4} and the mini-batch size as 2. In the inference phase, we follow the nnU-Net using the scheme of sliding window.

3.3 Quantitative Evaluation

To quantitatively evaluate our method, we compare with three CNN-based segmentation networks (nnU-Net [9], SegResNet [15], and Attention U-Net [16]), three Transformer-based networks (TransBTS [19], UNETR [8], and Swin UNETR [7]), one PET-CT lesion segmentation network (H-DenseFormer [18]), and two Mamba-based networks (U-Mamba [13], and SegMamba [21]). SegResNet, Attention U-Net, UNETR, and Swin UNETR are selected from MONAI[9]. We use Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95) as evaluation metrics.

[8] http://pytorch.org/

[9] https://monai.io/

Table 1. Quantitative results of different methods on **public** dataset. The best results are highlighted in red and the second best results are highlighted in blue.

Methods	DSC ↑	HD95 (mm) ↓
autoPET		
nnU-Net (2018) [9]	0.7421±0.1123	39.6±10.6
SegResNet (2018) [15]	0.7315±0.0934	40.0±11.2
Attention U-Net (2018) [16]	0.6687±0.1162	74.3±22.0
TransBTS (2021) [19]	0.7099±0.0915	40.1±15.1
UNETR (2018) [8]	0.7168±0.1272	53.4±14.6
Swin UNETR (2022) [7]	0.7298±0.0769	49.6±9.5
H-DenseFormer (2023) [18]	0.7131±0.1326	34.5±10.1
U-Mamba Enc (2024) [13]	0.7359±0.0862	39.2±13.4
SegMamba (2024) [21]	0.7304±0.1025	48.0±9.2
LM-UNet	0.7543±0.0801	34.5±8.6

Table 2. Quantitative results of different methods on **in-house** dataset. The best results are highlighted in red and the second best results are highlighted in blue.

Methods	DSC ↑		HD95 (mm) ↓	
	CT	PET	CT	PET
In-house Dataset				
nnU-Net (2018) [9]	0.6705±0.1212	0.8273±0.0652	53.1±20.6	23.6±7.8
SegResNet (2018) [15]	0.6511±0.1343	0.8331±0.0748	55.9±24.9	22.8±8.1
Attention U-Net (2018) [16]	0.6294±0.1736	0.8097±0.0410	59.2±22.6	21.1±6.8
TransBTS (2021) [19]	0.6392±0.2013	0.8143±0.0578	56.8±22.9	24.2±8.1
UNETR (2018) [8]	0.5910±0.1437	0.8102±0.0477	52.0±25.0	22.7±8.1
Swin UNETR (2022) [7]	0.6427±0.1826	0.8204±0.0328	51.0±24.5	19.2±6.9
H-DenseFormer (2023) [18]	0.6302±0.1598	0.8111±0.0463	60.2±30.7	26.1±10.6
U-Mamba Enc (2024) [13]	0.6648±0.1291	0.8201±0.0491	49.6±22.2	22.6±8.3
SegMamba (2024) [21]	0.6683±0.1062	0.8344±0.0627	52.8±26.0	23.3±8.2
LM-UNet	0.6998±0.1111	0.8492±0.0499	50.9±23.2	20.3±7.2

We summarize the quantitative results in Table 1 and Table 2. It is shown that our LM-UNet outperforms the state-of-the-art methods by a large margin, especially for CT lesion segmentation on our in-house dataset. Besides, we can find that all the Mamba-based methods have shown promising performance. It is worth noting that for the autoPET dataset, since it contains single-modality-based annotation, we use the structure of the CT branch as the decoder in our LM-UNet. For the models that are originally designed for single-task segmentation, we use the same decoder as their original work for both annotation branches on our in-house dataset.

3.4 Qualitative Evaluation

Besides quantitative evaluation, we also conduct visual comparison for qualitative evaluation on our dual-modality-based annotated in-house dataset. We demonstrate the comparison in Fig. 3. We can see that our LM-UNet has superior capability in differentiating pathological from normal physiological metabolism

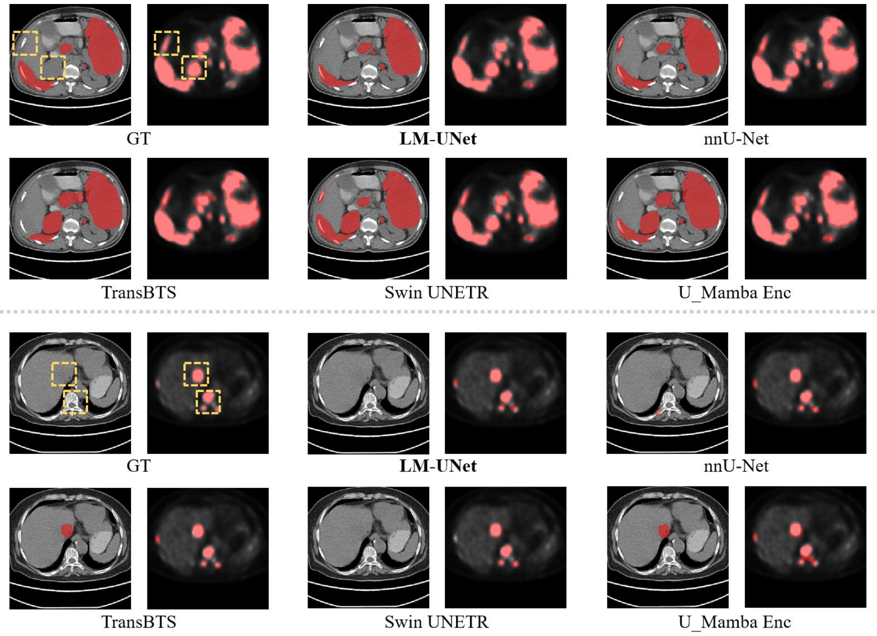

Fig. 3. Qualitative results for two representative cases on our in-house dataset. Upper two rows: Case 1; Bottom two rows: Case 2.

and provides modality-specific insights. For example, within the kidney, which is a region of daily high metabolic activity, our model is capable of precisely determining whether the observed metabolism is pathological. Besides, in the liver and bone regions, our model can accurately avoid potential misunderstanding based on the normality in CT images.

Table 3. The impact of AE module on segmentation performance.

Methods	DSC ↑		HD95 (mm) ↓	
	CT	PET	CT	PET
		In-house Dataset		
LM-UNet w/o AE	0.6812±0.1203	0.8471±0.0503	51.4±24.9	22.5±6.9
LM-UNet w/ AE	**0.6998±0.1111**	**0.8492±0.0499**	**50.9±23.2**	**20.3±7.2**

3.5 Ablation Studies

Effectiveness of AE Module. We conduct an ablation study on AE module to verify its effectiveness on our in-house dataset. The results are listed in Table 3.

We can see that AE improves the performance in the lesion segmentation of CT images in terms of both DSC and HD95.

4 Conclusion

In this work, we propose a threefold strategy for whole-body PET-CT lesion segmentation. Specifically, we develop a dual-modality-based annotation strategy to improve clinical applicability. Besides, we introduce a novel segmentation network, called LM-UNet, to accurately identify lesions by modeling long-range dependencies. Moreover, we present an anatomical enhancement module to better constrain anatomical shape of lesions. Experiments on both public and in-house datasets demonstrate that our LM-UNet *not only* achieves outstanding performance in whole-body automatic lesion segmentation from PET-CT, *but also* shows promising potential for the use of dual-modality-based annotation. Through these efforts, we aspire to contribute to the advancement of AI-assisted diagnostic applications, ultimately improving patient care and outcomes.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (grant numbers 62131015, 62250710165, U23A20295), the STI 2030-Major Projects (No. 2022ZD0209000), Shanghai Municipal Central Guided Local Science and Technology Development Fund (grant number YDZX20233100001001), and Science and Technology Commission of Shanghai Municipality (STCSM) (grant number 21010502600).

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

References

1. 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)
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
3. Farwell, M.D., Pryma, D.A., Mankoff, D.A.: Pet/ct imaging in cancer: current applications and future directions. Cancer **120**(22), 3433–3445 (2014)
4. Gatidis, S., Früh, M., Fabritius, M., Gu, S., Nikolaou, K., La Fougère, C., Ye, J., He, J., Peng, Y., Bi, L., et al.: The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging (2023)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)

7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI Brainlesion Workshop. pp. 272–284. Springer (2021)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)
9. 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)
10. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
11. Li, G.Y., Chen, J., Jang, S.I., Gong, K., Li, Q.: Swincross: Cross-modal swin transformer for head-and-neck tumor segmentation in pet/ct images. arXiv preprint arXiv:2302.03861 (2023)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
14. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
15. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. pp. 311–320. Springer (2019)
16. 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 preprint arXiv:1804.03999 (2018)
17. 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)
18. Shi, J., Kan, H., Ruan, S., Zhu, Z., Zhao, M., Qiao, L., Wang, Z., An, H., Xue, X.: H-denseformer: An efficient hybrid densely connected transformer for multimodal tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 692–702. Springer (2023)
19. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: Transbts: Multimodal brain tumor segmentation using transformer. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. pp. 109–119. Springer (2021)
20. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
21. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv preprint arXiv:2401.13560 (2024)