H2ASeg: Hierarchical Adaptive Interaction and Weighting Network for Tumor Segmentation in PET/CT Images

Jinpeng Lu[1], Jingyun Chen[1], Linghan Cai[2(✉)], Songhan Jiang[2], and Yongbing Zhang[2(✉)]

[1] School of Science, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
cailh@buaa.edu.cn, ybzhang08@hit.edu.cn

Abstract. Positron emission tomography (PET) combined with computed tomography (CT) imaging is routinely used in cancer diagnosis and prognosis by providing complementary information. Automatically segmenting tumors in PET/CT images can significantly improve examination efficiency. Traditional multi-modal segmentation solutions mainly rely on concatenation operations for modality fusion, which fail to effectively model the non-linear dependencies between PET and CT modalities. Recent studies have investigated various approaches to optimize the fusion of modality-specific features for enhancing joint representations. However, modality-specific encoders used in these methods operate independently, inadequately leveraging the synergistic relationships inherent in PET and CT modalities, for example, the complementarity between semantics and structure. To address these issues, we propose a Hierarchical Adaptive Interaction and Weighting Network termed H2ASeg to explore the intrinsic cross-modal correlations and transfer potential complementary information. Specifically, we design a Modality-Cooperative Spatial Attention (MCSA) module that performs intra- and inter-modal interactions globally and locally. Additionally, a Target-Aware Modality Weighting (TAMW) module is developed to highlight tumor-related features within multi-modal features, thereby refining tumor segmentation. By embedding these modules across different layers, H2ASeg can hierarchically model cross-modal correlations, enabling a nuanced understanding of both semantic and structural tumor features. Extensive experiments demonstrate the superiority of H2ASeg, outperforming state-of-the-art methods on AutoPet-II and Hecktor2022 benchmarks. The code is released at https://github.com/JinPLu/H2ASeg.

Keywords: PET/CT imaging · Multi-modal segmentation · Modality-cooperative spatial attention · Target-aware modality weighting

J. Lu and J. Chen—Contributed equally to this work.

1 Introduction

Positron emission tomography (PET) combined with computed tomography (CT) is considered the imaging modality of choice for the diagnosis, staging, and monitoring of treatment responses in various cancers [4,7,17]. Benefiting from the sensitivity to high metabolic areas, PET has demonstrated excellent capabilities in localizing tumors. Nevertheless, the low resolution of PET images prevents radiologists from relying solely on PET for outlining tumors. Conversely, despite CT cannot detect metabolic activity, it provides detailed structural information. Hence, the integration of PET and CT yields rich complementary information that enables accurate presentation of tumor regions [13]. In clinical practice, manual annotation is a time-consuming process, limiting the effect of PET/CT examinations. Fortunately, the development of deep learning [1, 20, 22, 25, 27] presents new avenues for automatic tumor segmentation in PET/CT.

In the multi-modal segmentation task, early convolutional neural networks (CNNs) such as UNet-3D [5], VNet [16], and ResUNet-3D [3] face limitations in capturing contextual information, resulting in high-metabolism organs, which have a similar presentation to tumor regions in PET scans, often being misidentified. Subsequent works (e.g., UNETR [10] and SwinUNETR [9]) leverage Transformer [6, 19] to explicitly model long-range spatial dependencies, significantly improving tumor segmentation performance. However, these methods rely on a simplistic concatenation operation for modality fusion that fails to model the non-linear relationship between PET and CT modalities. Thus, they cannot fully exploit the complementarity of each imaging modality in tumor segmentation.

Considering the specificity of different modalities, A2FSeg [21] tailors an encoder for each modality to derive modality-specific feature representations, and then adopts an adaptive fusion module to generate a joint feature representation for tumor segmentation. NestedFormer [23] investigates a modality-aware feature aggregation module to fuse high-level features and uses the fused features to adaptively select shallow features for better decoding. Despite significant advancements, these solutions [14,21,23,26] struggle to accurately segment tumors in PET/CT, as the modality-specific encoders are generally independent of each other, posing challenges in exploring the synergistic relations inherent in PET and CT modalities. To make use of the relation, Xue *et al.* [24] propose a multi-modal co-learning network that replaces the down-sampling blocks in the dual encoder with the shared down-sampling blocks to promote the feature interaction between PET and CT images. Although the intent behind this strategy makes sense, significant differences between PET and CT modalities often weaken the effectiveness of this module and make information transfer difficult.

To address these issues, we present a Hierarchical Adaptive Interaction and Weighting Network named H2ASeg for precise PET/CT tumor segmentation. Our motivation stems from the fact that during tumor annotation, radiologists repeatedly observe PET/CT images to roughly locate a tumor region, and subsequently use detailed information around the tumor to obtain the silhouette mask. We, therefore, argue that the hierarchical interaction of PET and CT images in a deep learning network is crucial for tumor localization, in addition to

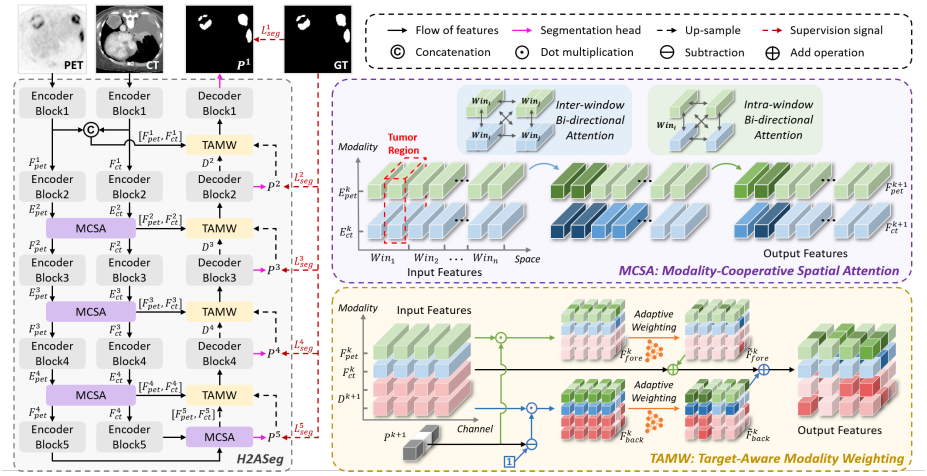

Fig. 1: Overview of the proposed H2ASeg, which is an encoder-decoder framework. MCSA modules perform modality interaction in the encoder. TAMW modules adjust multi-modal features for refining segmentation.

highlighting features related to the tumor for accurate boundary discrimination. To these ends, H2ASeg uses Modality-Cooperative Spatial Attention (MCSA) modules to execute modality interaction at different levels, in which the inter- and intra-window Bi-Directional Spatial Attention (BDSA) mechanisms bridge the connections between modalities globally and locally, enriching the complementary information. Meanwhile, a Target-Aware Modality Weighting (TAMW) module is developed to focus on tumor-related features for optimizing tumor boundaries. In a nutshell, our contributions can be summarized as follows. (1) We present an H2ASeg that hierarchically models the correlations between PET and CT, exploiting the complementary information for precise tumor segmentation. (2) We propose an MCSA module that can transfer valuable information across modalities at both local and global scales. Meantime, a TAMW module is developed to highlight tumor-related features from multi-modal features for segmentation refinement. (3) Extensive experiments demonstrate the superiority of H2ASeg, achieving state-of-the-art performance on two datasets.

2 Methodology

Fig. 1 illustrates the overall framework of **H2ASeg**, which adopts **two** modality-**A**daptive components into the **H**ierarchies of a dual-branch ResUNet-3D for tumor **Seg**mentation. At each level, MCSA executes feature interaction within and between modalities, collecting rich complementary information to improve feature representation. TAMW encourages the network to collect tumor-related features and then selects these features to refine tumor segmentation results.

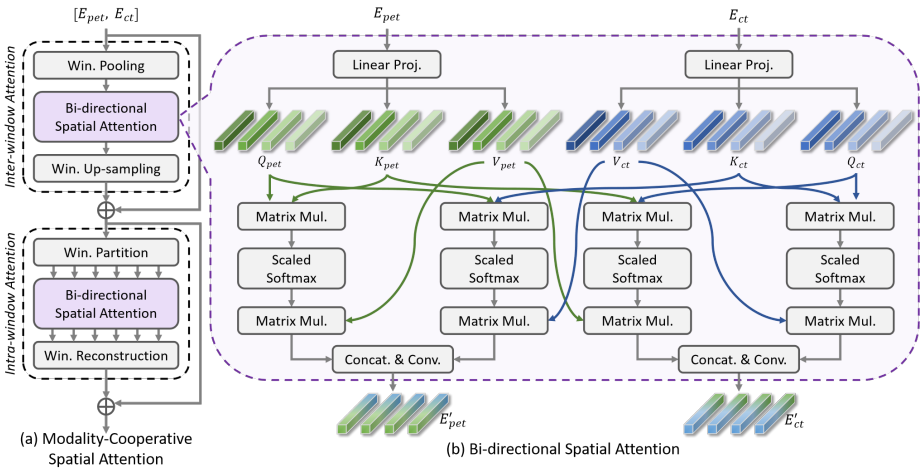

Fig. 2: Structure of modality-cooperative spatial attention module. (a) presents the overall architecture. (b) shows a bi-directional spatial attention mechanism. "Win." is the abbreviation of window, "Proj." is projection, "Mul." means multiply, "Concat. & Conv." denotes concatenation and convolution.

2.1 Modality-Cooperative Spatial Attention (MCSA)

Given the output features E^k_{ct}, E^k_{pet} from the k^{th} encoder block, MCSA treats them as two sets of 3D windows with the channel number of C^k and the window size of $(H^k_{win} \times W^k_{win} \times D^k_{win})$. As shown in Fig. 2 (a), MCSA firstly executes inter-window attention to establish the long-range dependencies between windows and then captures the details through intra-window attention. During the process, a Bi-Directional Spatial Attention (BDSA) mechanism is used for feature interaction within and between modalities to collect complementary information.

Concretely, in inter-window attention, each window of the inputs is pooled into a token by a convolutional layer with the kernel size and stride equal to the window size $(H^k_{win}, W^k_{win}, D^k_{win})$. In the process, the size of E^k_{pet} and E^k_{ct} changes from $(C^k \times H^k \times W^k \times D^k)$ to $(C^k \times \frac{H^k}{H^k_{win}} \times \frac{W^k}{W^k_{win}} \times \frac{D^k}{D^k_{win}})$. Subsequently, BDSA is used to model long-range dependencies across windows. After that, window up-sampling, a trilinear interpolation operation, reverses the above shape change to generate intermediate outputs $\hat{E}^k_{pet}$, $\hat{E}^k_{ct} \in \mathbb{R}^{C^k \times H^k \times W^k \times D^k}$.

In intra-window attention, window partition splits the feature into two ordered sets, S^k_{pet}, $S^k_{ct} \in \mathbb{R}^{N^k_h N^k_w N^k_d \times C^k \times H^k_{win} \times W^k_{win} \times D^k_{win}}$, of windows with preset shapes, where N^k_h, N^k_w, N^k_d denote the number of windows on the height, width, and depth axes respectively. Then, BDSA is implemented in each window to enhance local details. Finally, the windows are rearranged in order through window reconstruction to obtain the final output $F^k_{pet}, F^k_{ct} \in \mathbb{R}^{C^k \times H^k \times W^k \times D^k}$.

Bi-Directional Spatial Attention (BDSA). Fig. 2 (b) shows the structure of the BDSA mechanism, which can be simplified as two parallel unidirectional processes, using PET and CT as the source and the other as the target respectively. Given the source feature $E_s \in \mathbb{R}^{C\times H\times W\times D}$ and the target feature $E_t \in \mathbb{R}^{C\times H\times W\times D}$, BDSA performs $1\times 1\times 1$ convolutions to obtain two triples (Q_s, K_s, V_s) and (Q_t, K_t, V_t), where K_s and V_s are calculated from E_s, while Q_t, K_t, V_t, and Q_s are obtained from E_t. Secondly, BDSA flattens these features to $(C \times HWD)$, and computes self-attention A_{self} and cross-attention A_{cross}:

$$A_{self} = V_s \otimes Softmax(\frac{Q_s^T \otimes K_s}{\sqrt{C}})^T,\ A_{cross} = V_t \otimes Softmax(\frac{Q_t^T \otimes K_t}{\sqrt{C}})^T, \quad (1)$$

where "$\otimes$" is the matrix multiply operation. Finally, we concatenate A_{self} and A_{cross}, reshape the fused feature to $(C \times H \times W \times D)$, and then use a $3\times 3\times 3$ convolution to derive the improved target feature $E_t' \in \mathbb{R}^{C\times H\times W\times D}$.

2.2 Target-Aware Modality Weighting (TAMW)

Fig. 1 intuitively shows the TAMW module, which is designed to highlight tumor-related features from multi-modal features to refine segmentation. To be specific, the features $F_{pet}^k, F_{ct}^k \in \mathbb{R}^{C^k\times H^k\times W^k\times D^k}$, generated from MCSA, are concatenated with the up-sampled feature $\tilde{D}^{k+1} \in \mathbb{R}^{2C^{k+1}\times H^k\times W^k\times D^k}$ of the $(k+1)^{th}$ decoder block output to obtain $F^k \in \mathbb{R}^{2(C^k+C^{k+1})\times H^k\times W^k\times D^k}$. Next, the up-sampled prediction map $\tilde{P}^{k+1}$ generated from the $(k+1)^{th}$ decoder block is used to obtain the foreground feature F_{fore}^k and background feature F_{back}^k:

$$F_{fore}^k = \tilde{P}^{k+1} \odot F^k,\ F_{back}^k = (1-\tilde{P}^{k+1}) \odot F^k, \quad (2)$$

where $\odot$ denotes the dot product. Then, global average pooling is employed to F_{fore}^k and F_{back}^k to obtain the channel intensity, I_{fore}^k and $I_{back}^k \in \mathbb{R}^{2(C^k+C^{k+1})\times 1}$. We use an MLP with an activation function $tanh$ to derive the weights for foreground W_{fore}^k and background W_{back}^k, obtaining the enhanced feature F_{enh}^k:

$$W_{fore}^k = tanh(MLP(I_{fore}^k)),\ W_{back}^k = tanh(MLP(I_{back}^k)), \quad (3)$$

$$F_{enh}^k = F^k + W_{fore}^k \odot F_{fore}^k + W_{back}^k \odot F_{back}^k\ . \quad (4)$$

In TAMW, W_{fore}^k and W_{back}^k select foreground-related and background-related features adaptively for refinement. Take W_{fore}^k as an example, for a specific channel, when the corresponding foreground weight is greater than 0, it is *emphasized* by the network for optimizing tumor-related features; otherwise, it is *weakened*.

2.3 Loss Function

H2ASeg outputs five prediction maps corresponding to the five levels of our network. For the prediction map P^k of the k^{th} level decoder, we up-sample it

Table 1: Quantitative comparison (mean±std) on AutoPET-II and Hecktor2022. The best performance is shown in **bold**, and the second is underlined.

Methods	AutoPET-II				Hecktor2022			
	Dice (↑)	HD95 (↓)	Precision (↑)	Recall (↑)	Dice (↑)	HD95 (↓)	Precision (↑)	Recall (↑)
UNet-3D [5]	27.50±16.51	121.35±49.91	25.51±20.77	51.42±20.38	53.52±9.66	180.52±40.81	63.50±12.07	53.70±10.95
VNet [16]	55.07±16.21	72.90±43.01	61.94±18.10	64.17±15.10	51.36±8.12	185.69±30.28	54.60±10.66	58.50±12.13
ResUNet-3D [3]	39.78±20.53	69.11±42.95	33.16±22.53	60.70±17.15	45.52±12.83	184.74±36.19	61.05±9.01	43.97±13.52
UNETR [10]	39.45±18.15	92.37±36.25	42.54±23.67	52.01±18.25	46.22±10.05	207.02±31.91	57.57±7.81	47.49±9.04
SwinUNETR [9]	55.75±17.47	79.18±31.06	59.73±20.20	62.37±16.65	47.29±7.90	205.99±29.85	55.77±7.79	50.03±7.62
nnUNet [12]	55.93±12.31	70.12±28.71	58.37±14.97	61.20±15.09	50.04±6.04	140.80±23.77	56.37±6.89	57.11±6.02
NestedFormer [23]	57.33±16.76	67.38±34.04	68.85±21.07	61.48±18.64	51.65±7.99	217.33±30.67	56.85±9.02	57.17±8.64
A2FSeg [21]	52.41±19.07	67.24±37.75	57.97±22.03	59.32±18.65	54.77±6.56	144.50±25.84	**70.09**±7.09	51.90±6.98
SDB [24]	49.57±22.06	82.89±44.24	48.12±25.78	56.35±21.50	49.71±10.21	179.26±36.49	58.95±13.88	51.42±13.10
H2ASeg (Ours)	**60.03**±14.75	**63.09**±30.44	**68.91**±16.03	**65.44**±19.87	**59.69**±6.14	**131.92**±21.78	68.98±7.15	**58.54**±6.83

to the same size with ground truth GT and calculate binary cross-entropy loss $\mathcal{L}_{bce}$ and Dice loss $\mathcal{L}_{dice}$ [16]. The total loss $\mathcal{L}_{total}$ can be formulated as:

$$\mathcal{L}_{total} = \sum_{k=1}^{k=5}(6-k)(\mathcal{L}_{bce}(P^k, GT) + \mathcal{L}_{dice}(P^k, GT)). \tag{5}$$

3 Experiments

3.1 Experiments on PET/CT Tumor Segmentation

Datasets and Evaluation Metrics. We evaluate H2ASeg on two benchmarks: AutoPET-II [8] and Hecktor2022 [18]. **AutoPET-II** is a whole-body PET/CT tumor image dataset, including 1014 PET/CT scans of 900 patients. Each modality has a $3 \times 2 \times 2$ spacing and is already co-registered. Following the recent work [15], we first ensure that different scans from one patient do not exist in both the training and testing sets, and then randomly split the dataset in a ratio of training: validation: testing = 6: 2: 2. **Hecktor2022** is a head and neck tumor image dataset consisting of 524 PET/CT cases collected from 9 centers. Each modality has a $1 \times 1 \times 1$ spacing and is registered. Following the recent work [2], we randomly split the data in a ratio of training: validation: testing = 6: 2: 2.

We adopt four widely-used metrics for quantitative evaluation: foreground Dice score (Dice), 95% Hausdorff distance (HD95) [11], Precision, and Recall, where the Dice score, Precision, and Recall are presented as percentages (%) in experiments. Notably, since there are lesions-free samples in AutoPET-II, we only calculate Dice and HD95 in samples with lesions. To avoid randomness, we ran all experiments four times and reported the averaged metrics.

Implementation Details. Our method is implemented in PyTorch 2.0.0 on an NVIDIA GeForce RTX 3090 GPU. We employ the Adam optimizer to optimize the overall parameters with a learning rate of 1e-4 and a weight decay of 1e-5. In experiments, each image is cropped into patches with the size of $128 \times 128 \times 64$ for training. H2ASeg is trained end-to-end to 200 epochs with a batch size of 2. For MCSA, the window size is set to (8, 8, 4) for each layer of H2ASeg.

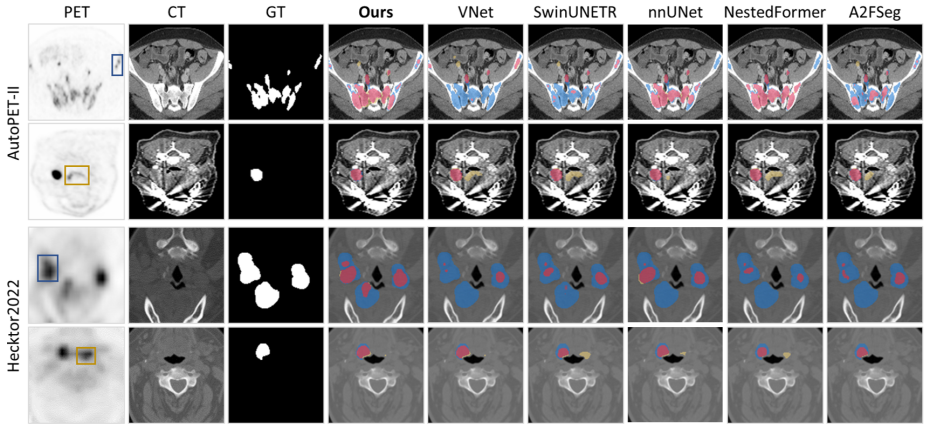

Fig. 3: Qualitative results of different methods. Boxes in PET highlight areas that are easily misclassified. For segmentation results, the red area is the true positive, the blue is the false negative, and the yellow is the false positive.

Quantitative and Qualitative Evaluation. We compare our H2ASeg with nine advanced methods, including UNet-3D [5], VNet [16], ResUNet-3D [3], UNETR [10], SwinUNETR [9], nnUNet [12], NestedFormer [23], A2FSeg [21], and SDB [24]. To ensure fairness, we obtain the experimental results by rerunning their released code based on the same training strategy.

Table 1 lists the quantitative comparison results, in which H2ASeg outperforms other methods on both datasets. On AutoPET-II, due to the limited global modeling ability, CNN-based methods like UNet-3D, ResUNet-3D are easily interfered with by the high metabolic areas, resulting in inferior performance to Transformer-based methods (e.g., UNETR, SwinUNETR). H2ASeg effectively uses Transformer to model the local and global relation across modalities, obtaining state-of-the-art result with Dice of 60.03%. On Hecktor2022, H2ASeg achieves significant improvements with a gain of 4.98% in Dice to the advanced multi-modal segmentation method A2FSeg. Fig. 3 provides visible comparison results. It can be seen that H2ASeg shows superior performance in localizing and segmenting tumors of different sizes under various scenes.

3.2 Ablation Study

To further evaluate the effectiveness of the proposed modules, we conducted ablation experiments and reported the quantitative results in Table 2. From the table, we can find the following. (1) MCSA and TAMW are effective components, which can significantly increase baseline performance by 12.20% and 10.94% in Dice respectively. (2) The joint use of MCSA and TAMW can further improve the baseline, achieving the state-of-the-art result on AutoPET-II. This reveals that MCSA and TAMW can synergistically enhance each other, enabling H2ASeg to accurately perform tumor segmentation in PET/CT images.

Table 2: Effects of MCSA and TAMW on AutoPet-II.

Modules		AutoPET-II			
MCSA	TAMW	Dice	HD95	Precision	Recall
		46.23	71.67	41.43	60.71
✓		58.43	69.04	63.94	64.17
	✓	57.19	65.91	**70.68**	58.45
✓	✓	**60.03**	**63.09**	68.91	**65.44**

Table 3: Effects (%) of PET/CT in TAMW for emphasis at different depths.

Depths	Foreground emphasis		Background emphasis	
	PET	CT	PET	CT
1	82.79 ± 19.69	95.77 ± 9.83	46.44 ± 21.61	39.81 ± 15.71
2	86.68 ± 11.15	65.36 ± 34.41	56.26 ± 26.83	43.31 ± 16.00
3	68.50 ± 31.11	53.84 ± 21.29	49.61 ± 26.89	55.39 ± 13.15
4	66.83 ± 0.29	61.42 ± 32.47	51.98 ± 20.60	48.15 ± 24.95

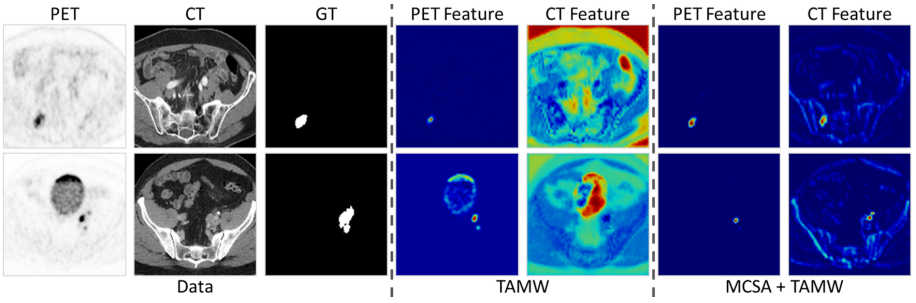

Fig. 4: Visualization of feature maps selected by TAMW for foreground emphasis.

To further prove the effectiveness of TAMW and MCSA modules, we visualize the feature maps (from F_{pet}^2 and F_{ct}^2) that are selected by TAMW for foreground emphasis in Fig. 4. It can be seen that the independent use of TAMW modules highlights tumor-related features in PET features, but fails to exploit CT features. The introduction of MCSA improves the phenomenon. This is because the modality interaction in MCSA enhances the feature representation of each modality, making the network effectively capture tumor-related features.

Table 3 shows the average weights of PET and CT for foreground emphasis and background emphasis in TAMW at different network depths. For the contribution of PET/CT in foreground emphasis, we observe that the H2ASeg tends to highlight CT features in the shallow layer (texture) and use PET features in deep layers (semantics). These behaviors are consistent with the actual role of PET and CT in clinical diagnosis, which demonstrates the hierarchical interaction of H2ASeg can effectively explore the complementary information contained in PET/CT, utilizing the tumor localization ability of PET and the detailed description ability of CT for precise segmentation.

4 Conclusion

In this paper, we propose a novel architecture H2ASeg that hierarchically models the correlations between modalities, exploiting the complementary information for precise PET/CT tumor segmentation. Specifically, we design a modality-cooperative spatial attention (MCSA) module to adaptively transfer potential information across modalities locally and globally. A target-aware modality

weighting (TAMW) module is developed for highlighting tumor-related features for segmentation refinement. Extensive experiments demonstrate the superiority of H2ASeg, with state-of-the-art results on two benchmarks. Moreover, MCSA and TAMW are flexible enough to be embedded into existing architectures with higher efficiency to excavate intrinsic correlations between modalities.

References

1. Andrearczyk, V., Oreiller, V., Boughdad, S., Le Rest, C.C., Tankyevych, O., Elhalawani, H., Jreige, M., Prior, J.O., Vallières, M., Visvikis, D., et al.: Automatic head and neck tumor segmentation and outcome prediction relying on fdg-pet/ct images: findings from the second edition of the hecktor challenge. Medical Image Analysis **90**, 102972 (2023)
2. Andrearczyk, V., Oreiller, V., Jreige, M., Castelli, J., Prior, J.O., Depeursinge, A.: Segmentation and classification of head and neck nodal metastases and primary tumors in pet/ct. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 4731–4735. IEEE (2022)
3. Bhalerao, M., Thakur, S.: Brain tumor segmentation based on 3d residual u-net. In: International MICCAI Brainlesion Workshop. pp. 218–225. Springer (2019)
4. Bussink, J., Kaanders, J.H., Van Der Graaf, W.T., Oyen, W.J.: Pet–ct for radiotherapy treatment planning and response monitoring in solid tumors. Nature Reviews Clinical Oncology **8**(4), 233–242 (2011)
5. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
7. Ell, P.: The contribution of pet/ct to improved patient management. The British journal of radiology **79**(937), 32–36 (2006)
8. Gatidis S, K.T.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions). The Cancer Imaging Archive **226** (2022)
9. 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)
10. 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)
11. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the hausdorff distance. IEEE Transactions on pattern analysis and machine intelligence **15**(9), 850–863 (1993)
12. 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)

13. Kapoor, V., McCook, B.M., Torok, F.S.: An introduction to pet-ct imaging. Radiographics **24**(2), 523–543 (2004)
14. Li, L., Zhao, X., Lu, W., Tan, S.: Deep learning for variational multimodality tumor segmentation in pet/ct. Neurocomputing **392**, 277–295 (2020)
15. Marinov, Z., Reiß, S., Kersting, D., Kleesiek, J., Stiefelhagen, R.: Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging. arXiv preprint arXiv:2303.07126 (2023)
16. 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)
17. Mu, W., Jiang, L., Zhang, J., Shi, Y., Gray, J.E., Tunali, I., Gao, C., Sun, Y., Tian, J., Zhao, X., et al.: Non-invasive decision support for nsclc treatment using pet/ct radiomics. Nature communications **11**(1), 5228 (2020)
18. Oreiller, V., Andrearczyk, V., Jreige, M., Boughdad, S., Elhalawani, H., Castelli, J., Vallieres, M., Zhu, S., Xie, J., Peng, Y., et al.: Head and neck tumor segmentation in pet/ct: the hecktor challenge. Medical image analysis **77**, 102336 (2022)
19. Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis p. 102802 (2023)
20. 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)
21. Wang, Z., Hong, Y.: A2fseg: Adaptive multi-modal fusion network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 673–681. Springer (2023)
22. Xiang, D., Zhang, B., Lu, Y., Deng, S.: Modality-specific segmentation network for lung tumor segmentation in pet-ct images. IEEE Journal of Biomedical and Health Informatics **27**(3), 1237–1248 (2022)
23. Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L.: Nestedformer: Nested modality-aware transformer for brain tumor segmentation (2022)
24. Xue, Z., Li, P., Zhang, L., Lu, X., Zhu, G., Shen, P., Ali Shah, S.A., Bennamoun, M.: Multi-modal co-learning for liver lesion segmentation on pet-ct images. IEEE Transactions on Medical Imaging **40**(12), 3531–3542 (2021)
25. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 107–117. Springer (2022)
26. Zhao, X., Li, L., Lu, W., Tan, S.: Tumor co-segmentation in pet/ct using multimodality fully convolutional neural network. Physics in Medicine & Biology **64**(1), 015011 (2018)
27. Zhou, H.Y., Guo, J., Zhang, Y., Han, X., Yu, L., Wang, L., Yu, Y.: nnformer: Volumetric medical image segmentation via a 3d transformer. IEEE Transactions on Image Processing (2023)