MST-UDA: A Unsupervised Domain Adaptation Framework via 2.5D Multi-Style Perceptual Translation Network and Self-Filtering for Cross-Modal Multi-Organ Segmentation

Zengmin Zhang[1 [0009−0000−4181−0403]], Yanjun Peng[1 (✉) [0000−0002−8444−0622]], Xiaoning Zhang[1 [0009−0002−4202−7933]], and Haoran Li[1 [2222−−3333−4444−5555]]

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
pengyanjuncn@163.cn

Abstract. Abdominal medical image segmentation is essential for clinical diagnosis and treatment planning. Although abdominal CT multi-organ segmentation has achieved significant progress, MRI and PET modalities face challenges of annotation scarcity and cross-modal domain gaps. Unsupervised domain adaptation (UDA) provides an effective solution. However, existing methods suffer from domain shift and training instability when adapting to multiple target styles. Thus, we propose a UDA framework for MRI/PET abdominal multi-organ segmentation based on multi-style perceptual translation and self-filtering. Achieved accurate segmentation of unlabeled MRI/PET only using CT annotations. Specifically: (1) We designed an enhanced 2.5D Multi-Style Perceptual Translation Network (MST-Net) synthesizing diverse fake MRI/PET images from CT; (2) We train a dense segmentation model using multi-style data to generate pseudo-labels for real MRI/PET images; (3) We filter fake images and pseudo-labels through accuracy and stability assessment to improve final train data quality; (4) Final, we employ a two-stage lightweight segmentation model for accurate and efficient MRI/PET segmentation. Experiments on FLARE2025 validation set show our method achieves excellent performance with fast, low-resource characteristics: MRI and PET average DSC reach 80.77% and 62.83%, with 3.47s average inference time and 2479MB peak memory consumption.

Keywords: Unsupervised domain adaptation · Cross-Modal Segmentation · Style Translation.

1 Introduction

Abdominal multi-organ segmentation is fundamental to computer-assisted diagnosis, treatment planning, and prognosis [11]. With advances in deep learning, CT-based abdominal multi-organ segmentation methods [36] [35] have achieved remarkable results. However, while CT imaging is widely adopted in clinical

practice, MRI and PET offer unique advantages for abdominal imaging and diagnosis. MRI provides superior soft tissue contrast and radiation-free imaging, enabling clearer visualization of parenchymal organ structures. PET imaging offers metabolic functional information that is irreplaceable for tumor detection, staging, and treatment response evaluation. However, the high cost and effort of obtaining expert annotations on MRI and PET hinder the development of robust segmentation models in these modalities. Thus, making effective cross-modal domain adaptation methods essential.

Direct application of CT-trained segmentation models to MRI and PET images faces significant challenges due to substantial domain gaps caused by different imaging mechanisms, scanning protocols, and tissue contrasts [14]. Unsupervised domain adaptation (UDA) offers a promising solution by leveraging labeled source domain data and unlabeled target domain data to improve target domain performance. Current UDA approaches primarily fall into two categories: (1) image-to-image translation methods using generative adversarial networks (GANs) to establish cycle consistency [8], and (2) feature alignment approaches that minimize distributional constraints at the feature level [27]. However, simple feature alignment cannot meet the needs of multi-factor differences, and it is also prone to causing negative transfer and performance degradation when adapting to MRI and PET, especially leading to domain shift and training instability in small organ.

Recent works have explored disentangled representation learning to address domain gaps in medical imaging. Yang et al. [32] combined structural and appearance features through image translation. Yao et al. [33] proposed multi-style image translation frameworks to mitigate domain alignment issues. However, most existing methods operate on 2D images and stack slice-wise predictions, potentially losing anatomical continuity in the depth dimension [12]. While some works like [25] incorporate inter-slice attention mechanisms, they typically apply attention only at the deepest layers, leaving multi-level global context underexploited. Furthermore, existing disentanglement-based methods employ relatively simple style feature extraction, which may be insufficient for modalities with large style variations like CT-to-MRI/PET translation. May cause stylistic ambiguity of certain tissue textures and organs.

To address these limitations, we propose an unsupervised cross-modal domain adaptation framework that integrates a novel 2.5D Multi-Style Translation Network (MST-Net) with self-filtering mechanisms for CT-to-MRI/PET segmentation. In the style translation stage, unlike existing approaches that employ separate models for different target modalities, our MST-Net simultaneously processes MRI and PET images within a unified framework, reducing training complexity while fully exploiting inter-modal correlations. In MST-Net, we design a 2.5D dynamic slice fusion module that learns adaptive weights among three adjacent slices, effectively capturing inter-slice spatial relationships. Besides, we incorporate a pyramid attention mechanism in the style encoder that extracts and aggregates multi-scale features from different encoding layers, enhancing style vector expression for cross-modal generalization. Moreover, we

introduce a Large-Small Convolution (LSConv [28]) module that mimics human visual perception by combining large kernel global context capture with small kernel local detail processing, improving feature discriminability across modalities. In the segmentation stage, we first establish reliable supervision by selecting stable CT cases based on overlap agreement among top-performing FLARE2022 [9] [29] models. Then, a dense segmentation model trained on real CT and fake MRI/PET data generates pseudo-labels for real target images. We then employ accuracy and stability-based self-filtering to select high-quality synthetic images and pseudo-labels, culminating in a two-stage coarse-to-fine segmentation framework for precise target domain adaptation segmentation.

Our main contributions are summarized as follows:

- We proposed a unified 2.5D multi-style perceptual translation and self-filtering UDA framework for cross-modal abdominal multi-organ segmentation.
- We introduced an enhanced multi-style translation network MST-Net, for improved style disentanglement and feature learning.
- We apply a comprehensive self-filtering strategy using overlap ratio, accuracy, and stability metrics for high-quality data selection.
- Superior segmentation performance on FLARE2025 with efficient resource utilization, demonstrating practical clinical applicability.

2 Method

As illustrated in Figure 1, our framework consists of three main stages: **(1) Data preparation:** to enhance training efficiency and ensure data quality, a hierarchical data selection strategy was adopted. For the source domain computed tomography (CT) data, 50 fully annotated CT images were selected from the FLARE2022 dataset. Additionally, 450 CT images with reliable pseudo-labels were screened from 2000 candidates through pseudo-label overlap calculation. For the target domain data, 180 unannotated MRI images were randomly selected from the AMOS dataset, and 320 unannotated MRI images (8 modalities $\times$ 40 cases) were obtained from the LLD dataset; meanwhile, 500 unannotated PET images were collected. These data were integrated to construct a multi-modal training set. **(2) Style translation:** the selected CT, MRI, and PET data were fed into the multi-style translation network (MST-Net) for cross-modal style translation training. To fully leverage 2.5D spatial information, 5 groups of consecutive 3-slice sequences were randomly sampled from each 3D input image, ensuring that all slice regions of the images were covered during the training process. Through the multi-style perceptual translation capability of MST-Net, synthetic CT images with MRI-style and PET-style characteristics were generated. **(3) Self-Filtering and Segmentation:** first, the MedNeXt-M [24] dense segmentation model was trained using the CT data selected in the previous stage and the fake MRI-style/PET-style images, to guarantee the accuracy of pseudo-label generation. By performing inference on the fake images

and comparing the results with the ground truth labels, 50 cases with the highest precision were selected as a part of the final high-quality training dataset. Simultaneously, inference was conducted on the real MRI and PET data, and 400 cases with the most stable performance were screened via category stability detection as another part of the final train set. Finally, a two-stage coarse-to-fine segmentation framework inspired by [15] was employed to achieve fast and accurate MRI/PET segmentation with low resource consumption.

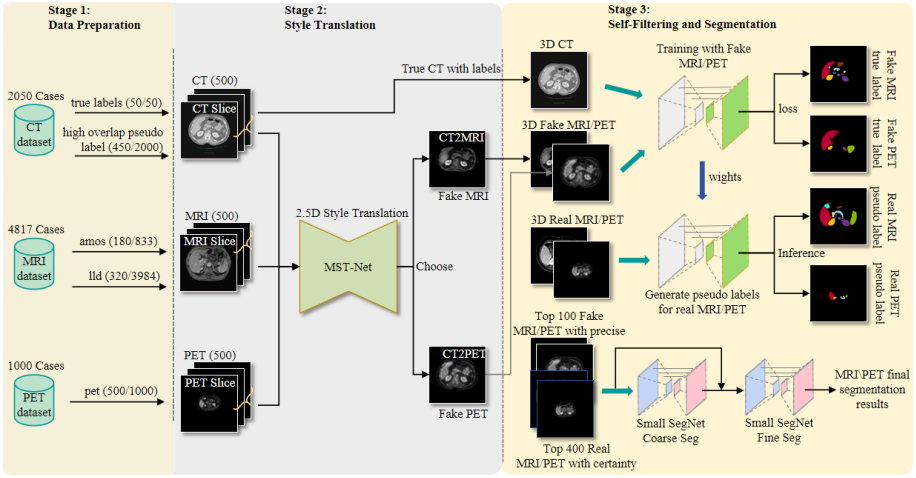

Fig. 1. Overview of our proposed unsupervised domain adaptation abdominal multi-organ segmentation framework based on 2.5D multi-style perceptual translation network and self-filtering.

2.1 Preprocessing

For the data preparation stage, in addition to random select part, we mainly did the following operations:

- From the 2000 CT case pseudo-labels from the best-performing team [9] [29] in the FLARE2022 dataset [21], the cases were ranked according to the average overlap rate across all categories, and the 450 cases with the highest annotation consistency were finally selected.
- From the LLD-MMRI dataset [16], we select 40 cases encompassing 8 imaging modalities, resulting in 320 image volumes.

For the style translation stage, we use the following data preprocessing steps:

- All CT, MRI, and PET images are resampled to uniform [4.0, 1.0, 1.0] mm spacing using trilinear interpolation.

– Images are zero-padded to achieve consistent dimensions of depth×512×512.
– All data were normalized. In addition, to ensure the accuracy of the CT structure in the translation model, the CT intensity was clipped to [-350:350].

For the segmentation phase, we employ the following data preprocessing steps:

– For LLD-MMRI dataset, we utilize C+Delay modality segmentation results as pseudo-labels for the remaining 7 modalities due to its good annotation quality and contrast characteristics.
– All data is redirected to the target direction.
– Target-size-based scaling replaces physical spacing-based resampling for computational efficiency. Dense and fine segmentation models use input dimensions of [96, 192, 192]. Coarse segmentation model employs reduced input size of [64, 64, 64] for faster processing.
– The resampled data were normalized to [0, 1] based on the mean and standard deviation.

2.2 2.5D Multi-Style Perceptual Translation Network (MST-Net)

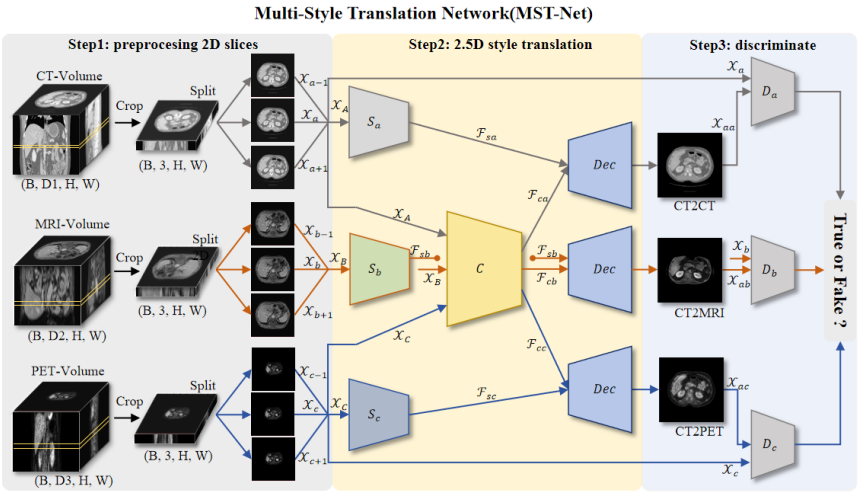

Fig. 2. The detailed structure of MST-Net. The translation model mainly consists of three independent style encoders, a shared content encoder, and a decoder. As well as three image discriminators and content discriminators

As shown in Figure 2, our proposed MST-Net starts with input from three consecutive slices of three modalities. We define the CT input slices as $\mathcal{X}_{a-1}$, $\mathcal{X}_a$, and $\mathcal{X}_{a+1}$, representing three longitudinally adjacent abdominal CT slices, and we denote $\mathcal{X}_{a-1:a+1}$ as $\mathcal{X}_A$. Similarly, we define $\mathcal{X}_B$ and $\mathcal{X}_C$ as the inputs

for MRI and PET modalities. After data augmentation operations such as flipping and rotation, they are respectively input into style encoders $\mathcal{S}_a$, $\mathcal{S}_b$, and $\mathcal{S}_c$ to extract rich style features from CT, MRI, and PET modalities. Additionally, we extract common structural and content features of the three modalities through a shared content encoder $\mathcal{C}$. Subsequently, the style vector $\mathcal{F}_{sa}$ from $\mathcal{S}_a$ and the CT content features $\mathcal{F}_{ca}$ from $\mathcal{C}$ are jointly input into the decoder, utilizing AdaIN [13] multi-level style embedding to complete the reconstruction from CT to CT modality. Similarly, style translation from CT to MRI and PET modalities can be obtained. To better constrain the style of generated images, we also perform image generation from MRI and PET to all three modalities. Finally, we feed all reconstructed images, transferred images, and content into the discriminator for discrimination to complete adversarial training. To ensure transfer stability, the loss function of our MST-Net is similar to [33], including reconstruction loss, adversarial loss, and discriminative loss.

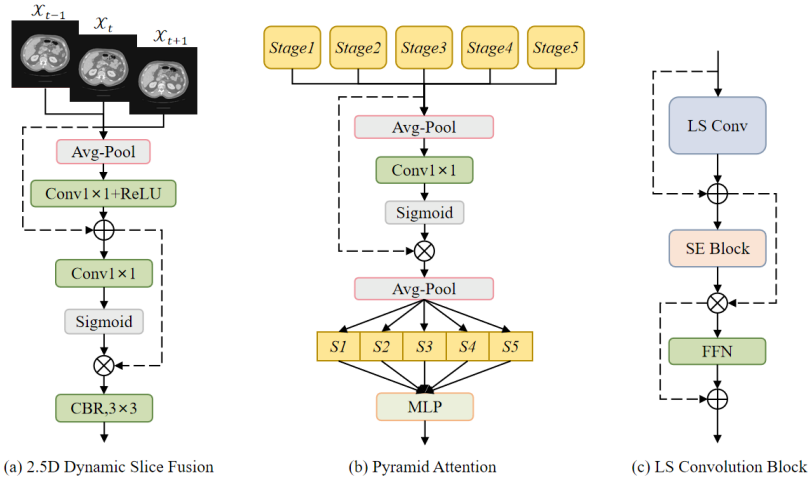

Fig. 3. Detailed structures of our proposed three modules in MST-Net.

2.5D Dynamic Slice Fusion(DSF): To model the depth-wise relationships between adjacent slices and tissue continuity and mitigate the inter-slice inconsistency problem associated with 2D methods, we propose the Dynamic Slice Fusion (DSF) module, which is integrated at the input stage of the style and content encoders. Specifically, the input is a 2.5D tensor $\mathcal{X}_T$ consisting of three consecutive slices $\mathcal{X}_{t-1:t+1}$. First, adaptive average pooling, 1×1 convolution, and ReLU activation are employed to extract the global semantic information of each slice. Subsequently, 1×1 convolution and Sigmoid activation are utilized to generate channel-adaptive weights; these weights are then applied to weight the context-enhanced features, enabling dynamic attention to different slices and channels. Finally, cross-slice feature fusion is achieved through 3×3 convolution,

batch normalization (BN), and ReLU activation, which enhances the representation quality of longitudinal structures.

Pyramid Attention(PA): To enhance the multi-style expression capability and generalization of the style encoder. Avoiding vector under-representation or excessive noise caused by a single scale, we design a Pyramid Attention (PA) module within the style encoder. Specifically, candidate style features are extracted from the multi-level encoder stages (Stage 1...5). For the feature vector of each level, adaptive pooling, 1×1 convolution, and Sigmoid activation are applied to generate channel attention weights for the corresponding scale. Subsequently, the multi-scale weighted features undergo global average pooling and flattening, respectively; the resulting vectors are concatenated to form a multi-scale aggregated style vector. Finally, a multi-layer perceptron (MLP) composed of two fully connected layers is employed for feature fusion and dimensionality reduction, resulting in a more robust style vector.

LS Convolution Block: To balance global context and local details during the content encoding stage, improve the transferability of cross-modal shared features, and enhance the stability of disentanglement and transfer, we introduce LS Convolution Block [28] in the content encoder that combines large-kernel and small-kernel convolutions. To simulate the collaborative perception of human vision toward global structures and local textures, thereby enhancing the ability to capture structural information.

Style Translation Visualization: As illustrated in Figure 4, we present the cross-domain translation results of MST-Net from CT to MRI and PET modalities. The first, second, and fourth column displays real CT, MRI, and PET images. MST-Net extracts global-to-detail structural features and multi-scale style vectors from the above images. The third and fifth columns demonstrate the synthesized MRI and PET images generated by fusing CT structural content with MRI and PET style characteristics. The visualization reveals that MST-Net prioritizes structural consistency throughout the cross-domain translation process by effectively disentangling content from style and progressively reconstructing the fused features.

2.3 Self-Filtering

To mitigate the impact of low-quality fake data and pseudo-labels on segmentation results, we implement a self-filtering strategy that processes generated fake MRI and PET images and pseudo-labels for real MRI and PET through two stages.

Accuracy calculation for fake images with real labels: To reduce the impact of domain shift on the real target domain, we perform inference prediction on 500 trained transferred images after completing precision model training. Since the transferred images are perfectly paired with corresponding CT images, we evaluate the Dice score between prediction results P_a and ground truth labels Y_a, filtering the top 100 results as high-quality transferred images for the final training set.

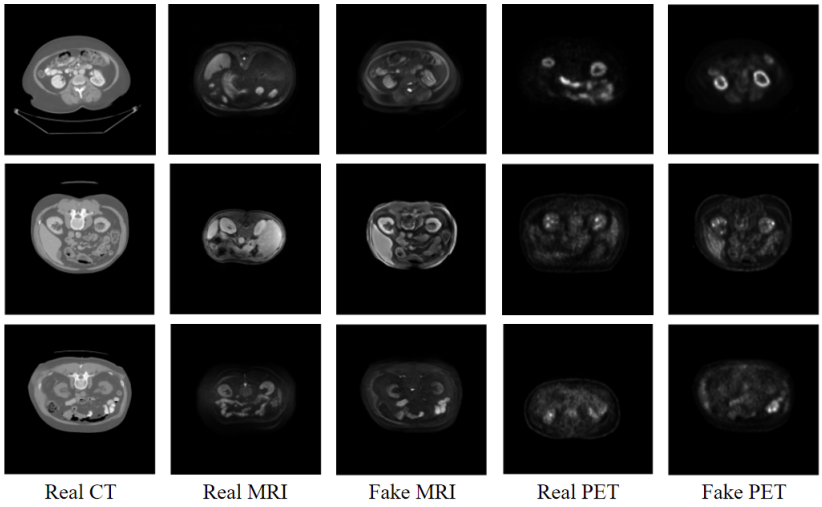

Fig. 4. CT-to-MRI and CT-to-PET style translation visualization.

$$\text{Accuracy}(i) = \frac{1}{C}\sum_{c=1}^{C}\text{Dice}(P_a^{(i,c)}, Y_a^{(i,c)}) \tag{1}$$

where C represents the number of classes, $P_a^{(i,c)}, Y_a^{(i,c)}$ mean the prediction and ground truth for image i, class c.

Stability calculation for real images with pseudo-labels: To reduce the impact of pseudo-label quality, we perform class instability detection on all unlabeled MRI and PET images. Following [9], we adopt segmentation target-oriented detection by calculating the proportion of non-overlapping regions between two prediction results for each class, selecting the top 400 images with highest stability rankings as another part of the final training set.

$$\text{Instability}(i) = \frac{1}{C}\sum_{c=1}^{C}\frac{|P_1^{(i,c)} \triangle P_2^{(i,c)}|}{|P_1^{(i,c)} \cup P_2^{(i,c)}|} \tag{2}$$

where $P_1^{(i,c)}, P_2^{(i,c)}$ represents two prediction results for image i, class c, $\triangle$ mean symmetric difference operator.

2.4 Final Segmentation Train and Inference

Regarding the coarse and fine segmentation models trained, both models are PH-Trans, which adopt a U-shaped encoding design consisting of a parallel mixture of convolution and Swin Transformer. For more details, please refer to [15].

Loss function: we use the summation between Dice loss and cross-entropy loss because compound loss functions have been proven to be robust in various medical image segmentation tasks [17].

2.5 Post-processing

We use leverage connected-component analysis. For each organ class, we keep only the largest predicted connected region and discard all remaining components, thereby suppressing spurious fragments and reducing false positives.

3 Experiments

3.1 Dataset and evaluation measures

The training dataset is curated from more than 30 medical centers under the license permission, including TCIA [2], LiTS [1], MSD [26], KiTS [6,7], autoPET [5,4], AMOS [11], LLD-MMRI [16], TotalSegmentator [30], and AbdomenCT-1K [23], and past FLARE Challenges [20,21,22]. The training set includes 2050 CT scans, 4817 MRI scans and 1000 PET scans. The core set includes 100 MRI and 100 PET scans sampled from the original training set. The validation set includes 160 MRI scans and 50 PET scans. The organ annotation process used ITK-SNAP [34], nnU-Net [10], MedSAM [18], and Slicer Plugins [3,19].

The evaluation metrics encompass two accuracy measures—Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD)—alongside two efficiency measures—running time and area under the GPU memory-time curve. These metrics collectively contribute to the ranking computation. Furthermore, the running time and GPU memory consumption are considered within tolerances of 15 seconds and 4 GB, respectively.

3.2 Implementation details

Data augmentation: We use online data augmentation strategies, including random rotation and scaling, adding white Gaussian noise, applying Gaussian blur, adjusting brightness and contrast, low-resolution simulation, gamma transformation, and elastic deformation.
Environment settings: The development environments and requirements are presented in Table 1.
Training settings: To maximize the quality of pseudo labels, we adopt the larger segmentation model MedNeXt-M with more parameters—as the high-precision generator; its training configurations are provided in Table 2. To reduce resource consumption and improve inference speed while maintaining accuracy, the final deployed coarse-to-fine SegNet uses a configuration with fewer parameters and lower FLOPs, resulting in faster training; the corresponding settings are listed in Table 2.

Table 1. Development environments and requirements.

System	Ubuntu 20.04.3 LTS
CPU	18 vCPU AMD EPYC 9754 CPU@2.70GHz
RAM	60GB
GPU (number and type)	NVIDIA 4090D 24G
CUDA version	11.3
Programming language	Python 3.8.10
Deep learning framework	torch 1.13, torchvision 0.13.1
Specific dependencies	connected-components-3d, MedPy, batchgenerators etc.
Code	https://github.com/zzm3zz/FLARE2025

Table 2. Training protocols for the dense and final segment model.

Model	Dense / Coarse / Fine
Network initialization	"He" normal initialization
Batch size	2 / 4 / 2
Patch size	96×192×192 / 64×64×64 / 96×192×192
Total epochs	300
Optimizer	AdamW
Initial learning rate (lr)	5e-4
Lr decay schedule	Cosline Annealing LR
Training time	26 / 2 / 20 hours
Loss function	Cross entropy + Dice
Number of model parameters	18.27 / 4.2 / 4.2M[1]
Number of flops	558.09 / 18.60 / 251.19G [2]
CO_2eq	6.2415 / 0.0973 / 1.8788 Kg[3]

Table 3. Quantitative evaluation results of MRI scans.

Target	Validation		Testing	
	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	95.95 ± 2.68	95.27 ± 5.81		
Right kidney	94.70 ± 5.97	95.30 ± 8.29		
Spleen	93.37 ± 12.21	93.31 ± 13.47		
Pancreas	77.97 ± 18.42	87.85 ± 19.43		
Aorta	92.16 ± 7.06	94.80 ± 7.71		
Inferior vena cava	86.24 ± 7.44	89.72 ± 8.66		
Right adrenal gland	62.61 ± 18.69	79.31 ± 20.04		
Left adrenal gland	60.35 ± 25.74	74.04 ± 29.08		
Gallbladder	72.17 ± 29.74	68.63 ± 31.02		
Esophagus	72.77 ± 17.73	89.65 ± 17.59		
Stomach	82.09 ± 15.76	86.03 ± 17.33		
Duodenum	64.98 ± 19.08	84.14 ± 19.25		
Left kidney	94.61 ± 6.24	95.79 ± 8.90		
Average	80.77 ± 13.16	87.22 ± 8.68		

Table 4. Quantitative evaluation results of PET scans.

Target	Validation		Testing	
	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	89.34 ± 3.27	80.15 ± 8.29		
Right kidney	71.27 ± 19.24	64.84 ± 19.96		
Spleen	34.10 ± 30.78	39.65 ± 26.66		
Left kidney	56.61 ± 34.65	61.96 ± 37.22		
Average	62.83 ± 28.50	61.65 ± 25.73		

Table 5. Ablation Study Results of our method.

ID	Train Data	Model			Fake Image	Pseudo Label	MST-Net	Self-Filtering	DSC(%)	
		Dense	Coarse	Fine					MRI	PET
v1	50CT	nnUNet-B	-	-	-	-	-	-	35.36	24.71
v2	50CT+Core	nnUNet-B	-	-	✓	-	-	-	67.85	41.78
v3	50CT+Core	PHTrans-L	-	-	✓	-	-	-	69.12	43.53
v4	50CT+Core	MedNeXt-M	-	-	✓	-	-	-	70.06	44.57
v5	50CT+Core	MedNeXt-M	-	-	✓	✓	-	-	72.64	46.33
v6	50CT+Core	MedNeXt-M	MedNeXt-S	MedNeXt-S	✓	✓	-	-	72.03	45.78
v7	50CT+Core	MedNeXt-M	PHTrans-S	PHTrans-S	✓	✓	-	-	72.12	46.07
v8	50CT+Core	MedNeXt-M	PHTrans-S	PHTrans-S	✓	✓	✓	-	76.46	52.21
v9	all C+M+P	MedNeXt-M	PHTrans-S	PHTrans-S	✓	✓	✓	-	78.38	55.65
ours	500C+M+P	MedNeXt-M	PHTrans-S	PHTrans-S	✓	✓	✓	✓	**80.77**	**62.83**

4 Results and discussion

4.1 Quantitative results on validation set

The quantitative results on the FALRE25 MRI validation set are summarized in Table 3. On the 110-case multimodal MRI validation cohort, our framework achieved a mean DSC of 80.77% with a standard deviation of 13.16%, and a mean NSD of 87.22% with a standard deviation of 8.68%. These results demonstrate effective domain adaptation, indicating successful transfer of knowledge from the source domain to the target domain. Additionally, the results on the 50-case PET validation set from FLARE25 are summarized in Table 4. Across the four-organ segmentation task, the framework achieved a mean DSC of 62.83% (standard deviation 28.50%) and a mean NSD of 61.65% (standard deviation 25.73%). Although the segmentation performance for the left-sided spleen and left kidney was suboptimal, competitive results were obtained for the right-sided liver and right kidney. These findings indicate that our framework also exhibits competitive performance on the PET modality.

To validate our framework design, we conducted comprehensive ablation experiments as shown in Table 5. (v1): Training nnUNet on 50 labeled CT cases achieved only 35.36% and 24.71% DSC on MRI and PET, respectively, confirming the substantial domain gap challenge. (v2): Introducing unsupervised domain adaptation with style-content translation model [33] on 50 CT cases and 100 core MRI/PET samples significantly improved performance to 67.85% and 41.78%, demonstrating the effectiveness of disentangled image translation. (v3,v4): We

evaluated different dense models, ultimately selecting MedNeXt-M [24] for optimal performance. (v5): Incorporating pseudo-labels from unlabeled target images provided modest improvements. (v6, v7): For computational efficiency, we adopted a two-stage coarse-to-fine approach using PHTrans-S [15], which maintained competitive accuracy while enhancing inference speed. (v8): Our proposed MST-Net substantially improved translation quality, boosting target domain performance to 76.46% and 52.21%. (v9): Scaling to the full dataset (all C+M+P) further increased performance to 78.38% and 55.65%, confirming the benefits of data augmentation for generalization. (ours): Finally, implementing self-filtering for quality control achieved the best results of 80.77% and 62.83% for MRI and PET, respectively. Notably, PET showed greater improvement from filtering due to its inherently lower structural clarity compared to MRI.

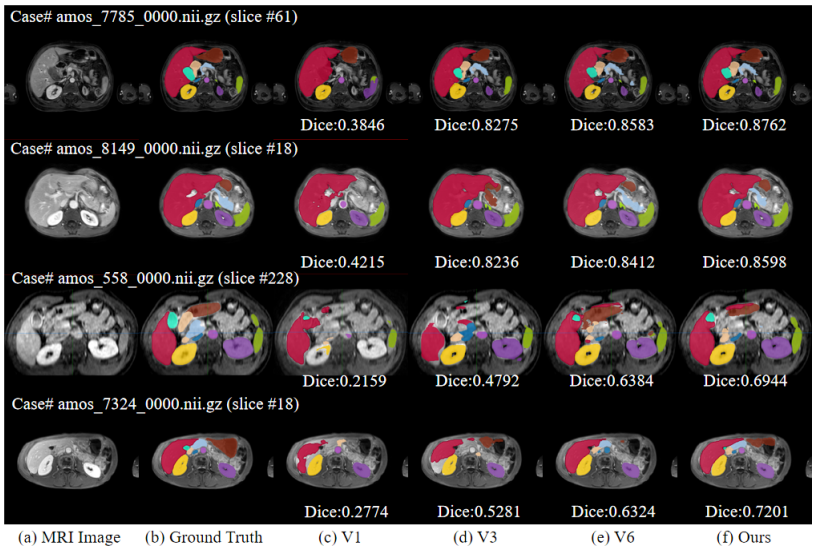

Fig. 5. Visualization of segmentation results of the MRI validation set.

4.2 Qualitative results on validation set

As illustrated in Figures 5 and 6, we present segmentation results on the validation set across four representative cases. The first two rows demonstrate successful segmentation examples. As observed in column (c), direct application of CT-trained models to MRI and PET images produces suboptimal segmentation results, with many small organs remaining undetected. Progressive implementation of our proposed methods yields gradual improvements, with our final approach (f) achieving comprehensive identification of small organs that closely approximates ground truth annotations. However, the bottom two rows

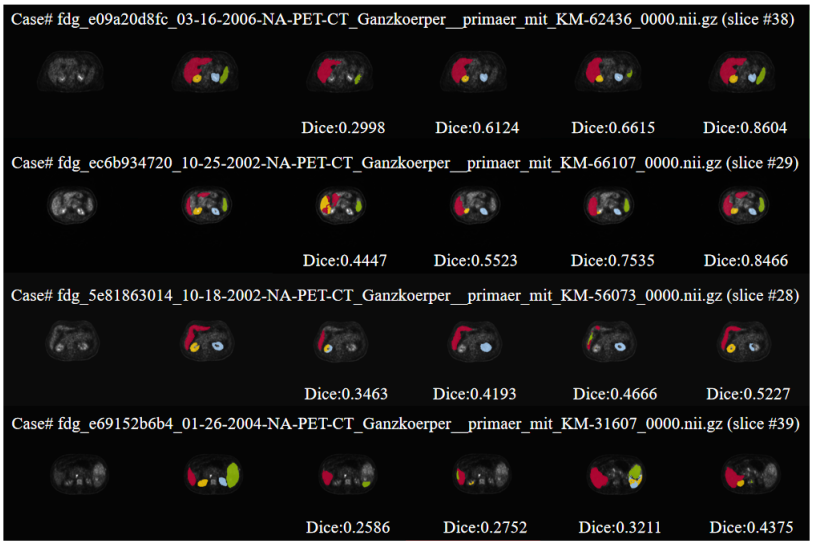

Fig. 6. Visualization of segmentation results of the PET validation set.

of Figures 5 and 6 showcase challenging cases where segmentation performance remains suboptimal across both ablation variants and our complete framework. We attribute these failures to several factors: inadequate image resolution (Figure 5, third row), ambiguous inter-organ boundaries (Figure 5, fourth row), and poor modal contrast (Figure 6, third and fourth rows). Despite these individual challenging cases, quantitative analysis of average performance demonstrates that our method achieves accurate segmentation across the majority of samples, validating the incremental contributions of each proposed component.

4.3 Segmentation efficiency results on validation set

We conducted a performance analysis of the inference process for the validation set using our in-house platform. Our method achieved an average inference time of 3.78 seconds per case on the 110-case MRI validation set, with a peak GPU memory usage of 2,749 MB. It also demonstrated an average inference time of 1.67 seconds per case on the 50-case PET validation set, with a peak GPU memory usage of 2,747 MB. Table 6 presents detailed performance metrics for representative samples, including runtime, maximum GPU memory consumption, and total GPU memory utilization.

4.4 Results on final testing set

This is a placeholder. We will send you the testing results during MICCAI 2025.

Table 6. Quantitative evaluation of segmentation efficiency in terms of the running them and GPU memory consumption. Total GPU denotes the area under GPU Memory-Time curve. Evaluation GPU platform: NVIDIA RTX4090D (24G).

Case ID	Image Size	Running Time (s)	Max GPU (MB)	Total GPU (MB)
amos_0540	(192, 192, 100)	9.87	1461.12	7307.72
amos_7324	(256, 256, 80)	10.12	1557.12	10377.41
amos_0507	(320, 290, 72)	10.16	1695.12	8476.23
amos_7236	(400, 400, 115)	11.62	1621.12	9413.21
amos_7799	(432, 432, 40)	10.10	1057.12	5487.65
amos_0557	(512, 152, 512)	16.51	1639.12	13534.68
amos_0546	(576, 468, 72)	12.38	1639.12	10146.25
amos_8082	(1024, 1024, 82)	18.46	1639.12	18747.32
fdg_605369e88d	(400, 400, 92)	9.76	1639.12	8012.56
fdg_d951eeb735	(400, 400, 58)	9.84	1163.12	5732.67
psma _af293f5b5149087a	(200, 200, 121)	8.12	1639.12	6712.43

4.5 Limitation and future work

The proposed method has several limitations: 1) The MST-Net design is relatively simplistic in detail and should incorporate more advanced research developments. 2) Due to the 2.5D architecture, the style translation process inevitably utilizes only partial slices from each sample, resulting in insufficient target domain style learning and limited generalization capability. 3) We have not explored more lightweight and accurate state-of-the-art segmentation models. 4) PET modality segmentation results show substantial room for improvement, potentially due to the absence of physics-based spacing resampling, causing many cases to fail in identifying the left kidney and spleen. In future work, we will address these limitations to enhance UDA segmentation performance.

5 Conclusion

This work presents an unsupervised domain adaptation framework for cross-modal abdominal multi-organ segmentation in MRI and PET modalities. Our approach integrates a 2.5D Multi-Style Perceptual Translation Network (MST-Net) with self-filtering mechanisms to achieve CT-to-MRI/PET adaptation, demonstrating strong performance on FLARE2025 with 80.77% and 62.83% average DSC for MRI and PET, respectively, while maintaining efficiency with 3.47s inference time and 2479MB peak memory usage. The framework combines multi-style translation, enhanced feature learning through pyramid attention and Large-Small Convolution modules, comprehensive self-filtering, and a two-stage segmentation pipeline. Future work will address these limitations through more sophisticated architectures, full 3D processing capabilities, advanced segmentation models, and domain-specific preprocessing strategies to further enhance cross-modal adaptation performance in clinical applications.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2025 challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers. The proposed solution is fully automatic without any manual intervention. We thank all data owners for making the CT scans publicly available and CodaLab [31] for hosting the challenge platform.

Disclosure of Interests

The authors declare no competing interests.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., Lohöfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., Lipková, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., Hülsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., Jäger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 9
2. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of Digital Imaging **26**(6), 1045–1057 (2013) 9
3. Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., et al.: 3d slicer as an image computing platform for the quantitative imaging network. Magnetic Resonance Imaging **30**(9), 1323–1341 (2012) 9
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. preprint at Research Square (Nature Portfolio) (2023). https://doi.org/https://doi.org/10.21203/rs.3.rs-2572595/v1 9
5. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 9

6. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 9
7. Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging. American Society of Clinical Oncology **38**(6), 626–626 (2020) 9
8. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. Pmlr (2018) 2
9. Huang, Z., Wang, H., Ye, J., Niu, J., Tu, C., Yang, Y., Du, S., Deng, Z., Gu, L., He, J.: Revisiting nnu-net for iterative pseudo labeling and efficient sliding window inference. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation. pp. 178–189. Springer (2022) 3, 4, 8
10. 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
11. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., Luo, P.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022) 1, 9
12. Jiang, J., Hu, Y.C., Tyagi, N., Rimner, A., Lee, N., Deasy, J.O., Berry, S., Veeraraghavan, H.: Psigan: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based mri segmentation. IEEE transactions on medical imaging **39**(12), 4071–4084 (2020) 2
13. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 6
14. Li, J., Chen, Q., Ding, H., Liu, H., Wan, L.: A 3d unsupervised domain adaptation framework combining style translation and self-training for abdominal organs segmentation. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 209–224. Springer (2024) 2
15. Liu, W., Xu, W., Yan, S., Wang, L., Li, H., Yang, H.: Combining self-training and hybrid architecture for semi-supervised abdominal organ segmentation. In: MICCAI challenge on fast and low-resource semi-supervised abdominal organ segmentation, pp. 281–292. Springer (2022) 4, 8, 12
16. Lou, M., Ying, H., Liu, X., Zhou, H.Y., Zhang, Y., Yu, Y.: Sdr-former: A siamese dual-resolution transformer for liver lesion classification using 3d multi-phase imaging. arXiv preprint arXiv:2402.17246 (2024) 4, 9
17. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021) 9
18. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 9

19. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025) 9
20. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Štern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze, P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022) 9
21. Ma, J., Zhang, Y., Gu, S., Ge, C., Ma, S., Young, A., Zhu, C., Meng, K., Yang, X., Huang, Z., Zhang, F., Liu, W., Pan, Y., Huang, S., Wang, J., Sun, M., Xu, W., Jia, D., Choi, J.W., Alves, N., de Wilde, B., Koehler, G., Wu, Y., Wiesenfarth, M., Zhu, Q., Dong, G., He, J., the FLARE Challenge Consortium, Wang, B.: Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. Lancet Digital Health (2024) 4, 9
22. Ma, J., Zhang, Y., Gu, S., Ge, C., Wang, E., Zhou, Q., Huang, Z., Lyu, P., He, J., Wang, B.: Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge. arXiv preprint arXiv:2408.12534 (2024) 9
23. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022) 9
24. Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H.: Mednext: transformer-driven scaling of convnets for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 405–415. Springer (2023) 3, 12
25. Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., Hwang, D.: Sdc-uda: Volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7412–7421 (2023) 2
26. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 9
27. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014) 2
28. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Lsnet: See large, focus small. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9718–9729 (2025) 3, 7
29. Wang, E., Zhao, Y., Wu, Y.: Cascade dual-decoders network for abdominal organs segmentation. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation. pp. 202–213. Springer (2022) 3, 4
30. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 9

31. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 15
32. Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S.: Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 255–263. Springer (2019) 2
33. Yao, K., Su, Z., Huang, K., Yang, X., Sun, J., Hussain, A., Coenen, F.: A novel 3d unsupervised domain adaptation framework for cross-modality medical image segmentation. IEEE Journal of Biomedical and Health Informatics **26**(10), 4976–4986 (2022) 2, 6, 11
34. Yushkevich, P.A., Gao, Y., Gerig, G.: Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3342–3345 (2016) 9
35. Zhang, Z., Peng, Y., Duan, X.: Amots: Partially supervised framework for abdominal multi-organ and tumor segmentation via aspect-aware complementary. Artificial Intelligence in Medicine **168**, 103224 (2025) 1
36. Zhang, Z., Peng, Y., Duan, X., Hou, Q., Li, Z.: Dual-axis generalized cross attention and shape-aware network for 2d medical image segmentation. Biomedical Signal Processing and Control **107**, 107791 (2025) 1

Table 7. Checklist Table. Please fill out this checklist table in the answer column.

Requirements	Answer
A meaningful title	Yes
The number of authors (≤6)	4
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Figure.1
Pre-processing	Pages 4-5
Strategies to use the partial label	Pages 5-7
Strategies to use the unlabeled images.	Pages 7-8
Strategies to improve model inference	Pages 8-9
Post-processing	Page 9
The dataset and evaluation metric section are presented	Pages 9-10
Environment setting table is provided	Table.1
Training protocol table is provided	Table.2
Ablation study	Pages 11-12
Efficiency evaluation results are provided	Table.6
Visualized segmentation example is provided	Figure.5, 6
Limitation and future work are presented	Yes
Reference format is consistent.	Yes