Unsupervised Domain Adaptation for Cross-modality Abdominal Organ Segmentation via Organ Attention Style Transfer and Dual-stage Pseudo Label Filtering

Huamin Wang[1][0009−0001−7872−6499], Jianghao Wu[2][0000−0001−9743−9316], Guotai Wang[1][0000−0002−8632−158X], Xianhao Zhou[1][0009−0001−2014−3098], and Jinlong He[1][0009−0009−8891−554X]

[1] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, 611731, China
[2] Monash University, Australia
jianghao.wu@monash.edu

Abstract. Unsupervised domain adaptation (UDA) for abdominal organ segmentation from labeled Computed Tomography (CT) to unlabeled Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) presents a significant challenge due to large cross-modality gaps. To address this, we propose a comprehensive multi-stage framework that synergistically combines structure-preserving image synthesis, a robust segmentation architecture, and an advanced self-training pipeline. Initially, we leverage an Organ Attention CycleGAN to synthesize anatomically-faithful MRI and PET images from labeled CTs. These synthetic images first train a Coarse-to-Fine segmentation network, which is then refined through a sophisticated self-training scheme. This scheme features a novel dual-stage pseudo-label filtering pipeline that first selects plausible samples based on anatomical consistency and then generates high-precision consensus labels via model ensembling. Evaluated on the FLARE 2025 Task 3 validation set, our complete framework achieves a mean Dice score of 81.21% on MRI and 81.43% on PET, demonstrating the efficacy of our approach in bridging the domain gap without requiring any target-domain annotations.

Keywords: Unsupervised domain adaptation · Abdominal organ segmentation · Dual stage pseudo filtering

1 Introduction

Abdominal organ segmentation in medical images is a cornerstone for numerous clinical applications, including computer-aided diagnosis, surgical planning, and radiotherapy treatment [26]. While deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved remarkable success in this area, their performance heavily relies on large-scale, accurately annotated datasets [18].

The majority of these advancements have been concentrated on Computed Tomography (CT) imaging, for which annotated data is relatively abundant. However, other imaging modalities like Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) offer unique and complementary clinical advantages. MRI provides superior soft-tissue contrast without ionizing radiation, while PET offers crucial functional and metabolic information, making multi-modal segmentation a task of significant clinical value [19].

Despite their importance, the development of robust segmentation models for MRI and PET is severely hampered by the scarcity of annotated data. As exemplified by this year's FLARE 2025 challenge, the task involves segmenting organs in both MRI and PET scans, yet publicly available, large-scale annotated datasets for these modalities remain rare. This data imbalance presents a formidable challenge: how can we leverage the rich annotations from the CT domain (source) to build high-performance segmentation models for unlabeled MRI and PET scans (targets)? This problem falls into the realm of Unsupervised Domain Adaptation (UDA), where the goal is to overcome the significant "domain gap" between the source and target modalities, which arises from fundamentally different imaging physics and appearance characteristics [3].

Image-to-image translation, often powered by Generative Adversarial Networks (GANs) like CycleGAN [28], has emerged as a popular UDA strategy. By synthesizing target-style images from labeled source images, one can train a segmentation model in a fully supervised manner. However, a critical drawback of standard translation methods is their tendency to distort or lose fine anatomical details during the cross-modality synthesis process, which is detrimental to segmentation accuracy [24]. To mitigate this issue, we previously developed an Organ Attention CycleGAN, which incorporates a lightweight segmentation task head to guide the generator, ensuring better preservation of anatomical structures during translation [25]. Furthermore, given the inherent complexity of accurately delineating multiple organs, a direct end-to-end segmentation approach can be suboptimal. A multi-stage strategy, such as the coarse-to-fine framework, the winning solution of the FLARE 2024 challenge [11], has demonstrated superior performance by first localizing the region of interest and then performing precise segmentation.

Building upon these foundations, this work tackles the multi-target UDA problem for abdominal organ segmentation from CT to both MRI and PET. We hypothesize that by combining our structure-preserving synthesis method with a robust coarse-to-fine segmentation pipeline, and further enhancing it with a sophisticated self-training mechanism, we can effectively bridge the cross-modality gap without requiring any annotations for the target MRI and PET domains.

The main contributions of this work are as follows:

- We employ Organ Attention CycleGAN framework with lightweight segmentation task head for cross-modality image translation, which effectively preserves anatomical structures during the translation from CT to various MRI and PET modalities.

- We demonstrate the effectiveness of Coarse-to-Fine segmentation framework that first localizes the region of interest and then performs fine segmentation, emphasizing the importance of a dual-stage approach for fully supervised training and pseudo label generating in self-training.
- We propose a dual-stage pseudo label filtering pipeline that first filtering reliable samples using a multi-strategy approach and then aggregating consensus labels at pixel-wise level, significantly improving segmentation accuracy on real-world MRI and PET data.

2 Method

Our proposed methodology for unsupervised multi-modal abdominal organ segmentation is structured as a comprehensive pipeline (Fig. 1) designed to bridge the significant domain gap between labeled CT scans and unlabeled MRI and PET scans. The framework is organized into three main stages, each corresponding to a core contribution of our work. First, we address the cross-modality image synthesis challenge by employing our Organ Attention CycleGAN framework. This stage focuses on generating high-fidelity, structure-preserving synthetic MRI and PET images from CT data, forming the foundation for supervised training. Second, we utilize these synthetic images to train a robust segmentation model based on a Coarse-to-Fine architecture, which enhances accuracy by first localizing organ regions and then performing precise delineation. Finally, to adapt the model to real-world data, we introduce a sophisticated self-training scheme that leverages a novel dual-stage pseudo-label filtering pipeline. This final stage iteratively refines the model's performance on the unlabeled target domains. The subsequent subsections will elaborate on each of these components in detail.

2.1 Style Transfer with Organ Attention

A model trained on labeled source images (CT), $\mathcal{D}_s$, performs poorly on unlabeled target domains (MRI/PET), $\mathcal{D}_t$, due to the significant domain shift [23]. While CycleGAN [28] is a common unsupervised domain adaptation approach using adversarial ($\mathcal{L}_{gan}$) and cycle-consistency ($\mathcal{L}_{cyc}$) losses, it often introduces anatomical distortions that are detrimental to medical segmentation tasks.

To mitigate this, we leverage our Organ Attention CycleGAN [25], which integrates a task-specific guidance mechanism to explicitly preserve organ structures. The core idea is to generate a spatial attention map for each source image X_i^s using a pre-trained segmentation network S. This map, $A(X_i^s)$, highlights all foreground organ regions by effectively subtracting the predicted background probability from one:

$$A(X_i^s) = 1 - \mathrm{softmax}(S(X_i^s)) \tag{1}$$

This attention map is then fused with the translated image, forcing the generator to focus on preserving structural details within these highlighted areas during synthesis. This process ensures the generation of an anatomically reliable dataset for training the downstream segmentation network.

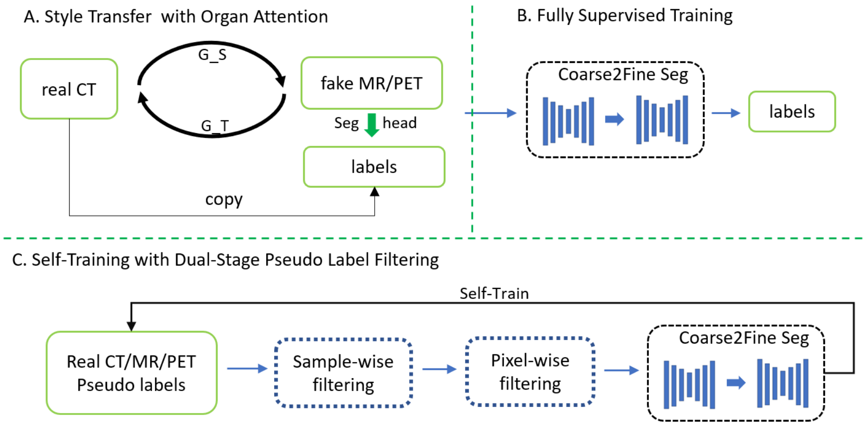

Fig. 1. Overview of Unsupervised Domain Adaptation for Cross-modality Abdominal Organ Segmentation via Organ Attention Style Transfer and Dual-stage Pseudo Label Filtering.

2.2 Fully Supervised Training with Coarse-to-Fine Framework

We train the segmentation model on our synthetic data using a two-stage Coarse-to-Fine framework [11], which effectively decomposes the task into localization and then precise delineation. First, a "coarse" model trained on down-sampled 3D volumes identifies a Region of Interest (ROI) encompassing all target organs. Second, a "fine" model performs detailed, multi-class segmentation on high-resolution patches cropped from within this ROI. The entire pipeline is implemented using the powerful and automated `nnU-Net` v2 framework [9], providing a robust foundation for the subsequent self-training phase.

2.3 Self-Training with Dual-Stage Pseudo Label Filtering

While the models trained on synthetic data provide a strong baseline, a performance gap to real-world data invariably remains due to the imperfect nature of image synthesis. To bridge this final gap and adapt our models to the true data distributions of unlabeled MRI and PET scans, we introduce a sophisticated self-training pipeline. The core of this pipeline is a dual-stage pseudo-label filtering mechanism designed to iteratively generate high-quality pseudo-labels from the unlabeled target domain data. The process flows from initial pseudo-label generation to sample-wise filtering, then to pixel-wise filtering, and finally, the refined labels are used to retrain our Coarse-to-Fine segmentation models.

Stage 1: Sample-Wise Filtering for Plausibility. The first stage aims to filter out entire image samples whose pseudo-labels are anatomically implausible or generated with low confidence. For each unlabeled image X_j^t (preprocessed to match the median spacing of the source CT dataset), we first generate an initial pseudo-label $\tilde{Y}_j$ using our best-performing model trained on synthetic data. Each sample is then subjected to two key filtering criteria:

- **Anatomical Volume Consistency:** We check if the predicted organ sizes are realistic. For each foreground organ class c, we compute its *relative volume* R_c, defined as the ratio of the organ's volume to the volume of the minimal 3D bounding box encompassing all foreground organs.

$$R_c(\tilde{Y}_j) = \frac{\mathcal{V}(\tilde{Y}_j^c)}{\mathcal{V}\left(\mathcal{B}\left(\bigcup_{k\in \mathrm{fg}} \tilde{Y}_j^k\right)\right)}, \tag{2}$$

  where $\mathcal{V}(\cdot)$ is the volume operator, $\mathcal{B}(\cdot)$ computes the minimal bounding box, and fg is the set of all foreground classes. We pre-calculate the mean (μ_c) and standard deviation (σ_c) of these relative volumes from the ground truth labels in the source CT dataset $\mathcal{D}_s$. A pseudo-labeled sample $\tilde{Y}_j$ is considered plausible only if the relative volume of every organ falls within a trusted interval:

$$\forall c \in \mathrm{fg}, \quad \mu_c - 3\sigma_c \leq R_c(\tilde{Y}_j) \leq \mu_c + 3\sigma_c. \tag{3}$$

- **Prediction Confidence Score:** To measure model confidence, we use the top two checkpoints from our validation set to generate two independent predictions for each sample. We then compute the mean Dice score across all foreground classes between these two predictions. A sample is retained only if it is among the top-k ranked samples (e.g., top 100) and its mean Dice score exceeds a high threshold (e.g., 0.9 for MRI, 0.82 for PET), indicating strong agreement between the models.

Stage 2: Pixel-Wise Filtering for Consensus. Samples that pass the initial filtering stage are considered reliable at a macroscopic level. The second stage refines these labels at the pixel level to generate a high-precision consensus label. To achieve this, we employ an ensemble of N diverse models. This ensemble includes multiple checkpoints from our training runs, models trained with different data augmentations (e.g., non-linear transformations, etc), and pseudo labels from the FLARE22 winning algorithm [8] and the best-accuracy-algorithm [21].

For each voxel v in a selected sample, the final pseudo-label for a class c, denoted $\hat{Y}_j^c(v)$, is determined by a majority vote across the N model predictions:

$$\hat{Y}_j^c(v) = \mathbb{I}\left(\sum_{n=1}^{N} \tilde{Y}_{j,n}^c(v) > \frac{N}{2}\right), \tag{4}$$

where $\tilde{Y}_{j,n}^c(v)$ is the binary prediction for class c at voxel v from the n-th model in the ensemble, and $\mathbb{I}(\cdot)$ is the indicator function. This process effectively filters out noisy or uncertain predictions at the pixel level, resulting in a cleaner and more accurate set of consensus labels.

This dual-stage filtering pipeline yields a high-confidence pseudo-label dataset, $\mathcal{D}_{PL}$. This refined dataset is then used to retrain our Coarse-to-Fine segmentation models, completing one cycle of self-training and significantly improving segmentation performance on real-world MRI and PET data.

2.4 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 [13].

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 [20], KiTS [6,7], autoPET [5,4], AMOS [10], LLD-MMRI [12], TotalSegmentator [22], and AbdomenCT-1K [17], and past FLARE Challenges [14,15,16].

Training Set: The training set includes 2050 labeled CT scans, 4817 unlabeled MRI scans and unlabeled 1000 PET scans. Among CT scans, 50 cases are provided with high-quality, manually-curated ground-truth segmentation masks. The remaining 2,000 cases are annotated with reliable pseudo-labels, which were generated by the FLARE22 winning algorithm [8] and the best-accuracy-algorithm [21].

Validation Set: The public validation set is multi-modal and comprises a total of 160 cases, including 110 MRI and 50 PET scans. This set is used for hyperparameter tuning, model selection, and evaluating the effectiveness of our domain adaptation and self-training strategies before the final testing phase.

Evaluation Measures: We use Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD) for accuracy evaluation, and running time and area under the GPU memory-time curve for model efficiency. Furthermore, the running time for each case should within 60 seconds.

3.2 Implementation details

Environment settings: The development environments and requirements are presented in Table 1.

Preprocessing: For the style transfer framework, we adopted the preprocessing steps from our previous work on 3D medical images[25]. Specifically, CT images were clipped to a window of [-600, 600], while MRI and PET scans were clipped

Table 1. Development environments and requirements.

System	Ubuntu 20.04.3 LTS
CPU	Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz
RAM	125GB
GPU (number and type)	NVIDIA GeForce RTX 2080 Ti (11GB)
CUDA version	11.8
Programming language	Python 3.10.4
Deep learning framework	PyTorch 2.3.0+cu118, torchvision 0.18.0
Specific dependencies	antspyx, einops, etc

to intensity values between the 1st and 99th percentiles. All images were then linearly normalized to the range [-1, 1]. For the segmentation network, we followed the same preprocessing as in[11], including cropping, z-score normalization, and gamma transformation. When generating pseudo labels for self-training, we resampled all unlabeled MRI and PET images to the median spacing of the labeled CT dataset to mitigate spatial discrepancies across datasets.

Training protocols: The training parameters are listed in Table 2. Additional training details for style transfer and segmentation can be found in our previous works on Organ Attention[25] and the two-stage framework [11].

Post-processing: Post-processing, including connected component analysis and test-time augmentation, remains identical to that in last year's winning solution [11].

Table 2. Training protocols.

Network initialization	
Batch size	1
Patch size	96×192×192
Total epochs	200
Optimizer	AdamW
Initial learning rate (lr)	1e-5, 1e-6(self-train)
Lr decay schedule	Cosine Annealing LR
Training time	10 hours
Loss function	Cross entropy + Dice
Number of model parameters	4.2M[3]
Number of flops	251.19G[4]

Table 3. Quantitative evaluation results of MRI and PET scans on the validation dataset.

Target	MRI		PET	
	DSC(%)	NSD(%)	DSC(%)	NSD (%)
Liver	93.77 ± 2.29	95.50 ± 3.75	88.41 ± 4.17	79.50 ± 11.78
Right Kidney	94.55 ± 3.33	95.54 ± 5.09	78.45 ± 11.72	68.33 ± 15.24
Left Kidney	95.32 ± 1.99	96.96 ± 2.59	80.26 ± 13.00	73.98 ± 14.66
Spleen	93.68 ± 9.24	96.06 ± 9.92	78.59 ± 14.99	65.97 ± 18.91
Pancreas	80.79 ± 10.57	92.12 ± 11.09	–	–
Aorta	91.48 ± 8.76	94.99 ± 8.88	–	–
Inferior Vena Cava	85.65 ± 7.38	90.23 ± 8.00	–	–
Right Adrenal Gland	60.19 ± 15.77	77.06 ± 16.80	–	–
Left Adrenal Gland	66.02 ± 19.18	79.46 ± 21.67	–	–
Gallbladder	77.40 ± 24.37	73.95 ± 24.84	–	–
Esophagus	68.29 ± 14.62	85.21 ± 15.05	–	–
Stomach	82.74 ± 16.01	86.95 ± 17.35	–	–
Duodenum	65.88 ± 13.61	87.02 ± 12.93	–	–
Average	81.21 ± 12.11	88.54 ± 7.47	81.43 ± 4.09	71.94 ± 5.24

Table 4. Sample-wise Pseudo Label Selection Quantities for Unlabeled Dataset. Where (N*) represents the number of modalities, and $\underline{\text{N}}$ denotes approximate quantities.

Strategy	CT	MR-AMOS	MR-LLD(8*)	PET(2*)
total	2000	800	4000	1000
+ class number	$\underline{1500}$	$\underline{500}$	$\underline{3700}$	$\underline{500}$
+ relative volume	$\underline{900}$	$\underline{260}$	$\underline{2200}$	$\underline{140}$
+ top 2 sorted dice	100	100	618	14

Table 5. Performance metrics for different baselines on MR and PET, where C2F: Coarse-to-Fine, OA: Organ Attention, DSF: Dual-Stage Filtering for pseudo labels, RCT: Real CT, FI: Fake Images, PL: Pseudo Label.

Baseline	Strategy			Training Data			MR		PET	
	C2F	OA	DSF	RCT	FI	PL	DSC (%)	NSD (%)	DSC (%)	NSD (%)
baseline 1	√			√			80.05	87.08	79.26	69.14
baseline 2	√	√		√	√		80.51	87.78	80.87	71.14
ours	√	√	√	√	√	√	81.21	88.54	81.43	71.94

4 Results and discussion

4.1 Quantitative results on validation set

We evaluated our framework on the FLARE 2025 validation set. Quantitative results are summarized in Table 3, detailing the final organ-wise performance, and Table 5, which presents a progressive ablation study.

Our final model (Table 3) demonstrates strong performance, achieving excellent DSC scores for large organs on MRI (e.g., 95.32% for Left Kidney) but predictably lower scores for smaller structures (e.g., 60.19% for Right Adrenal Gland). On PET, the model also performs well, reaching a DSC of 88.41% for the Liver. Overall, the final model achieves an average DSC of 81.21% on MRI and 81.43% on the evaluated PET organs.

The ablation study in Table 4 and Table 5 confirms the efficacy of our components. A baseline Coarse-to-Fine (C2F) model trained only on real CT data achieves 80.05% DSC on MR and 79.26% on PET. Integrating our Organ Attention (OA) mechanism for synthetic image generation boosts performance, particularly on PET (80.87% DSC). Finally, applying our Dual-Stage Filtering (DSF) in a self-training loop further improves the scores to our final results of 81.21% (MR) and 81.43% (PET). This incremental improvement validates the significant contribution of both the structure-preserving synthesis and the pseudo-label filtering pipeline.

4.2 Qualitative results on validation set

Figure 2 presents two examples of successful segmentation and two examples of poor segmentation from the validation set. Furthermore, undersegmented foreground regions are indicated by green arrows, and false positive segmentation regions are enclosed in yellow boxes.

4.3 Segmentation efficiency results on validation set

We also present efficiency metrics in Table 6, including running time and GPU memory consumption. These metrics were evaluated on an NVIDIA GeForce RTX 2080 Ti GPU with 11 GB of memory.

4.4 Results on final testing set

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

4.5 Limitation and future work

The current UDA approach for style transfer involves training separate CycleGAN models for each MRI and PET modality to preserve modality-specific anatomical features. While effective, this design incurs additional computational

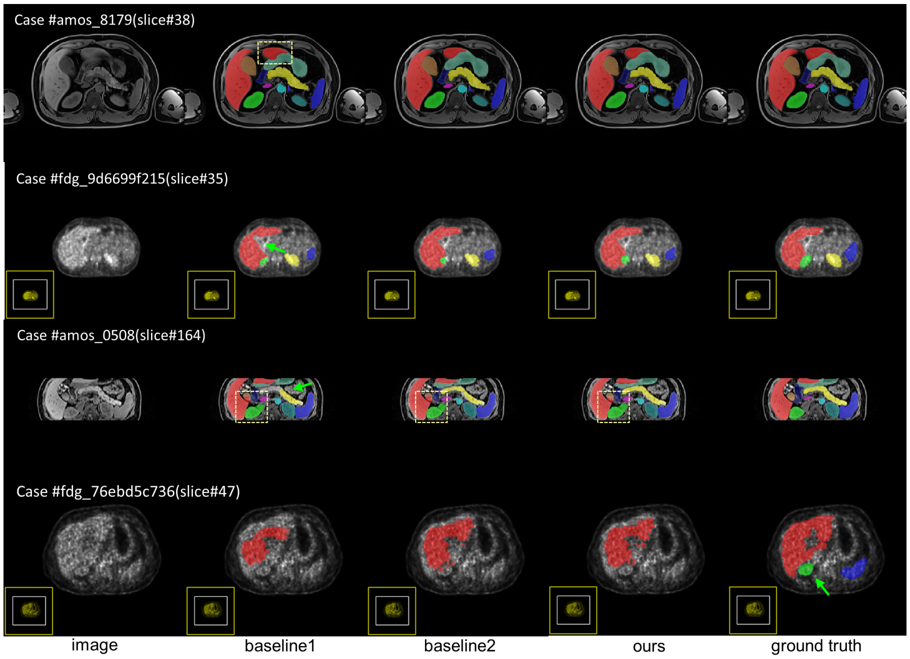
Case #amos_8179(slice#38)
Case #fdg_9d6699f215(slice#35)
Case #amos_0508(slice#164)
Case #fdg_76ebd5c736(slice#47)
image
baseline1
baseline2
ours
ground truth

Fig. 2. Visualization of two examples with good segmentation results and two examples with bad segmentation results in the validation set.

Table 6. Quantitative evaluation of segmentation efficiency in terms of the running time and GPU memory consumption. Total GPU denotes the area under GPU Memory-Time curve. Evaluation GPU platform: NVIDIA GeForce RTX 2080 Ti (11GB).

Case ID	Image Size	Running Time (s)	Max GPU (MB)	Total GPU (MB)
amos_0540	(192, 192, 100)	13.51	3083	13075
amos_7324	(256, 256, 80)	13.97	2641	13728
amos_0507	(320, 290, 72)	13.76	2641	12880
amos_7236	(400, 400, 115)	14.48	2641	13985
amos_7799	(432, 432, 40)	13.99	2641	13438
amos_0557	(512, 152, 512)	17.14	3083	17185
amos_0546	(576, 468, 72)	15.69	3079	15668
amos_8082	(1024, 1024, 82)	28.37	3083	31920
fdg_605369e88d	(400, 400, 92)	14.27	2837	13084
fdg_d951eeb735	(400, 400, 58)	14.30	3081	13310
psma_af293f5b5149087a	(200, 200, 121)	13.40	2641	11855

costs and may limit scalability. Furthermore, the dual-stage pseudo label filtering is an offline strategy, rendering it unsuitable for real-time self-improvement during the self-training phase. In the future, integrating a self-training pipeline that dynamically improves pseudo label quality could yield more precise segmentation.

5 Conclusion

In this work, we presented a comprehensive unsupervised domain adaptation framework that combines structure-preserving image synthesis, a coarse-to-fine segmentation strategy, and a novel dual-stage pseudo-label filtering pipeline to segment abdominal organs in unlabeled MRI and PET scans. Our extensive experiments demonstrated that each component provides significant, incremental performance gains, validating the effectiveness of our multi-stage design. Ultimately, this synergistic approach successfully bridges the large domain gap between CT, MRI, and PET, yielding a robust segmentation model without requiring any annotations in the target domains.

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 [27] 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) 6

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) 6
3. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis **83**, 102628 (2023) 2
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 6
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) 6
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) 6
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) 6
8. 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) 5, 6
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) 4
10. Ji, Y., Bai, H., GE, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhang, 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) 6
11. 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, 4, 7
12. 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) 6

13. 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) 6
14. 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) 6
15. 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) 6
16. 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) 6
17. 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) 6
18. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al.: Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 1
19. Savenije, M.H., Maspero, M., Sikkes, G.G., van der Voort van Zyp, J., Kotte, T., Alexis, N., Bol, G.H., van den Berg, T., Cornelis, A., et al.: Clinical implementation of mri-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiation oncology **15**(1), 1–12 (2020) 2
20. 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) 6
21. 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) 5, 6
22. 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) 6
23. Wu, J., Gu, R., Dong, G., Wang, G., Zhang, S.: FPL-UDA: Filtered Pseudo Label-Based Unsupervised Cross-Modality Adaptation for Vestibular Schwannoma Segmentation. In: ISBI. pp. 1–5. IEEE (2022) 3
24. Wu, J., Guo, D., Wang, L., Yang, S., Zheng, Y., Shapey, J., Vercauteren, T., Bisdas, S., Bradford, R., Saeed, S., et al.: TISS-Net: Brain tumor image synthesis and segmentation using cascaded dual-task networks and error-prediction consistency. Neurocomputing p. 126295 (2023) 2

25. Wu, J., Zhang, G., Qi, X., Wang, H., Liu, X., Wang, G.: Unsupervised domain adaptation for abdominal organ segmentation using pseudo labels and organ attention cyclegan. In: MICCAI Challenge on Fast and Low-Resource Semi-supervised Abdominal Organ Segmentation, pp. 225–242. Springer (2024) 2, 3, 6, 7
26. Xu, X., Chen, Y., Wu, J., Lu, J., Ye, Y., Huang, Y., Dou, X., Li, K., Wang, G., Zhang, S., et al.: A novel one-to-multiple unsupervised domain adaptation framework for abdominal organ segmentation. Medical Image Analysis **88**, 102873 (2023) 1
27. 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) 11
28. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017) 2, 3

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

Requirements	Answer
A meaningful title	Yes/No
The number of authors (≤6)	5
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	Page 6,7
Strategies to use the partial label	Page 3-5
Strategies to use the unlabeled images.	Page 3-5
Strategies to improve model inference	Page 4-7
Post-processing	Page 7
The dataset and evaluation metric section are presented	Page 6
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	Page 8,9
Efficiency evaluation results are provided	Table 6
Visualized segmentation example is provided	Figure 2
Limitation and future work are presented	Yes
Reference format is consistent.	Yes