A Novel Pseudo Label-based Unsupervised Multiple Target Domain Adaptation Framework for Abdominal Organ Segmentation

Yuntao Zhu[1(✉)][0000−0003−2816−2709], Liwen Zou[1][0000−0003−4085−4003], Pengxu Wen[1][0009−0000−5211−4876], Ziwei Nie[1], Luying Gui[2], and Xiaoping Yang[1]

[1] Department of Mathematics, Nanjing University, Nanjing, China
YuntaoZhu7@smail.nju.edu.cn

[2] Department of Mathematics, Nanjing University of Science and Technology, Nanjing, China

Abstract. Obtaining labeled data from medical images is very expensive and labor intensive. We consider the problem of unsupervised domain adaptation (UDA) in cross-modality medical image segmentation, aiming to perform segmentation on the unannotated target domain with the help of labeled source domain. We present a framework that combines unsupervised domain adaptation, registration, and pseudo-label learning to effectively and flexibly adapt to multiple target domains. We introduce a novel image translation method based on anatomy space and a novel operation of matching and registration to improve pseudo-labels, which effectively mitigates the large cross-modality domain gap. Experiments demonstrate that our method achieves the average Dice Similarity Coefficient of 0.6053 and Normalized Surface Dice of 0.6462 on 13 abdominal organ segmentation tasks. Moreover, it significantly improves the inference speed, with an average running time of 20.7 seconds, and uses only an average of 1107332.6 MB of total GPU memory on final testing set.

Keywords: Unsupervised domain adaptation · Accelerate inference · Lightweight network.

1 Introduction

Deep learning has revolutionized the field of medical image segmentation, enabling accurate segmentation of various anatomical structures and lesions [21]. However, well-trained deep models usually perform poorly in real scenarios due to the severe data distribution difference between training and test sets caused by different imaging modalities, scanning protocols, and/or demographic properties. Domain adaptation (DA) is a promising solution to address the problem of dramatic performance degradation across modalities at inference time. It attempts to establish a mapping between the source and target domains so that models trained in the source domain can perform well in the target domain. Among DA, unsupervised domain adaptation (UDA) is more attractive and feasible, where the ground truth in the target domain is not required. Various methods have been proposed to deal with this problem by aligning the source and target domains in terms of image appearance, feature distribution, or output structure. CycleGAN [30] and Contrastive Unpaired Translation (CUT) [20], focus on aligning image appearance between the source and target domains. However, these methods may introduce distortions to the anatomical structure of the images, which can hinder accurate segmentation [26]. Long et al. [11] focused on alignment at the feature level. FPL+ [25] aim to align the domains at both the image and feature levels by using pseudo labels. However, an overly simplified assumption for style transfer, which is a commonly used strategy of image alignment, is adopted in most previous work, where they model this process as a deterministic one-to-one mapping with a mean style of each domain. In fact, image styles in a single domain may are quite different [28].

In this work, we propose a multiple target domain adaptation framework for UDA in 3D medical image segmentation. First, an Anatomy Space Image Translation (ASIT) using CycleGAN enables the alignment of CT and MRI images, which builds a bi-directionality mapping between CT and MRI. In this way, we can eliminate the impact of the domain shift on networks. Second, we can train a segmentation network using the fake MRI generated by CT and ground truth. Then, The trained network infers real MRI to gain pseudo-labels. Since the MRI data have multiple sequences, e.g. T1, T2, DWI, some sequences show better pseudo-label quality than others. We conduct a mixed operation (multi-domain matching, MDM) of matching and registration to improve pseudo-labels. Third, after we own an MRI dataset with high-quality pseudo-labels, we do self-training to enhance network performance by filtering some bad cases. Finally, we use the selected MRI dataset to train a small network and do a fast inference process, that can speed the inference and reduce computational resources.

The contributions of this work are summarized in three aspects:

- We propose a novel UDA framework 3D medical image segmentation based on generating high-quality pseudo labels in the multiple target domains.
- We introduce a novel image translation method (ASIT) based on Anatomy Space and a novel operation (MDM), which effectively mitigates the large cross-modality domain gap.
- The experiment shows that our method improves the ability of segmentation networks for the unsupervised domains adaptation. This outperforms the single target domain adaptation by 23 percentage points for the Dice Similarity Coefficient (DSC).

2 Method

As shown in Figure 1, our method has four stages: 1) Image translation using CycleGAN enables the alignment of CT and MRI images, which builds a bi-directionality mapping between CT and MRI. In this way, we can eliminate the impact of the domain shift on networks. 2) We can train a segmentation network using the fake MRI generated by CT and ground truth. Then, The

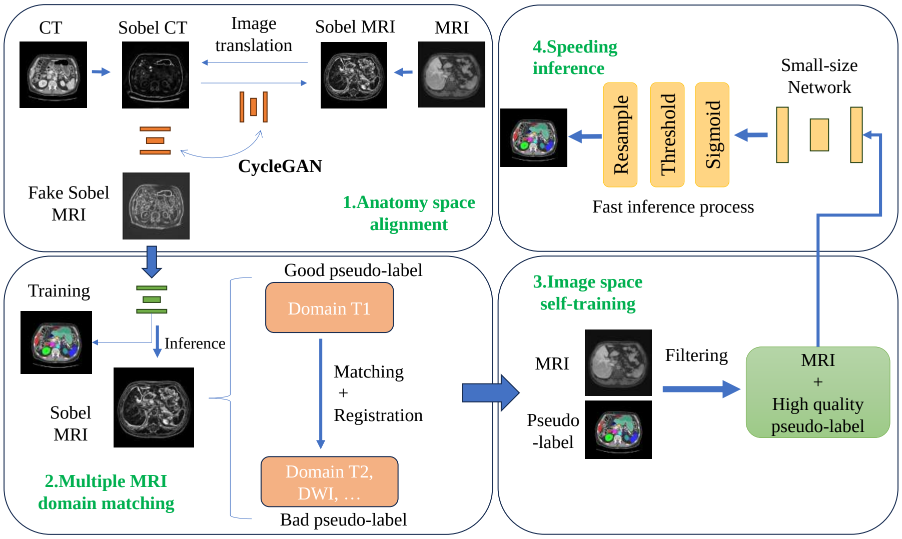


Fig. 1: The proposed method framework.

trained network infers real MRI to gain pseudo-labels. Since the MRI data have multiple sequences, e.g. T1, T2, DWI, some sequences show better pseudo-label quality than others. We conduct a mixed operation of match and registration to improve pseudo-labels. 3) After we own an MRI dataset with high-quality pseudo-labels, we do self-training to enhance network performance by filtering some bad cases. 4) Finally, we use the selected MRI dataset to train a small network and do a fast inference process, that can speed the inference and reduce computational resources. In addition, we used the pseudo labels generated by the FLARE22 winning algorithm [8] and the best-accuracy-algorithm [23] as labels of CT.

2.1 Anatomy space image translation

To achieve UDA for 3D medical image segmentation, our idea first obtains high-quality pseudo labels for training images in the target domain and then trains a segmentation model in that domain by learning from pseudo labels. We find an interesting fact that CT and MRI have consistent anatomic structures in the body. Undoubtedly, an organ's shape, such as the liver, kidneys, etc, does not change with the modal changing, it is an invariant feature. Utilizing this feature, we can build a network to transfer segmentation ability between various modalities.

Let $\mathcal{D}_s$ and $\mathcal{D}_t$ denote a set of labeled source-domain images and a set of unlabeled target-domain images, respectively. Let X_i^s and X_j^t denote the i-th image from $\mathcal{D}_s$ and the j-th image from $\mathcal{D}_t$, respectively, where the label of X_i^s is Y_i^s. Note that the source domain and target domain are from different patient groups, i.e., X_i^s and X_j^t are unpaired. Due to the domain shift between $\mathcal{D}_s$ and $\mathcal{D}_t$, training a model with $\mathcal{D}_s$ to generate pseudo labels for $\mathcal{D}_t$ will lead to a poor performance. To improve the quality of pseudo labels for $\mathcal{D}_t$, we propose Anatomy Space Image Translation (ASIT) to augment $\mathcal{D}_s$ before training the pseudo label generator.

Specifically, we utilize an image style translator T_t to translate a labeled source-domain image X_i^s into a pseudo target-domain image $X_i^{s\to t} = T_t(X_i^s)$, and use another image style translator T_s to translate $X_i^{s\to t}$ back to the source domain, leading to a pseudo source-domain image $X_i^{s'} = T_s(X_i^{s\to t})$. Note that T_t and T_s are often trained jointly for learning from unpaired training sets, as used in CycleGAN [30]. As the training sets are unpaired, it is difficult to make $X_i^{s\to t}$ and $X_t^{s'}$ exactly match the ground truth target-modality and source-modality images, respectively. However, we employ the Sobel filter on CT and MRI first to obtain Sobel-CT and Sobel-MRI, the modality difference achieves a large degree of reduction. As illustrated in Figure 2, the Anatomy Space Image Translation shows a good performance.

In this work, the image translators T_s and T_t are implemented based on CycleGAN [30] with two discriminators D_s and D_t for the two domains, respectively. The training involves two adversarial losses $\mathcal{L}_{gan}^t$, $\mathcal{L}_{gan}^s$, a cycle consistency loss $\mathcal{L}_{cyc}$ and an identity mapping loss $\mathcal{L}_{\text{identity}}$. The target-domain adversarial loss $\mathcal{L}_{gan}^t$ is:

$$\begin{aligned}\mathcal{L}_{gan}^t(T_t, D_t) =& \mathbb{E}_{X_j^t \sim \mathcal{D}_t}[\log D_t(X_j^t)] \\ &+ \mathbb{E}_{X_i^s \sim \mathcal{D}_s}[\log(1 - D_t(X_i^{s\to t}))].\end{aligned} \tag{1}$$

The source-domain adversarial loss $\mathcal{L}_{gan}^s$ is defined similarly based on T_s, D_s and $X_i^{s'}$, and the consistency loss is:

$$\begin{aligned}\mathcal{L}_{cyc}(T_s, T_t) =& \mathbb{E}_{X_i^s \sim \mathcal{D}_s}[\|T_s(T_t(X_i^s)) - X_i^s\|_1] \\ &+ \mathbb{E}_{X_j^t \sim \mathcal{D}_t}[\|T_t(T_s(X_j^t)) - X_j^t\|_1].\end{aligned} \tag{2}$$

Following in the footsteps of CycleGAN, we utilize the identity mapping loss, which is:

$$\mathcal{L}_{\text{identity}}(T_s, T_t) = \mathbb{E}_{X_i^s \sim \mathcal{D}_s}[\|T_t(X_i^s) - X_i^s\|_1] + \mathbb{E}_{X_j^t \sim \mathcal{D}_t}[\|T_s(X_j^t) - X_j^t\|_1]. \tag{3}$$

Our full objective function is:

$$\begin{aligned}\mathcal{L}_{full}(T_s, T_t, D_s, D_t) =& \mathcal{L}_{gan}^s(T_s, D_s) + \mathcal{L}_{gan}^t(T_t, D_t) \\ &+ \lambda_1 \mathcal{L}_{cyc}(T_s, T_t) + \lambda_2 \mathcal{L}_{\text{identity}}(T_s, T_t),\end{aligned} \tag{4}$$

where λ_1 and λ_2 are weighting parameters.

2.2 Pseudo label by multi-domain matching

We train a segmentation network using pseudo target-domain training set $\mathcal{D}_{st} = \{(X_i^{s\to t}, Y_i^s))\}$. Then we generate pseudo-label $\bar{Y}_i^t$ by inference on Sobel-MRI sample X_i^t. Due to an observed fact, the pseudo-label quality of the MRI T1 sequence is generally better than other MRI sequences. The MRI dataset should be divided into several sub-datasets, i.e. there are multi-target domains. For simplicity, we split MRI dataset $\mathcal{D}_t$ into two sub-datasets $\mathcal{D}_{t1}, \mathcal{D}_{t2}$, where $\mathcal{D}_{t1}$ have better pseudo-label than $\mathcal{D}_{t2}$. Now, we hope to build a registration problem to adapt the label of $\mathcal{D}_{t1}$ to $\mathcal{D}_{t2}$. A simple way to match a moving image X_j^{t1} and a fixed image X_i^{t2} is to select the minimum value of two sample's volume differences.

$$\min_{\mathbf{u}\in\mathcal{D}_{t1}} \|vol(X_i^{t2}) - vol(u)\|_1 \tag{5}$$

A common way to model the deformable registration problem is to consider the minimization of an energy functional (6) over the set of plausible deformations ϕ,

$$\mathcal{J}(\boldsymbol{\phi}; q_S, q_T) = \underbrace{\mathrm{D}(\boldsymbol{\phi}, q_S, q_T)}_{\text{Data term}} + \underbrace{\mathcal{R}(\boldsymbol{\phi})}_{\text{Regularizer}} \tag{6}$$

where D(.,.) is the data term that measures the discrepancy between the deformed shape $\phi.q_S$ and the target shape q_T. The term $\mathcal{R}(.)$ plays the role of a regularizer and thus controls the plausibility of the solution ϕ^*. In this work, the deformations ϕ is an affine transform, $\mathrm{D}(A, B) = \|A - B\|_2$, and $\mathcal{R}(.) = 0$. The solution ϕ^* can transfer the pseudo-label of q_S to q_T. The updated dataset still denotes $\mathcal{D}_t$.

2.3 Self-training

We use nnU-Net to train a segmentation model. As the trained nnU-Net may not perform well in all the images of $\mathcal{D}_t$ by using pseudo-labels, some unreliable pseudo labels may harm the training of small nnU-Net. We employ a simple method from [8] to filter the unreliable pseudo labels based on the stability during different training iterations. The method assumes that the generated pseudo labels should be stable during iterative training. If some pseudo labels vary greatly in different iterations, it indicates that the model is very uncertain about these pseudo labels. We should not use them for training. We calculate the uncertainty of pseudo labels using the following equation:

$$u = \frac{1}{K-1}\sum_{i=2}^{K} \frac{SUM(y_i \neq y_{i-1})}{SUM(y_i > 0)} \tag{7}$$

where u is the uncertainty and K is the total number of iterations, y_i is the pseudo label generated in iteration i. Further, we ensemble the pseudo-label of selected cases by doing a union operation.

2.4 Speeding inference

In order to improve inference speed and reduce resource consumption, we use a small-size network structure in reference [8]. And we change the default resampling function and order, which effectively speeds up the inference. The setup of

Table 1: Network architecture and inference process.

Channels in the first stage	16
Convolution number per stage	2
Patch size	40×224×192
Downsampling times	4
inference process	(Sigmoid, Threshold, Resample)
Deep supervision	True

network architecture and inference process are presented in Table 1. Comparison of different strategy settings in Table 2.

Table 2: Comparison of different strategy settings. The order of axes of input patch size and spacing is (z,y,x).

Settings	Default	Small
Channels in the first stage	32	16
Convolution number per stage	2	2
Patch size	40×224×192	40×224×192
Downsampling times	5	4
Input spacing	(2.5, 0.81, 0.81)	(2.5, 0.81, 0.81)

2.5 Data processing

Preprocessing For image prepossessing, all of our settings follow the default nnU-NetV2.

- Statistical analysis is conducted on data pertaining to volume spacing and foreground intensity.
- Images are clipped at the 0.5 and 99.5 percentiles of foreground voxels.
- All images are normalized through the subtraction of the mean and division by the standard deviation.
- The volume is then resampled to a target spacing of (2.5, 0.81, 0.81).

Post-processing We do not perform any post-processing, such as connected component analysis or testing time augmentation, during our inference.

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 [22], KiTS [6,7], autoPET [5,4], AMOS [10], LLD-MMRI [12], TotalSegmentator [24], and AbdomenCT-1K [19], and past FLARE Challenges [16,17,18]. The training set includes 2050 abdomen CT scans and over 4000 MRI scans. The validation and testing sets include 110 and 300 MRI scans, respectively, which cover various MRI sequences, such as T1, T2, DWI, and so on. The organ annotation process used ITK-SNAP [29], nnU-Net [9], MedSAM [14], and Slicer Plugins [3,15].

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

Environment settings The development environments and requirements are presented in Table 3.

Table 3: Development environments and requirements.

System	Ubuntu 20.04.5 LTS
CPU	Intel(R) Xeon(R) Gold 6354 CPU @ 3.00GHz
RAM	16×4GB; 1600MT/s
GPU (number and type)	1 × NVIDIA A100 40G
CUDA version	11.7
Programming language	Python 3.10.11
Deep learning framework	Pytorch 2.0.0, torchvision 0.2.2
Specific dependencies	nnU-Net 2.0

Table 4: Training protocols.

Network initialization	“He” normal initialization
Batch size	2
Patch size	40×224×192
Total epochs	1000
Optimizer	SGD with nesterov momentum (μ =0.99)
Initial learning rate (lr)	0.01
Lr decay schedule	Poly learning rate policy: $(1 - epoch/1000)^{0.9}$
Training time	10 hours
Loss function	Cross entropy loss and dice loss

Training protocols The training protocols of the small nnU-Net are listed in Table 4. For the training set, We select 300 cases from the unlabeled MRI images and 100 cases from labeled CT images. 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]. We used the pseudo labels generated by the FLARE22 winning algorithm [8] and the best-accuracy-algorithm [23] as labels of CT. We employ the same data augmentation as the default setting of nnU-Net, which includes additive brightness, gamma, rotation, scaling, and elastic deformation on the fly during training. During inference, the model does not perform test time augmentation (TTA) of flipping. The patch sampling strategy is foreground over-sampling. Finally, we choose the model that obtains the fast and best accuracy on the validation set.

4 Results and discussion

Table 5: Quantitative evaluation results in terms of DSC(%) and NSD(%).

Target	Validation	
	DSC(%)	NSD(%)
Liver	86.97 ± 6.480	78.74 ± 10.44
Right kidney	84.79 ± 11.42	83.32 ± 13.45
Spleen	82.62 ± 18.05	79.54 ± 20.46
Pancreas	44.28 ± 20.63	56.44 ± 24.83
Aorta	79.77 ± 12.10	83.17 ± 14.91
Inferior vena cava	57.49 ± 15.87	56.16 ± 15.19
Right adrenal gland	38.16 ± 17.54	53.63 ± 23.91
Left adrenal gland	47.13 ± 18.12	61.88 ± 21.67
Gallbladder	46.71 ± 29.53	35.55 ± 28.70
Esophagus	35.33 ± 18.51	46.90 ± 23.22
Stomach	60.08 ± 17.51	58.93 ± 18.67
Duodenum	38.68 ± 17.29	61.24 ± 21.90
Left kidney	84.89 ± 16.62	84.55 ± 17.00
Average	60.53 ± 16.90	64.62 ± 19.57

4.1 Quantitative results on validation set

In Table 5, we report the DSC and NSD of the final docker commit results. The average of the 110 public validation achieve a DSC of about 0.6053 and an NSD of 0.6462. In general, large organs such as the liver, spleen, kidney and stomach have high accuracy. However, the accurate identification of small and complex objects, such as the pancreas, the adrenal glands, and the duodenum, poses a huge challenge. It requires more attention, especially when dealing with extremely small and blurred boundaries.

In Table 7, we have done a careful ablation analysis of each module and can see that training the model using only CT data is the least effective. After using

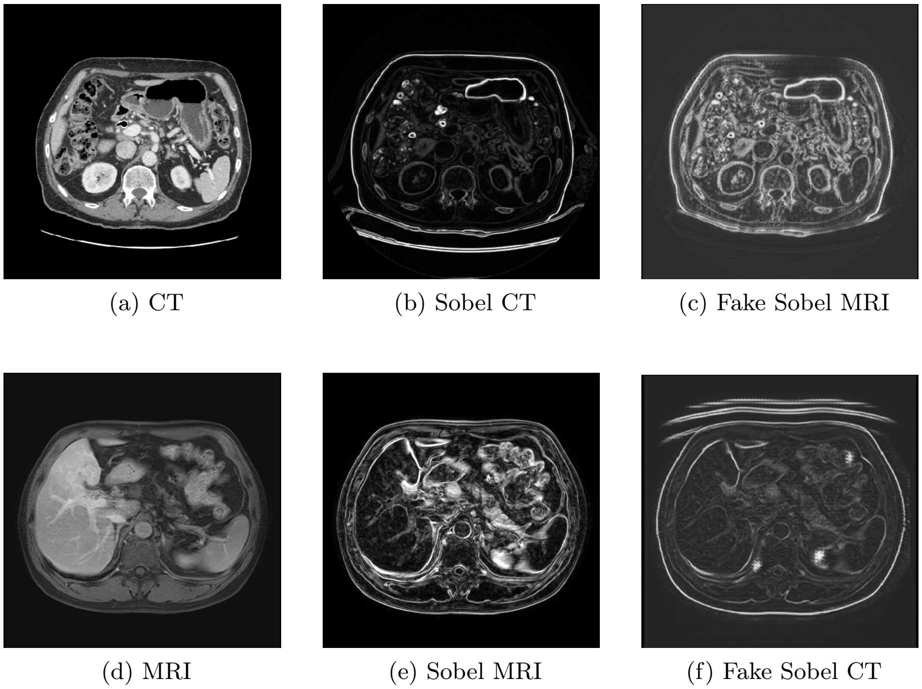
Table 6: Ablation studies of online validation quantitative evaluation results in terms of DSC(%) and NSD(%). DA is the results of using fake MRI generated by ASIT, and MDM represent to refine pseudo-labels by multi-domain matching.

Target	w/o DA		w/ DA		w/ DA + MDM	
	DSC(%)	NSD(%)	DSC(%)	NSD (%)	DSC(%)	NSD (%)
Liver	17.56	13.79	53.68	42.37	86.63	78.08
Right kidney	13.44	12.12	29.96	29.96	66.05	62.38
Spleen	5.990	4.290	46.40	39.70	79.41	73.01
Pancreas	11.13	11.39	12.13	17.02	38.64	48.73
Aorta	37.52	37.27	58.94	61.97	80.91	84.84
Inferior vena cava	16.05	15.43	28.28	26.99	54.29	54.00
Right adrenal gland	4.590	5.590	11.75	17.94	24.55	37.36
Left adrenal gland	10.26	11.39	10.91	16.56	24.45	33.98
Gallbladder	16.52	11.66	30.38	25.55	34.79	30.91
Esophagus	17.22	21.53	27.92	44.14	49.61	71.56
Stomach	13.21	8.810	34.61	35.92	59.55	60.62
Duodenum	15.83	20.06	10.45	18.34	23.78	40.18
Left kidney	19.85	17.66	28.46	28.65	65.52	61.75
Average	15.32	14.69	29.53	31.16	52.93	56.72

Table 7: Each module's quantitative evaluation results in terms of DSC(%) and NSD(%). ASIT is the result of using fake MRI for training on Sobel filter images, and MDM represent to refine pseudo-labels by multi-domain matching. PLF indicates that pseudo-label filtering was used.

Module	Validation	
	DSC(%)	NSD(%)
Only CT	15.32	14.69
CycleGAN	20.01	16.40
ASIT	29.53	31.16
ASIT + MDM	52.93	56.72
ASIT + MDM + PLF	60.53	64.62

the fake MR data generated by CycleGAN, the effect is improved. However, the generation in the anatomical space after the application of Sobel filtering gives better results. We report the validation results of the model without DA, with ASIT, and with ASIT + MDM in Table 6. The model using ASIT resulted in an increase of the DSC from 0.1532 to 0.2953. Additionally, MDM alone increased the DSC by approximately 0.23. Finally, pseudo-label filtering (PLF) further increased the DSC to 60.53. These results illustrate that it is difficult to achieve high performance for multiple target domains using only image translation.


(a) CT

(b) Sobel CT

(c) Fake Sobel MRI

(d) MRI

(e) Sobel MRI

(f) Fake Sobel CT

Fig. 2: 3D visualization of ASIT on CT and MRI scans.

4.2 Qualitative results on validation set

Figure 3 presents easy and difficult validation set examples for segmentation. Promising results were observed for Case #amos_8179 and Case #amos_0546, but the segmentation of Case #amos_7324 and Case#amos_0530 was poor. Some lesions, such as tumors, cause network segmentation errors. More importantly, the pseudo-labels generated by image translation are of poorer quality for small organs, which results in a much lower average segmentation performance.

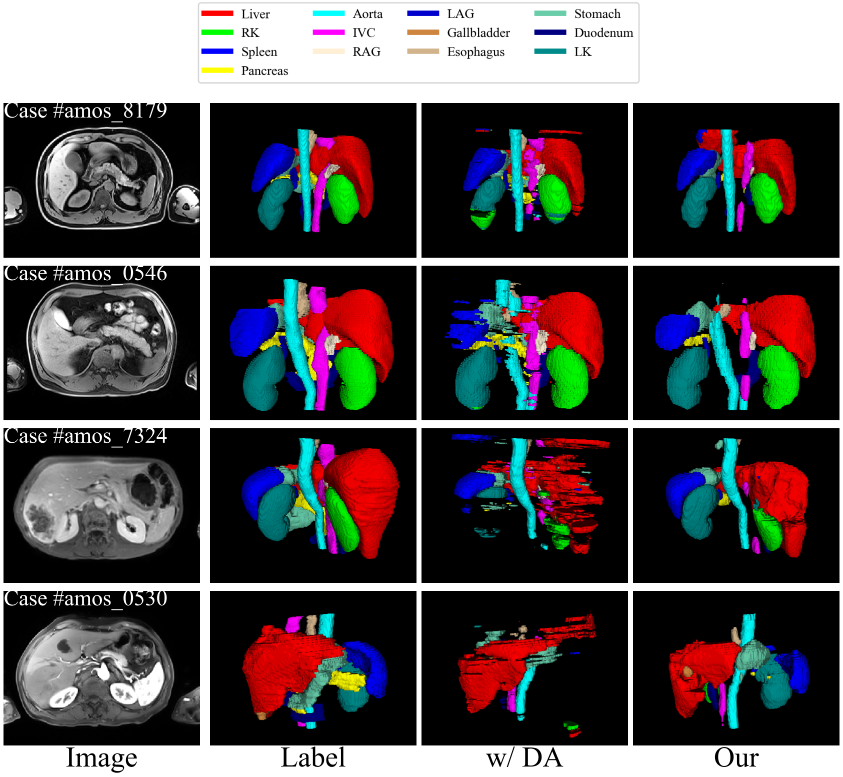


Fig. 3: 3D visualization on two easy cases (Case #amos_8179 with DSC of 0.64 and Case #amos_0546 with DSC of 0.60) and two hard cases (Case #amos_7324 with DSC of 0.45 and Case #amos_0530 with DSC of 0.48).

Table 8: Quantitative evaluation of segmentation efficiency in terms of the running them and GPU memory consumption on final testing set. Total GPU denotes the area under GPU Memory-Time curve. Evaluation GPU platform: NVIDIA QUADRO RTX5000 (16G).

Type of statistics	Running Time (s)	Total GPU (MB)
25th percentile value	18.3	967283.7
50th percentile value	19.5	1035840.0
75th percentile value	21.8	967283.7
Average value	20.7 ± 5.4	1107332.6 ± 316456.3

Figure 2 shows the picture generated using ASIT, which is effective in shrinking the inter-domain gap.

4.3 Segmentation efficiency results on final testing set

In Table 8, we report the results of the quantitative evaluation of the running efficiency on the final testing set. The average inference time and GPU occupied memory are 20.7 seconds and 1107332.6 MB.

4.4 Results on final testing set

Table 9: Quantitative evaluation results in terms of DSC(%) and NSD(%).

Target	Testing	
	DSC(%)	NSD(%)
25th percentile value	19.8	17.0
50th percentile value	40.0	36.0
75th percentile value	53.6	53.8
Average value	37.4 ± 19.5	36.3 ± 21.3

In Table 9, we report the DSC and NSD of the final docker commit results on testing set. The average of the final testing set achieve a DSC of about 37.4% and an NSD of 36.3%. The metric values on the testing set have a large decrease than on the validation set, which indicates that there is a gap between the data distribution of the testing and validation sets, and it is a challenging task to build a model with excellent generalizability.

4.5 Limitation and future work

This work have several limitations as follows. First, there are many ways to further improve the network inference process, such as a more efficient sliding window. Second, the challenge provided more than 3000 MRI cases, but we only utilized 300 cases and did not adequately utilize the data. Third, we perform image style translation to obtain pseudo-labeling in the decomposed space, but the quality is poor for organs with complex shapes and smaller sizes. How to effectively maintain the morphology of these organs deserves further exploration.

5 Conclusion

In this paper, we present a framework that combines unsupervised domain adaptation, registration and pseudo-labeling learning, which is effective and flexible for a variety of situations. In addition, we use a small nnU-Net and improve the inference process, effectively reducing its required computational resources and inference time. Because the amount of data used in training is small, performance on the full data will be explored in the future. The approach in this paper will be a good baseline result for exploring unsupervised domain adaptation.

Acknowledgements The authors of this paper declare that the segmentation method they implemented for participation in the FLARE 2024 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. 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) 7
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) 3, 5, 8
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) 7
10. 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) 6
11. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Unsupervised domain adaptation with residual transfer networks. In: NeurIPS. p. 136–144 (2016) 2
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) 8
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 7
15. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322 (2024) 7
16. 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
17. 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

18. 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
19. 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
20. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: ECCV. pp. 319–345 (2020) 2
21. Roth, H.R., Shen, C., Oda, H., Oda, M., et al.: Deep learning and its application to medical image segmentation. Medical Imaging Technology **36**(2), 63–71 (2018) 1
22. 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
23. 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, 8
24. 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
25. Wu, J., Guo, D., Wang, G., Yue, Q., Yu, H., Li, K., Zhang, S.: Fpl+: Filtered pseudo label-based unsupervised cross-modality adaptation for 3d medical image segmentation. IEEE Transactions on Medical Imaging pp. 1–1 (2024). https://doi.org/10.1109/TMI.2024.3387415 2
26. 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
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) 13
28. Yao, K., Su, Z., Huang, K., Coenen, F.: A novel 3d unsupervised domain adaptation framework for cross-modality medical image segmentation. IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS **26**(10) (2022) 2
29. 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) 7
30. 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, 4