Learning to Segment Multiple Organs from Multimodal Partially Labeled Datasets

Hong Liu[2,3], Dong Wei[3], Donghuan Lu[3], Jinghan Sun[2,3], Hao Zheng[3],
Yefeng Zheng[3](✉), and Liansheng Wang[1,2](✉)

[1] School of Informatics, Xiamen University, Xiamen, China
lswang@xmu.edu.cn
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China
{liuhong,jhsun}@stu.xmu.edu.cn
[3] Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
{donwei,caleblu,howzheng,yefengzheng}@tencent.com

Abstract. Learning to segment multiple organs from partially labeled datasets can significantly reduce the burden of manual annotations. However, due to the large domain gap, learning from *partially labeled datasets of different modalities* has not been well addressed. In addition, the anatomic prior knowledge of various organs is spread in multiple datasets and needs to be more effectively utilized. This work proposes a novel framework for learning to segment multiple organs from multimodal partially labeled datasets (*i.e.*, CT and MRI). Specifically, our framework constructs a cross-modal a priori atlas from training data, which implicitly contains prior knowledge of organ locations, shapes, and sizes. Based on the atlas, three novel modules are proposed to address the joint challenges of unlabeled organs and inter-modal domain gaps: 1) to better utilize unlabeled organs for training, we propose an atlas-guided pseudo-label refiner network to improve the quality of pseudo-labels; 2) we propose an atlas-conditioned modality alignment network for cross-modal alignment in the label space via adversarial training, forcing cross-modal segmentations of organs labeled in a different modality to match the atlas; and 3) to further align organ-specific semantics in the latent space, we introduce modal-invariant class prototype anchoring modules supervised by the refined pseudo-labels, encouraging domain-invariant features for each organ. Extensive experiments demonstrate the superior performance of our framework to existing state-of-the-art methods and the efficacy of its components.

Keywords: Multimodal partial label · Multi-organ segmentation · Probabilistic atlas.

1 Introduction

Accurate segmentation of various organs in medical images is valuable to clinical applications. In clinical practice, multiple image modalities, *e.g.*, computed

H. Liu and D. Wei——Contributed equally; H. Liu contributed to this work during an internship at Tencent.

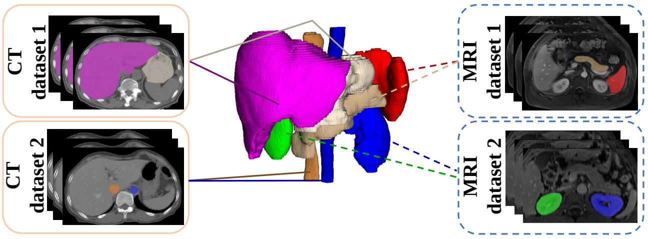

Fig. 1. Illustration of *multimodal* partially labeled multi-organ segmentation. This task aims to segment multiple organs in multiple modalities using a network trained on several partially labeled datasets, each providing segmentations of one or a few particular organs. While previous work explored partially labeled monomodal CT datasets (the solid lines), this work aims to fulfill the more challenging multimodal task to include MRI datasets (the dashed lines), another standard imaging modality in clinic routine.

tomography (CT) and magnetic resonance imaging (MRI), are commonly used together to provide complementary perspectives [17, 18, 23]. Most research focused on designing optimal, modality-specific segmentation models for either of them [2, 11]. However, these models often would yield suboptimal performance when applied to the modality not originally used for training due to the large domain gap, even though the content imaged by different modalities is the same. Some work [8, 33] proposed to address the problem by training a single model with joint representation learning/alignment in a unified latent space from multimodal datasets and achieved improved performance across modalities. However, these methods required all organs of interest to be annotated in every modality, which can be difficult and costly considering the expertise and labor needed.

Recently, researchers started to investigate using partially labeled datasets (*i.e.*, only a subset of all organs of interest is annotated in a specific image) to train a model that could segment all organs of interest together. This can significantly reduce the burden of manual annotation. Some studies [4, 6, 10, 27, 38] proposed adaptive and conditional loss functions and networks specifically designed for partial-label training. Nevertheless, they ignored unlabeled organs and treated them as background. Other works relied on pseudo-labels to train on unlabeled organs [14, 26]. Zhou *et al.* [40] proposed a prior-aware neural network (PaNN) that explicitly incorporated the anatomical prior of organ sizes. However, these methods only considered the settings where monomodal datasets were partially labeled for different tasks (*i.e.*, monomodal multitask setting), and their extension and efficacy in multimodal settings still needed to be studied. Unsupervised domain adaptation (UDA) can narrow the gaps between different domains [3, 9, 15, 29], often involving sophisticated feature disentanglement and domain-invariant feature extraction [34, 35, 37]. The performance may drop when the domain discrepancy is too large, especially for the cross-modal setting.

This paper proposes a novel framework for learning to segment multiple organs from multimodal partially labeled datasets (Fig. 1). Above all, we construct a cross-modal a priori probabilistic atlas [7, 13, 24] from training data, which implicitly contains rich prior knowledge about the organs, such as location, shape, and size. The atlas is exploited by three novel modules to address the critical challenges due to unlabeled organs and inter-modal domain gaps. First, to improve the quality of pseudo-labels for unlabeled organs in multimodal settings, we introduce an atlas-guided pseudo-label refiner network (APRN). The refined pseudo-labels are used to supervise the training of the main segmentation network so that the latter can better utilize valuable information from unlabeled organs and mistreat them less as background. Second, we propose an atlas-conditioned modality alignment network (AMAN) for cross-modal alignment via adversarial training, where a discriminator judges whether a segmentation is for an organ labeled in the current modality while the main segmentation network is forced to produce cross-modal segmentations able to fool the discriminator, both conditioned on the atlas. In this way, the cross-modal segmentation is aligned with the atlas in the label space. Third, to further align organ-specific semantics in the latent space, we introduce several modal-invariant class prototype anchoring modules (MICPAMs) into the decoder of the main segmentation network. Supervised by the APRN-refined pseudo labels, MICPAMs anchor the features of unlabeled organs in an image to the modal-invariant prototypes extracted from images in which these organs are labeled. Guided by the atlas, our AMAN and MICPAM modules are expected to align different modalities of significant domain gaps better than UDA. Extensive experiments demonstrate (1) our framework's superiority to existing state-of-the-art methods and (2) the efficacy of its novel modules.

2 Method

Problem Setting. Consider a set of M unpaired, partially labeled datasets $\{\mathcal{D}^{(i)}\}_{i=1}^{M}$ of different modalities (*e.g.*, CT and MRI) and segmentation tasks (*i.e.*, segmentation targets may vary with datasets). Further, each $\mathcal{D}^{(i)} = \{(x^{(i,j)}, y^{(i,j)})\}_{j=1}^{N_i}$, where N_i is the number of images in $\mathcal{D}^{(i)}$, $x^{(i,j)} \in \mathbb{R}^{D\times H\times W}$ is the j^{th} image in $\mathcal{D}^{(i)}$, D, H and W are the depth, height, and width of the image, respectively, $y^{(i,j)} \in \{0,1\}^{|\mathcal{C}^{(i)}|\times D\times H\times W}$ is the binary pixel-wise label of $x^{(i,j)}$, $\mathcal{C}^{(i)}$ is the set of labeled classes in $\mathcal{D}^{(i)}$ (different organs), and $|\mathcal{C}^{(i)}|$ is the number of classes in $\mathcal{C}^{(i)}$. Following existing literature [6, 10, 38], $\mathcal{C}^{(i)} \cap \mathcal{C}^{(j)} = \emptyset$ for all $i \neq j$, which is the most challenging setting for partially labeled scenarios, and the union set $\mathbb{C} = \{\mathcal{C}^{(i)}\}_{i=1}^{M}$ comprises all organs of interest to segment. The goal is to learn a single model from $\{\mathcal{D}^{(i)}\}$ to segment all classes in $\{\mathcal{C}^{(i)}\}$ for any modality in $\{\mathcal{D}^{(i)}\}$. Below, the superscripts i, j will be ignored without confusion.
Overview. Fig. 2 overviews our framework. As a premise, an a priori probabilistic atlas α for all organs of interest is constructed from training data. In each mini-batch, images are randomly sampled from $\{\mathcal{D}\}$ and fed into the main segmentation network, yielding a prediction p for all organs. Then, the APRN refines the prediction for *unlabeled organs* in each image (denoted by $p_{\mathbb{C}\setminus\mathcal{C}}$) guided by the same organs in the atlas (denoted by $\alpha_{\mathbb{C}\setminus\mathcal{C}}$), yielding refined pseudo-labels $\hat{y}$ to supervise the main segmentation network via $\mathcal{L}_{pseudo}$. Meanwhile,

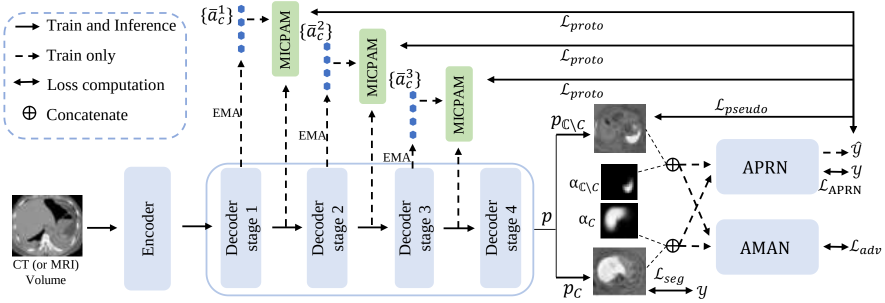

Fig. 2. Overview of our framework. A CT image, in which the liver is labeled while the spleen is not, is used for illustration.

the AMAN encourages the cross-modal segmentation of unlabeled organs $p_{\mathbb{C}\setminus\mathcal{C}}$ to harmonize with the atlas $\alpha_{\mathbb{C}\setminus\mathcal{C}}$ in the label space via $\mathcal{L}_{adv}$. Last but not least, several MICPAMs further align the features of unlabeled organs with a set of modal-invariant anchors $\{\bar{a}_c^1\}$–$\{\bar{a}_c^3\}$ in the latent space via $\mathcal{L}_{proto}$.

Cross-modal a Priori Atlas Construction. As the data used in this study primarily concern abdominal organs with roughly consistent fields of view, we implement a straightforward pipeline for volumetric image alignment and atlas construction from training data. First, we re-slice each image to the same resolution of 2×1×1 mm^3 and use Otsu thresholding [39] to identify and crop out the foreground sub-volumes (*i.e.*, abdominal torso) as preprocessing. Then, we resize the foreground sub-volumes to a uniform size of 143×233×338 voxels, the mean size of all training images after preprocessing. These steps roughly align all the images in effect and are applied to the labels, too. Next, we average all training labels of a specific class to obtain a class-wise a priori probabilistic atlas. Finally, the class-wise atlases for all classes of interest are concatenated to compose a cross-modal atlas $\alpha \in \mathbb{R}^{|\mathbb{C}|\times 143\times 233\times 338}$. Example atlases for the liver and spleen (denoted by $\alpha_{\mathcal{C}}$ and $\alpha_{\mathbb{C}\setminus\mathcal{C}}$, respectively) can be seen in Fig. 2.

Atlas-guided Pseudo-label Refiner Network (APRN). To utilize the unlabeled organs of partially labeled datasets rather than misleadingly treating them as background, a common solution is to adopt semi-supervised learning with model-generated pseudo-labels for training [28]. However, due to the large discrepancy between different modalities, the pseudo-labels may become highly unreliable for data from another modality. Therefore, we propose the APRN (Fig. 2), a lightweight U-shape network trained on labeled organs, to refine the pseudo-labels for unlabeled organs guided by the atlas. Denote the after-softmax probabilities predicted by the main segmentation network for an image by $p \in \mathbb{R}^{|\mathbb{C}|\times D\times H\times W}$. We slice the prediction and the atlas to extract the sub-tensors corresponding to the labeled organs (*e.g.*, the liver labeled in the CT image in Fig. 2), denoted by $p_{\mathcal{C}}$ and $\alpha_{\mathcal{C}}$, respectively, where $p_{\mathcal{C}}, \alpha_{\mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|\times D\times H\times W}$. Then, $p_{\mathcal{C}}$ and $\alpha_{\mathcal{C}}$ are concatenated and input to APRN to generate refined segmentations for the labeled organs, which are compared to the ground truth labels

y to train APRN with a loss $\mathcal{L}_{APRN}$.[4] Thus, APRN learns to refine the main segmentation network's prediction guided by the atlas. To apply it to pseudo-label refinement of unlabeled organs (*e.g.*, the spleen unlabeled in the CT image in Fig. 2), we slice to obtain $p_{\mathbb{C}\backslash\mathcal{C}}$ and $\alpha_{\mathbb{C}\backslash\mathcal{C}}$, where $\mathbb{C}\backslash\mathcal{C}$ indicates the difference set between $\mathbb{C}$ and $\mathcal{C}$, *i.e.*, the set of classes unlabeled in a specific image. Next, $p_{\mathbb{C}\backslash\mathcal{C}}$ and $\alpha_{\mathbb{C}\backslash\mathcal{C}}$ are concatenated and input to APRN to produce atlas-refined pseudo-labels $\hat{y}$, which are used to supervise the main segmentation network on the unlabeled organs via a loss $\mathcal{L}_{pseudo}$. To reduce the computational cost, APRN operates on a downsampled scale of 1/2 of the original input image.

Atlas-conditioned Modality Alignment Network (AMAN). Although an organ may present huge discrepancies in appearance (*e.g.*, intensity, contrast, and texture) in different modalities, its location, shape, and size should persist statistically. Nevertheless, it is often difficult to explicitly describe and incorporate such property consistency into the training of deep neural networks. Instead, we propose to generalize the consistency constraint of each organ via adversarial training [22] conditioned on the a priori probabilistic atlas.

The AMAN (Fig. 2) takes in the segmentation predicted by the main segmentation network paired (concatenated) with the atlas for a specific organ and determines whether the prediction is for an organ labeled in the current modality (1 for true and 0 for false):

$$\mathcal{L}_D(p|\alpha) = -\sum\nolimits_{c\in\mathbb{C}} \mathbb{1}(c)\log f_D(p_c|\alpha_c) + \big(1-\mathbb{1}(c)\big)\log\big(1-f_D(p_c|\alpha_c)\big), \qquad (1)$$

where f_D is the discriminator network, $c\in\mathbb{C}$ is an organ class, $p_c, \alpha_c \in \mathbb{R}^{1\times D\times H\times W}$ are the corresponding segmentation and atlas for the specific organ, respectively, and $\mathbb{1}(c)$ is an indicator function which equals 1 if the organ c is labeled in the current image's modality and 0 otherwise. Given the domain gap between modalities, the intramodal segmentation is expected to be better in quality and thus closer to the atlas than cross-modal segmentation. To confuse the discriminator, the main segmentation network is trained to produce cross-modal segmentations that better match the atlas with an adversarial loss: $\mathcal{L}_{adv}(p|\alpha) = -\sum_{c\in\mathbb{C}}\big(1-\mathbb{1}(c)\big)\log\big(f_D(p_c|\alpha_c)\big)$.

Modality-invariant Class Prototype Anchoring Module (MICPAM). The AMAN described above aligns different modalities in the *label space*. Below, we further improve the alignment using the MICPAMs in the *latent space*.

Class Prototype Extraction: During training, we maintain a set of modal-invariant class prototypes $\mathbb{A}^{(s)} = \{\bar{a}_c^{(s)}\}_{c\in\mathbb{C}}$ for each feature scale s of the main segmentation network's decoder. Without loss of generality, we describe the prototype extraction and anchoring mechanisms with a generic scale and denote the corresponding feature tensor in this scale by $\mathbf{F}\in\mathbb{R}^{n\times d\times h\times w}$. Then, the prototype for class c can be computed by masked average pooling:

$$a_c = \sum\nolimits_{(z,y,x)} m_{c,(z,y,x)}\mathbf{F}_{:,z,y,x} \Big/ \sum\nolimits_{(z,y,x)} m_{c,(z,y,x)}, \qquad (2)$$

[4] The gradients in APRN are not backpropagated to the main segmentation network.

where (z, y, x) enumerates all coordinates in a $d \times h \times w$ volume, $\mathbf{F}_{:,z,y,x}$ is the n-dimension feature vector at (z, y, x), and m_c is a binary mask indicating belongingness to class c (1 for true, 0 for false). To ensure the quality of the representations, we only update the prototypes for labeled classes in each mini-batch. Further, m_c is set to the consensus regions between the ground truth label y and the main segmentation network's predicted mask. The rationale is that only features in correctly predicted regions can represent the organs well. To prevent the prototypes from being radically affected by potential outliers and stabilize the training process, the exponential moving average (EMA) [28] is employed to update each anchor progressively and smoothly: $\bar{a}_c = \epsilon \times \bar{a}_c + (1 - \epsilon) \times a_c$, where ϵ is set to 0.99 and $\bar{a}_c$ is initialized randomly.

Anchoring to Prototypes by Deep Supervision: Next, we compute the cross-attention between the decoder features and the prototypes via dot product [30]. For batch processing, the prototypes of all classes comprise a prototype matrix $A \in \mathbb{R}^{|\mathbb{C}| \times n}$ where each row is a class's prototype, and the feature tensor $\mathbf{F}$ is reshaped to a matrix $F \in \mathbb{R}^{(dhw) \times n}$. Then, the attention is computed by

$$\text{Attn} = \text{softmax}\left(AW^A(FW^F)^T\right), \tag{3}$$

where W^A and W^F are linear projection matrices. Then, we reshape Attn to $|\mathbb{C}| \times d \times h \times w$, which is the per-pixel class-wise semantic-aware map from the features to all class prototypes. From a different perspective, the attention map can be considered a distance-metric-based segmentation concerning the distances from the features to the prototypes [5]. Therefore, we impose a segmentation loss $\mathcal{L}_{proto}$ on the unlabeled classes in Attn supervised by the atlas-refined pseudo-label $\hat{y}$ (resizing applied as needed). Not only does $\mathcal{L}_{proto}$ shape the modal-invariant class prototypes, but it also helps the main segmentation network align features of different modalities by anchoring them to the prototypes.

As shown in Fig. 2, the MICPAM is inserted at every intermediate feature scale of our main segmentation network's decoder for deep supervision [20].

Objective Function, Training, and Inference. The objective function of the main segmentation network is

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_{pseudo} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{proto}, \tag{4}$$

where $\mathcal{L}_{seg}$ is the supervised segmentation loss on labeled organs, and λ_1, λ_2 and λ_3 are weights. For $\mathcal{L}_{seg}$, the widely used Dice loss [21] plus cross entropy loss are used. For $\mathcal{L}_{pseudo}$, $\mathcal{L}_{proto}$, and $\mathcal{L}_{APRN}$, the Dice loss is used.

In each training mini-batch, APRN is firstly updated to optimize $\mathcal{L}_{APRN}$, then the main segmentation network and MICPAMs are updated to optimize Eqn. (4), and lastly, AMAN is updated to optimize Eqn. (1). The detailed algorithm is provided in the supplementary material. Note that the APRN, AMAN, and MICPAMs are only used during training. For inference, we directly take the main segmentation network's prediction as segmentation results, which is as efficient as a vanilla encoder-decoder architecture in terms of computation cost.

3 Experiments

Dataset and Experimental Settings. The AMOS22 dataset [16] contains 200 abdominal CT and 40 abdominal MRI scans for training. Voxel-wise annotations of 13 organs are provided in both CT and MRI. We randomly split the 200 CT (40 MRI) images into 162 (30) and 38 (10) subjects for training and testing, respectively. Further, the training CT data is randomly split into two partially labeled datasets of equal sizes, each for a non-overlapping four-organ segmentation task: liver, stomach, aorta, and esophagus are labeled in the first, whereas inferior vena cava, right adrenal gland, left adrenal gland, and duodenum are labeled in the second. Meanwhile, the remaining five organs are labeled in the MRI data: spleen, right kidney, left kidney, gallbladder, and pancreas. Therefore, we formulate a three-task (*i.e.*, $M = 3$) multimodal partially labeled segmentation problem. The Dice similarity coefficient (DSC) in percentage (%) is used for performance evaluation, and the Wilcoxon signed rank test is employed for analysis of statistical significance.

Implementation. The PyTorch framework (1.7.1) [25] is used for experiments. We use the same main segmentation network as Zhang *et al.* [38], essentially a 3D U-Net comprising a single encoder and a single decoder employing residual blocks [12] and group normalization [32]. The refiner (APRN) is a similar but smaller segmentation network, whereas the discriminator (AMAN) is a 3D classification network. Three Tesla V100 GPUs are used for training, with a batch size of three volumes. The Adam optimizer [19] is employed with an initial learning rate of 0.0005 and decayed according to a polynomial policy $lr = lr_{init} \times (1 - k/K)^{0.9}$ for $K = 600$ epochs. To match an input image and the a priori atlas, we apply the same preprocessing steps to the image as in the atlas construction and resize the atlas to the size of the preprocessed image. To standardize all volumes, CT images are normalized by clipping to $[-325, 325]$ Hounsfield units followed by linear scaling to the range of $[-1,1]$, whereas MRI images are normalized by subtracting the volume mean and dividing by the standard deviation. During training, we use random sub-volumes of 64×192×192 voxels as input. No other data augmentation is implemented, as by Zhang *et al.* [38]. We split ∼12.5% of training data for validation, *i.e.*, selecting the optimal model for testing. The loss weights in Eqn. (4) are empirically set to 1, 0.01, and 0.1 for λ_1, λ_2, and λ_3, respectively. The source code is available at: https://github.com/ccarliu/multimodal-PL.

Performance Comparison with State-of-the-art (SOTA). We compare our framework with nine SOTA methods: MH-Net [4], Cond-Dec [6], DoDNet [38], PaNN [40], U^{2}PL [31], VAT [22], PCL [1], DAR-UNet [35], and UniSeg [36]. We also include single-task models for reference. Table 1 shows the average intra- and cross-modal performance (*i.e.*, an organ is labeled and evaluated in the same modality or different ones) of all compared methods. We make the following observations. First, while the intramodal performance of the three monomodal partial-label methods (MH-Net, Cond-Dec, and DoDNet) is generally competent (especially DoDNet, which is *highly competitive*), their cross-modal performance is poor. Second, DAR-UNet, PaNN, U^{2}PL, VAT, PCL, and UniSeg yield tremendously improved cross-modal performance over the previous group of methods,

Table 1. Multimodal (CT and MRI) partially labeled abdominal multi-organ segmentation performance comparison with SOTA methods in DSC (mean±std%). The more challenging cross-modal evaluation (*i.e.*, an organ is labeled and evaluated in different modalities) is highlighted with dark shading. Detailed organ-wise results are provided in the supplementary material. $*$: $p < 0.05$ for pairwise comparison with our method.

Label	*Test modality: CT*									
modality	Single	MH-Net [4]	Cond-Dec [6]	DoDNet [38]	DAR-UNet [35]	PaNN [40]	U^{2}PL [31]	PCL [1]	UniSeg [36]	Ours
CT	78.8±17.5*	76.9±18.1*	79.0±19.2*	79.8±18.5	76.6±18.7*	74.7±19.1*	78.6±18.1*	79.6±17.2*	79.4±17.4*	**80.6**±16.7
MRI	25.2±13.3*	3.3±8.4*	15.2±23.8*	0.0±0.0*	70.3±23.0*	60.1±25.3*	67.1±20.1*	72.9±19.4*	73.0±25.3*	**81.4**±14.9
	Test modality: MRI									
CT	36.2±24.7*	12.4±18.1*	12.0±17.5*	2.4±8.0*	53.0±23.4*	51.7±26.3*	57.7±24.8*	57.1±24.5*	57.5±19.3*	**67.8**±21.5
MRI	79.4±23.3*	76.9±25.8*	81.2±20.4*	81.6±24.0*	76.0±24.3*	78.5±20.5*	79.3±21.0*	80.8±21.0*	83.1±21.7*	**86.1**±17.2

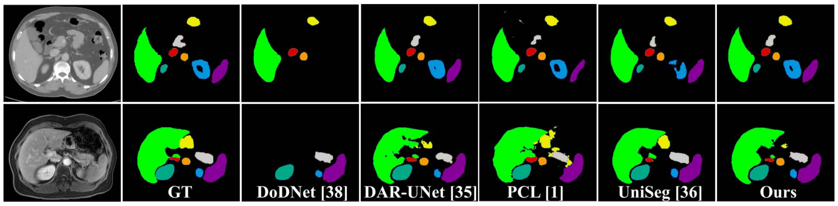

Fig. 3. Example segmentation results (top: CT; bottom: MRI) by our and several representative comparison methods. Organs labeled in CT: liver (🟩), stomach (🟨), inferior vena cava (🟥), and aorta (🟧); and organs labeled in MRI: spleen (🟪), left kidney (🟦), right kidney (■), and pancreas (⬜). Best viewed in color.

by margins of ∼40–70% in average cross-modal DSC. Last, our method performs best for all four intra- and cross-modal average DSCs. For intramodal, it outperforms the second-best method (DoDNet and UniSeg) by 0.8% on CT and 3.0% on MRI. For cross-modal, it outperforms the second-best methods (UniSeg and U^{2}PL) by 8.4% (MRI→CT) and 10.1% (CT→MRI), respectively. These results demonstrate the strong capability of our method in multimodal partially labeled multi-organ segmentation. Fig. 3 shows example segmentation results.

Table 2. Ablation study on the efficacy of our framework's three novel modules in DSC (mean±std%). $*$: $p < 0.05$ for pairwise comparison with the full model.

Ablation	(a)	(b)	(c)	(d)	(e)	(f)	(g)	Full
AMAN	×	✓	×	×	✓	×	✓	✓
APRN	×	×	✓	×	✓	✓	×	✓
MICPAM	×	×	×	✓	×	✓	✓	✓
Intramodal	80.1±18.8*	80.8±17.4*	82.6±17.3*	82.1±17.6	82.1±18.2*	82.8±17.0	83.3±16.5	**83.4**±16.9
Cross-modal	63.6±26.3*	67.2±20.3*	68.8±19.5*	69.6±24.8*	70.4±20.6*	69.5±24.9*	69.1±20.7*	**74.6**±18.2

Ablation Study. We conduct ablative experiments to validate the efficacy of our framework's novel components (Table 2). The baseline (a) is the main segmentation network trained only on labeled organs in each image. (b), (c), and (d) add AMAN, APRN, and MICPAMs to the baseline, respectively, and improve the performance by 0.7–2.5% in intramodal and 3.6–6.0% in cross-modal evaluation. (e)–(g) evaluate pairwise combinations of the modules, where at least either intra- or cross-modal performance is improved upon adding a single module. Lastly, our full model integrating all three modules achieves the best intra- and cross-modal performances—especially for the latter, which is 11% higher than the baseline. These results suggest that not only are the three modules effective individually, but they are also compatible, boosting each other together.

4 Conclusion and Future Work

This paper presented a novel probabilistic-atlas-guided framework for learning to segment multiple organs from multimodal partially labeled datasets. Extensive experiments demonstrated its superiority to existing SOTA approaches and the efficacy of its novel components. In the future, it would be useful to extend the framework for lesion segmentation, too. In addition, more advanced registration methods can be employed for volume alignment and atlas construction. Lastly, it would be interesting to explore partially labeled datasets with overlapping organ annotations, which may be exploited to bridge the domain gaps.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant No. 62371409).

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

References

1. Alonso, I., Sabater, A., Ferstl, D., Montesano, L., Murillo, A.C.: Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: ICCV. pp. 8219–8228 (2021)
2. Cao, H., Wang, Y., Chen, J., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV. pp. 205–218. Springer (2022)
3. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: AAAI. vol. 33, pp. 865–872 (2019)
4. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
5. Cui, H., Wei, D., Ma, K., Gu, S., Zheng, Y.: A unified framework for generalized low-shot medical image segmentation with scarce data. IEEE TMI **40**(10), 2656–2671 (2020)
6. Dmitriev, K., Kaufman, A.E.: Learning multi-class segmentations from single-class datasets. In: CVPR. pp. 9501–9511 (2019)
7. Dong, C., Chen, Y.w., Foruzan, A.H., et al.: Segmentation of liver and spleen based on computational anatomy models. Computers in Biology and Medicine **67**, 146–160 (2015)
8. Dou, Q., Liu, Q., Heng, P.A., Glocker, B.: Unpaired multi-modal segmentation via knowledge distillation. IEEE TMI **39**(7), 2415–2425 (2020)

9. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A.: Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss. arXiv preprint arXiv:1804.10916 (2018)
10. Fang, X., Yan, P.: Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE TMI **39**(11), 3619–3629 (2020)
11. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: MICCAI. pp. 61–71. Springer (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Huang, H., Zheng, H., Lin, L., et al.: Medical image segmentation with deep atlas prior. IEEE TMI **40**(12), 3519–3530 (2021)
14. Huang, R., Zheng, Y., Hu, Z., Zhang, S., Li, H.: Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In: MICCAI. pp. 146–155. Springer (2020)
15. Huo, Y., Xu, Z., Moon, H., et al.: SynSeg-Net: Synthetic segmentation without target modality ground truth. IEEE TMI **38**(4), 1016–1025 (2018)
16. Ji, Y., Bai, H., Yang, J., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)
17. Karim, R., Blake, L.E., Inoue, J., et al.: Algorithms for left atrial wall segmentation and thickness–evaluation on an open-source CT and MRI image database. Med. Image Anal. **50**, 36–53 (2018)
18. Kavur, A.E., Gezer, N.S., Barış, M., et al.: Chaos challenge-combined (CT-MR) healthy abdominal organ segmentation. Med. Image Anal. **69**, 101950 (2021)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial Intelligence and Statistics. pp. 562–570. Pmlr (2015)
21. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: International Conference on 3D Vision. pp. 565–571. IEEE (2016)
22. Miyato, T., Maeda, S.i., Koyama, M., Ishii, S.: Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE TPAMI **41**(8), 1979–1993 (2018)
23. Nikolaou, K., Alkadhi, H., Bamberg, F., Leschka, S., Wintersperger, B.J.: MRI and CT in the diagnosis of coronary artery disease: indications and applications. Insights into Imaging **2**(1), 9–24 (2011)
24. Park, H., Bland, P.H., Meyer, C.R.: Construction of an abdominal probabilistic atlas and its application in segmentation. IEEE TMI **22**(4), 483–492 (2003)
25. Paszke, A., Gross, S., Massa, F., et al.: PyTorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
26. Petit, O., Thome, N., Soler, L.: Iterative confidence relabeling with deep ConvNets for organ segmentation with partial labels. Computerized Medical Imaging and Graphics (2021)
27. Shi, G., Xiao, L., Chen, Y., Zhou, S.K.: Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. **70**, 101979 (2021)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NeurIPS **30** (2017)
29. Tsai, Y.H., Hung, W.C., Schulter, S., et al.: Learning to adapt structured output space for semantic segmentation. In: CVPR. pp. 7472–7481 (2018)

30. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. NeurIPS **30** (2017)
31. Wang, Y., Wang, H., Shen, Y., et al.: Semi-supervised semantic segmentation using unreliable pseudo-labels. In: CVPR. pp. 4248–4257 (2022)
32. Wu, Y., He, K.: Group normalization. In: ECCV. pp. 3–19 (2018)
33. Yang, J., Zhu, Y., Wang, C., Li, Z., Zhang, R.: Toward unpaired multi-modal medical image segmentation via learning structured semantic consistency. In: MIDL (2023)
34. Yang, J., Dvornek, N.C., Zhang, F., et al.: Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In: MICCAI. pp. 255–263. Springer (2019)
35. Yao, K., Su, Z., Huang, K., et al.: A novel 3D unsupervised domain adaptation framework for cross-modality medical image segmentation. IEEE JBHI **26**(10), 4976–4986 (2022)
36. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y.: Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. In: MICCAI. pp. 508–518. Springer Nature Switzerland, Cham (2023)
37. Zeng, G., Lerch, T.D., Schmaranzer, F., et al.: Semantic consistent unsupervised domain adaptation for cross-modality medical image segmentation. In: MICCAI. pp. 201–210. Springer (2021)
38. Zhang, J., Xie, Y., Xia, Y., Shen, C.: DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In: CVPR. pp. 1195–1204 (2021)
39. Zhang, J., Hu, J.: Image segmentation based on 2D Otsu method with histogram analysis. In: International Conference on Computer Science and Software Engineering. vol. 6, pp. 105–108. IEEE (2008)
40. Zhou, Y., Li, Z., Bai, S., et al.: Prior-aware neural network for partially-supervised multi-organ segmentation. In: ICCV. pp. 10672–10681 (2019)