MagicNet: Semi-Supervised Multi-Organ Segmentation via Magic-Cube Partition and Recovery

Duowen Chen[1] Yunhao Bai[1] Wei Shen[2] Qingli Li[1] Lequan Yu[3] Yan Wang[1]*
[1]Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University [3]The University of Hong Kong
duowen_chen@hotmail.com, yhbai@stu.ecnu.edu.cn, wei.shen@sjtu.edu.cn,
qlli@cs.ecnu.edu.cn, lqyu@hku.hk, ywang@cee.ecnu.edu.cn

Abstract

We propose a novel teacher-student model for semi-supervised multi-organ segmentation. In teacher-student model, data augmentation is usually adopted on unlabeled data to regularize the consistent training between teacher and student. We start from a key perspective that ***fixed relative locations*** *and* ***variable sizes*** *of different organs can provide distribution information where a multi-organ CT scan is drawn. Thus, we treat the prior anatomy as a strong tool to guide the data augmentation and reduce the mismatch between labeled and unlabeled images for semi-supervised learning. More specifically, we propose a data augmentation strategy based on partition-and-recovery N^3 cubes cross- and within- labeled and unlabeled images. Our strategy encourages unlabeled images to learn organ semantics in relative locations from the labeled images (cross-branch) and enhances the learning ability for small organs (within-branch). For within-branch, we further propose to refine the quality of pseudo labels by blending the learned representations from small cubes to incorporate local attributes. Our method is termed as MagicNet, since it treats the CT volume as a magic-cube and N^3-cube partition-and-recovery process matches with the rule of playing a magic-cube. Extensive experiments on two public CT multi-organ datasets demonstrate the effectiveness of MagicNet, and noticeably outperforms state-of-the-art semi-supervised medical image segmentation approaches, with +7% DSC improvement on MACT dataset with 10% labeled images. Code is avaiable at https://github.com/DeepMed-Lab-ECNU/MagicNet.*

1. Introduction

Abdominal multi-organ segmentation in CT images is an essential task in many clinical applications such as computer-aided intervention [25, 33]. But, training an accurate multi-organ segmentation model usually requires a large amount of labeled data, whose acquisition process is time-consuming and expensive. Semi-supervised learning (SSL) has shown great potential to handle the scarcity of data annotations, which attempts to transfer mass prior knowledge learned from the labeled to unlabeled images. SSL attracts more and more attention in the field of medical image analysis in recent years.

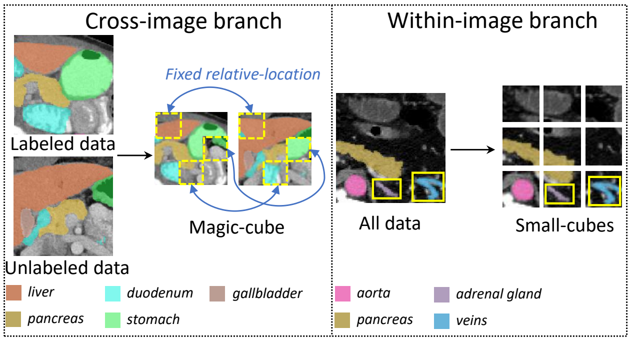

Figure 1. Two data augmentation strategies in MagicNet. **Left**: Although labeled and unlabeled images are not aligned, the latter can be regarded as a shifted version of the former. Co-shift of cubes transfers organ semantics in relative locations from the labeled data to unlabeled data. **Right**: Segmenting small organs from original images are difficult due to the cluttered background. Small cubes mitigate the impact from background and focus more on local attributes.

Popular SSL medical image segmentation methods mainly focus on segmenting a single target or targets in a local region, such as segmenting pancreas or left atrium [4, 9, 14, 15, 18, 23, 31, 35, 38, 39]. Multi-organ segmentation is more challenging than single organ segmentation, due to the complex anatomical structures of the organs, *e.g.*, the fixed relative locations (duodenum is always located at the head of the pancreas), the appearances of different organs, and the large variations of the size. Transfer-

[*] Corresponding Author.

ring current SSL medical segmentation methods to multi-organ segmentation encounters severe problems. Multiple organs introduce much more variance compared with a single organ. Although labeled and unlabeled images are always drawn from the same distribution, due to the limited number of labeled images, it's hard to estimate the precise distribution from them [32]. Thus, the estimated **distribution** between labeled and unlabeled images always suffers from **mismatch problems**, and is even largely increased by multiple organs. Aforementioned SSL medical segmentation methods lack the ability to handle such a large distribution gap, which requires sophisticated anatomical structure modeling. A few semi-supervised multi-organ segmentation methods have been proposed, DMPCT [43] designs a co-training strategy to mine consensus information from multiple views of a CT scan. UMCT [36] further proposes an uncertainty estimation of each view to improve the quality of the pseudo-label. Though these methods take the advantages of multi-view properties in a CT scan, they inevitably ignore the internal anatomical structures of multiple organs, resulting in suboptimal results.

Teacher-student model is a widely adopted framework for semi-supervised medical image segmentation [28]. Student network takes labeled images and unlabeled strongly augmented images as input, which attempts to minimize the distribution mismatch between labeled and unlabeled images from the model level. That is, data augmentation is adopted on unlabeled data, whose role is to regularize the consistent training between teacher and student. As mentioned, semi-supervised multi-organ segmentation suffers from large distribution alignment mismatch between labeled and unlabeled images. Reducing the mismatch mainly from the model level is insufficient to solve the problem. Thanks to the prior anatomical knowledge from CT scans, which provides the distribution information where a multi-organ CT scan is drawn, it is possible to largely alleviate the mismatch problem from the data level.

To this end, we propose a novel teacher-student model, called *MagicNet*, matching with the rule of playing a magic-cube. More specifically, we propose a partition-and-recovery N^3 cubes learning paradigm: (1) We partition each CT scan, termed as magic-cube, into N^3 small cubes. (2) Two data augmentation strategies are then designed, as shown in Fig. 1. *I.e.*, First, to encourage unlabeled data to learn organ semantics in relative locations from the labeled data, small cubes are mixed across labeled and unlabeled images while keeping their relative locations. Second, to enhance the learning ability for small organs, small cubes are shuffled and fed into the student network. (3) We recover the magic-cube to form the original 3D geometry to map with the ground-truth or the supervisory signal from teacher. Furthermore, the quality of pseudo labels predicted by teacher network is refined by blending with the learned representation of the small cubes. The cube-wise pseudo-label blending strategy incorporates local attributes *e.g.*, texture, luster and boundary smoothness which mitigates the inferior performance of small organs.

The main contributions can be summarized as follows:

- We propose a data augmentation strategy based on partition-and-recovery N^3 cubes cross- and within- labeled and unlabeled images which encourages unlabeled images to learn organ semantics in relative locations from the labeled images and enhances the learning ability for small organs.
- We propose to correct the original pseudo-label by cube-wise pseudo-label blending via incorporating crucial local attributes for identifying targets especially small organs.
- We verify the effectiveness of our method on BTCV [13] and MACT [11] datasets. The segmentation performance of our method exceeds all state-of-the-arts by a large margin, with 7.28% (10% labeled) and 6.94% (30% labeled) improvement on two datasets respectively (with V-Net as the backbone) in DSC.

2. Related Work

2.1. Semi-supervised Medical Image Segmentation

Semi-supervised medical image segmentation methods can be roughly grouped into three categories. (1) Contrastive learning based methods [35, 38], which learn representations that maximize the similarity among positive pairs and minimize the similarity among negative pairs. (2) Consistency regularization based methods [4, 9, 14, 18, 31, 36, 39], which attend to various levels of information for a single target via multi/dual-task learning or transformation consistent learning. (3) Self-ensembling/self-training based methods [1, 15, 23, 39, 43], which generate pseudo-labels for unlabeled images and propose several strategies to ensure the quality of pseudo-labels. But, most of these methods are mainly focusing on segmenting one target or targets in a local region or ROI, such as pancreas or left atrium [18,23,38,39], which encounters performance degradation when transferring to multi-organ segmentation, due to the lack of anatomical structure modeling ability.

2.2. Semi-supervised Multi-organ Segmentation

Due to the large variations of appearance and size of different organs [22, 26, 33, 37], multi-organ segmentation has been a popular yet challenging task. Only a few SSL methods specially for multi-organ segmentation have been proposed. DMPCT [43] adopted a 2D-based co-training framework to aggregate multi-planar features on a private dataset. UMCT [36] further enforces multi-view consistency on unlabeled data. These methods are pioneer works

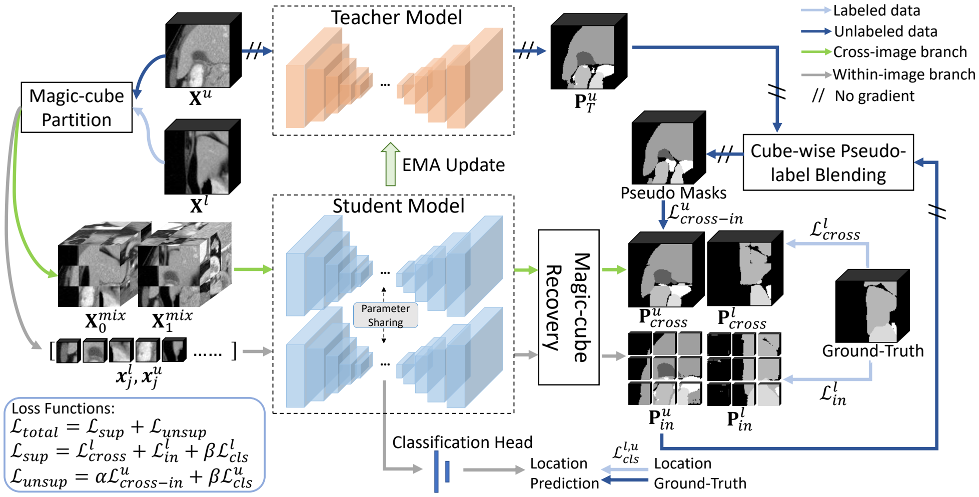

Figure 2. The architecture of MagicNet. Two essential components are proposed: (1) magic-cube partition and recovery cross- and within-image (see Sec. 3.1 and Fig. 3); (2) cube-wise pseudo-label blending, incorporating local attributes (see Sec. 3.2 and Fig. 4).

to handle semi-supervised multi-organ segmentation problem from the co-training perspective, which take advantages of the multi-view property of a CT volume, but the properties of multiple organs have not been well-explored.

2.3. Interpolation-based Semi-supervised Learning

Interpolation-based regularization [30, 40, 41] is quite successful in semi-supervised semantic segmentation. Methods such as Mixup [41] and CutMix [40] are good at synthesizing new training samples and are widely used in semi-supervised learning. FixMatch [24] design a weak-strong pair consistency to simplify semi-supervised learning. MixMatch [3] and ReMixMatch [2] generated pseudo labels for unlabeled data and incorporate gradually unlabeled data with reliable pseudo-label into the labeled set. ICT [30] proposed a interpolated-based consistency training method based on Mixup. GuidedMix-Net [29] utilized Mixup [41] in semi-supervised semantic segmentation. Thus, we also compare our approach with some popular interpolated-based methods in Experiments.

3. Method

We define the 3D volume of a CT scan as $\mathbf{X} \in \mathbb{R}^{W\times H\times L}$. The goal is to find the semantic label of each voxel $k \in \mathbf{X}$, which composes into a predicted label map $\widehat{\mathbf{Y}} \in \{0, 1, ..., C\}^{W\times H\times L}$, where $C = 0$ indicates background class, and $C \neq 0$ represents the organ class. The training set $\mathcal{D}$ consists of two subsets: $\mathcal{D} = \mathcal{D}^l \cup \mathcal{D}^u$, where $\mathcal{D}^l = \{(\mathbf{X}^l_i, \mathbf{Y}^l_i)\}^N_{i=1}$ and $\mathcal{D}^u = \{\mathbf{X}^u_i\}^{M+N}_{i=N+1}$. The training images $\mathbf{X}^l$ in $\mathcal{D}^l$ are associated with per-voxel annotations $\mathbf{Y}^l$ while those in $\mathcal{D}^u$ are not. In the rest of this paper, we denote the original and mixed CT scans as magic-cubes, and denote the partitioned small cubes as cubes for simplicity.

The overall framework of the proposed MagicNet is shown in Fig. 2, consisting of a student network and a teacher network. The teacher network is updated via weighted combination of the student network parameters with exponential moving average (EMA) [28]. Designed in this architecture, our whole framework includes two branches: Cross-image branch and within-image branch. The details will be illustrated in the following sections.

3.1. Magic-cube Partition and Recovery

Assume that a mini-batch $\mathcal{B}$ contains n images $\mathbf{X}^{\mathcal{B}} \in \mathbb{R}^{n\times W\times H\times L}$. For notational simplicity, let $n = 2$, and $\mathcal{B}$ contains one labeled image $\mathbf{X}^l$ and one unlabeled image $\mathbf{X}^u$, which are randomly sampled from $\mathcal{D}^l$ and $\mathcal{D}^u$, respectively. We partition $\mathbf{X}^l$ and $\mathbf{X}^u$ into N^3 magic-cubes $\{\mathbf{x}^u_j\}^{N^3}_{j=1}$ and $\{\mathbf{x}^l_j\}^{N^3}_{j=1}$, where $\mathbf{x}^l_j, \mathbf{x}^u_j \in \mathbb{R}^{W/N\times H/N\times L/N}$, j represents the relative location of the cube inside this magic-cube. This partition operation is called as $\mathcal{A}_{part}(\cdot)$.

Cross-image Partition and Recovery To encourage labeled and unlabeled images to learn comprehensive common semantics from each other, we mix these cubes across all the labeled and unlabeled images in a mini-batch. As shown in Fig. 3, these cubes $\mathbf{x}^l_j$ and $\mathbf{x}^u_j$ are mixed into two shuffled magic-cubes while keeping their original positions, which produce two interpolated images with mixed magic-cubes $\mathbf{X}^{mix}_0, \mathbf{X}^{mix}_1 \in \mathbb{R}^{W\times H\times L}$. The magic-cube mixing cross-image operations are termed as $\mathcal{A}^{cross}_{mix}(\cdot)$. The mixed images are then fed into the student network $\mathcal{F}(\cdot; \mathbf{\Theta}^s)$, followed by a softmax layer $\sigma(\cdot)$, obtaining the prediction maps $\mathbf{P}^{mix}_0, \mathbf{P}^{mix}_1 \in \mathbb{R}^{C\times W\times H\times L}$, where C denotes the number of classes. Next, we recover $\mathbf{P}^l_{cross}, \mathbf{P}^u_{cross} \in \mathbb{R}^{C\times W\times H\times L}$ from $\mathbf{P}^{mix}_0$ and $\mathbf{P}^{mix}_1$ via recovering the magic-cubes back to the original positions in their original

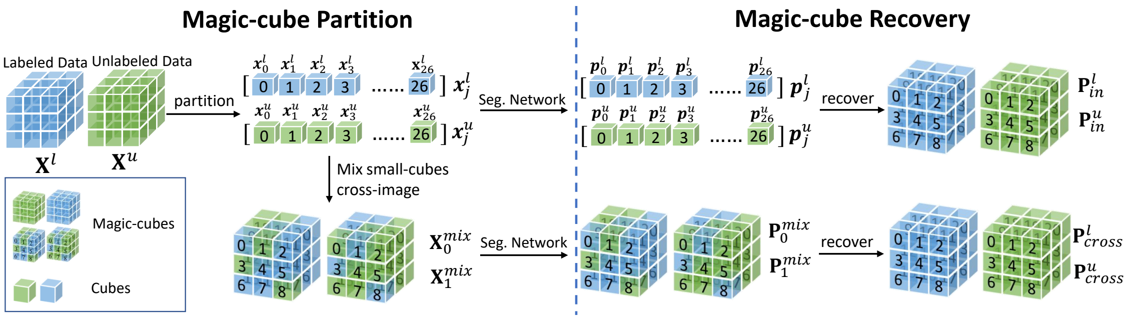

Figure 3. Magic-cube partition and recovery. Blue and green volumes represent labeled and unlabeled image, respectively. The number on cubes represents its relative-location in the original image.

images, and this operation is denoted as $\mathcal{A}_{rec}^{cross}(\cdot)$. We can simply denote the whole process as:

$$\mathbf{P}_{cross}^l, \mathbf{P}_{cross}^u = \mathcal{A}_{rec}^{cross}(\sigma(\mathcal{F}(\mathcal{A}_{mix}^{cross}(\mathcal{A}_{part}(\mathbf{X}^{\mathcal{B}})); \mathbf{\Theta}^s))). \quad (1)$$

Here, $\mathbf{P}_{cross}^l, \mathbf{P}_{cross}^u \in \mathbb{R}^{(C+1)\times W\times H\times L}$. The loss functions for labeled and unlabeled images are defined as

$$\mathcal{L}_{cross}^l(\mathcal{B}; \mathbf{\Theta}^s) = \ell_{dice}(\mathbf{P}_{cross}^l, \mathbf{Y}^l), \quad (2)$$

$$\mathcal{L}_{cross-in}^u(\mathcal{B}; \mathbf{\Theta}^s) = \ell_{dice}(\mathbf{P}_{cross}^u, \widehat{\mathbf{Y}}^u). \quad (3)$$

where ℓ_{dice} denotes multi-class Dice loss. $\mathbf{Y}^l \in \{0, 1, ..., C\}^{W\times H\times L}$ denotes per-voxel manual annotations for $\mathbf{X}^l$. $\mathcal{L}_{cross-in}^u$ indicates the refined pseudo-labels for $\mathbf{X}^u$ by blending cube-wise representations and more details can be seen in Sec. 3.2.

Within-image Partition and Recovery Besides magic-cubes partition and recovery across images, we also design a within-image partition and recovery branch for single images, which can better consider local features and learn local attributes for identifying targets, especially targets of small sizes. For $\mathbf{X}^l$, the partitioned j-th cube $\mathbf{x}_j^l$ is fed into $\mathcal{F}(\cdot; \mathbf{\Theta}^s)$ and the softmax layer, which can acquire a cube-wise prediction map $\mathbf{p}_j^l \in \mathbb{R}^{C\times W/N\times H/N\times L/N}$. Finally, we recover the magic-cube via mixing N^3 cube-wise prediction maps back to their original positions as illustrated in Fig. 3, and the recovered probability map is denoted as $\mathbf{P}_{in}^l \in \mathbb{R}^{C\times W\times H\times L}$. Similarly, for $\mathbf{X}^u$, the probability map is denoted as $\mathbf{P}_{in}^u \in \mathbb{R}^{C\times W\times H\times L}$. This operation is denoted as $\mathcal{A}_{rec}^{in}(\cdot)$. The whole process for obtaining $\mathbf{P}_{in}^l$ and $\mathbf{P}_{in}^u$ are:

$$\mathbf{P}_{in}^l, \mathbf{P}_{in}^u = \mathcal{A}_{rec}^{in}(\sigma(\mathcal{F}(\mathcal{A}_{part}(\mathbf{X}^{\mathcal{B}}); \mathbf{\Theta}^s))). \quad (4)$$

Since $\mathbf{P}_{in}^l$ and $\mathbf{P}_{in}^u$ are recovered via predictions of cubes, $\mathbf{P}_{in}^l$ and $\mathbf{P}_{in}^u$ are denoted as cube-wise representations in the rest of the paper. The loss function for the labeled image is:

$$\mathcal{L}_{in}^l(\mathcal{B}; \mathbf{\Theta}^s) = \ell_{dice}(\mathbf{P}_{in}^l, \mathbf{Y}^l). \quad (5)$$

For the unlabeled image, we propose to leverage the cube-wise representations by refining the pseudo-label instead of directly computing the loss function as in Eq. 5, which is illustrated in Sec. 3.2. We conduct ablation studies on various design choices of utilizing the cube-wise representations in the experiment part.

3.2. Cube-wise Pseudo-label Blending

As mentioned in the previous section, medical imaging experts reveal that local attributes *e.g.*, texture, luster and boundary smoothness are crucial elements for identifying targets such as tiny organs in medical images [42]. Inspired by this, we propose a within-image partition and recovery module to learn cube-wise local representations. Since the teacher network takes the original volume as input, and pays more attention to learn large organs in the head class, voxels that actually belong to the *tail* class can be incorrectly predicted as *head* class voxels, due to the lack of local attributes learning. To effectively increase the chance of the voxel predicted to the tail class, we design a cube-wise pseudo-label blending module, which blends the original pseudo-label with cube-wise features.

In detail, the image-level prediction map of $\mathbf{X}^u$ acquired from the teacher network is defined as $\mathbf{P}_T^u = \mathcal{F}(\mathbf{X}^u; \mathbf{\Theta}^t)$, where $\mathbf{\Theta}^t$ is the parameters of teacher network. The reconstructed cube-level representations of $\mathbf{X}^u$ acquired from the student network is $\mathbf{P}_{in}^u$. As mentioned above, cube-wise features pay more attention to learn local attributes which is important for tiny organs. Therefore, we propose a distribution-aware blending strategy, which is shown in Fig. 4(a), and formulated as:

$$\mathbf{P}_{blend}^u = (\mathbf{1} - \mathcal{R}(\mathbf{\Omega})) \odot \mathbf{P}_T^u + \mathcal{R}(\mathbf{\Omega}) \odot \mathbf{P}_{in}^u, \quad (6)$$

where $\odot$ indicates the element-wise multiplication, $\mathbf{1} \in \{1\}^{(C+1)\times W\times H\times L}$, and $\mathbf{\Omega} \in \mathbb{R}^{W\times H\times L}$ is a distribution-aware weight map.

To obtain the weight map, we firstly learn the class-wise distribution $\mathbf{v}$ during training. Suppose $\mathbf{v} \in \mathbb{R}^C$ is a vector, $\mathbf{v} = \{\mathbf{v}_0, ..., \mathbf{v}_{C-1}\}$, whose element $\mathbf{v}_c$ indicates the number of voxels belonging to $(c+1)$th organ, which is accumulated by counting the voxels *w.r.t.* pseudo-labels over a few previ-

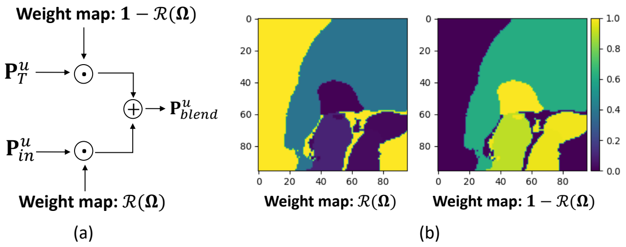
Weight map: $\mathbf{1}-\mathcal{R}(\boldsymbol{\Omega})$
$\mathbf{P}^u_T$
$\mathbf{P}^u_{blend}$
$\mathbf{P}^u_{in}$
Weight map: $\mathcal{R}(\boldsymbol{\Omega})$
(a)
0
20
40
60
80
Weight map: $\mathcal{R}(\boldsymbol{\Omega})$
Weight map: $\mathbf{1}-\mathcal{R}(\boldsymbol{\Omega})$
1.0
0.8
0.6
0.4
0.2
0.0
(b)

Figure 4. The cube-wise pseudo-label blending module. (a) Pipeline. (b) Illustration on $\mathcal{R}(\boldsymbol{\Omega}_{\mathbf{m}})$. If the pseudo-label $\widetilde{\mathbf{Y}}^u_{\mathbf{m}}$ on $\mathbf{m}$ is incorrectly assigned as a big organ, a big $\boldsymbol{\Omega}_{\mathbf{m}}$ is assigned to ensure possible correction from cube-wise representations.

ous iterations. Let $\widetilde{\mathbf{Y}}^u_{\mathbf{m}}$ denote the pseudo-label for the voxel on location $\mathbf{m}$. The value $\boldsymbol{\Omega}_{\mathbf{m}} \in \mathbb{R}^1$at each spatial location $\mathbf{m}$ on the weight map is derived as:

$$\boldsymbol{\Omega}_{\mathbf{m}} = \sum_{c=0}^{C-1} \frac{\mathbf{I}(\widetilde{\mathbf{Y}}^u_{\mathbf{m}} = c)\mathbf{v}_c}{\max \mathbf{v}} \qquad (7)$$

where $\mathbf{I}(\cdot)$ is an indicator function. Note that to match with the data dimension in Eq. 6, we replicate $\boldsymbol{\Omega}$ $C+1$ times to acquire $\mathcal{R}(\boldsymbol{\Omega}) \in \mathbb{R}^{(C+1)\times W\times H\times L}$.

Overall, since the teacher network takes the original volume as input, and pays more attention to learn large organs in the head class, the teacher pseudo-label may be biased. To unbias the pseudo-label, on the one hand, we keep the pseudo-label of small organs. On the other hand, we remedy the possible incorrect pseudo-labels of big organs via changing them to small-organ classes via Eq. 6 if $\boldsymbol{\Omega}_{\mathbf{m}}$ is large. This can effectively increase the chance of the voxels predicted to small-organs.

The final refined pseudo-label $\widehat{\mathbf{Y}}^u$ is obtained via $\widehat{\mathbf{Y}}^u_{\mathbf{m}} = \arg\max_{c\in\mathcal{C}} \mathbf{P}^u_{blend,c,\mathbf{m}}$, where $\mathcal{C}$ indicates the finite set of class labels, and $\mathbf{m}$ means the voxel location.

3.3. Magic-cube Location Reasoning

To fully leverage the prior anatomical knowledge of multi-organs, we further propose a magic-cube location reasoning method, which is in line with within-image partition. For one image $\mathbf{X}^l$, the partitioned j-th cube's $\mathbf{x}^l_j$ is fed into the encoder of student network $\mathcal{F}_{enc}(\cdot;\boldsymbol{\Theta}^s_{enc})$ to generate $\mathbf{f}^l_j \in \mathbb{R}^{D\times W/N\times H/N\times L/N}$. Then, the feature $\mathbf{f}^l_j$ is flattened into a one-dimensional vector, passing through a classification head $\mathcal{F}_{cls}(\cdot;\boldsymbol{\Theta}_{cls})$ to project the flattened feature vector into the same size as the number of cubes in one image, *i.e.*, N^3. The classification head is composed of two fully connected layers [10, 34]. Finally, we concatenate the output $\hat{\mathbf{y}}^l_j \in \mathbb{R}^{N^3}$ from the classification head for all N^3cubes in one image, which generates a vector in size $(N^3)^2$. This vector is further reshaped into a matrix of size $N^3 \times N^3$, whose rows indicate the probabilities of corresponding cubes belonging to the N^3 locations. The same applies to $\mathbf{X}^u$. For both $\mathbf{X}^l$ and $\mathbf{X}^u$ in $\mathcal{B}$, the cross-entropy loss is adopted to learn relative locations of cubes:

$$\mathcal{L}_{cls}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}_{cls}) = \frac{1}{|\mathcal{B}|}\sum_{\mathbf{X}\in\mathcal{B}} \ell_{ce}(\sigma(\mathcal{F}_{cls}(\mathcal{F}^s_{enc}(\mathcal{A}_{part}(\mathbf{X});\boldsymbol{\Theta}^s_{enc});\boldsymbol{\Theta}_{cls})),\mathbb{y}) \qquad (8)$$

where ℓ_{ce} denotes the cross-entropy loss, $\mathbb{y}$ represents cube's relative locations in $\mathbf{X}$.

3.4. Loss Function

Overall, we develop a total loss function based on a mini-batch for our MagicNet as follows:

$$\mathcal{L}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}^s_{seg},\boldsymbol{\Theta}_{cls}) = \mathcal{L}_{sup}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}^s,\boldsymbol{\Theta}_{cls}) + \mathcal{L}_{unsup}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}^s,\boldsymbol{\Theta}_{cls}), \qquad (9)$$

where $\mathcal{L}_{sup}(\cdot)$ and $\mathcal{L}_{unsup}(\cdot)$ denote the loss for labeled and unlabeled images in $\mathcal{B}$, respectively:

$$\mathcal{L}_{sup}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}^s,\boldsymbol{\Theta}_{cls}) = \mathcal{L}^l_{cross}(\mathcal{B};\boldsymbol{\Theta}^s) + \mathcal{L}^l_{in}(\mathcal{B};\boldsymbol{\Theta}^s) + \beta\mathcal{L}^l_{cls}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}_{cls}), \qquad (10)$$

where β is the weight factor to balance the location reasoning term with others, and

$$\mathcal{L}_{unsup}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}^s,\boldsymbol{\Theta}_{cls}) = \alpha\mathcal{L}^u_{cross-in}(\mathcal{B};\boldsymbol{\Theta}^s) + \beta\mathcal{L}^u_{cls}(\mathcal{B};\boldsymbol{\Theta}^s_{enc},\boldsymbol{\Theta}_{cls}), \qquad (11)$$

where α is another weight factor to balance the two terms. For $\mathcal{L}^l_{cls}$ and $\mathcal{L}^u_{cls}$ in Eq. 10 and Eq. 11, the following equation holds:

$$\mathcal{L}_{cls} = \mathcal{L}^l_{cls} + \mathcal{L}^u_{cls}. \qquad (12)$$

3.5. Testing Phase

Given a testing image $\mathbf{X}_{test}$, the probability map is obtained by: $\mathbf{P}_{test} = \sigma(\mathcal{F}(\mathbf{X}_{test};\boldsymbol{\Theta}^{s*}))$, where $\boldsymbol{\Theta}^{s*}$ denotes the well-trained student network parameters. The final label map can be determined by taking the *argmax* in the first dimension of $\mathbf{P}_{test}$.

4. Experiments

4.1. Datasets and Pre-processing

BTCV Multi-organ Segmentation Dataset BTCV multi-organ segmentation dataset is from *MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge* [13] which contains 30 subjects with 3779 axial abdominal CT slices with 13 organs annotation. In pre-processing, we follow [27] to re-sample all the CT scans to the voxel spacing $[1.5\times1.5\times2.0]mm^3$ and normalize them to have zero mean and unit variance. Strictly following [5, 6, 10], 18 cases are divided for training and the other 12 cases are for testing.

Labeled%	Methods	Spl	R.kid	L.kid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	RG	LG	Avg. DSC	Avg. NSD
30%	V-Net [20]	86.37	75.38	71.20	41.81	48.46	92.77	44.78	88.20	72.60	49.27	22.30	37.79	21.71	57.90	56.90
	MT [28]	79.89	77.56	78.08	38.31	58.99	92.26	48.73	88.61	79.36	52.73	28.42	54.16	21.30	61.42	60.85
	UA-MT [39]	87.75	73.87	72.97	42.08	50.79	93.22	42.07	88.30	71.80	55.27	36.61	43.22	37.67	61.20	60.99
	ICT [30]	88.78	78.00	83.93	50.09	53.35	88.32	64.19	89.02	79.60	63.51	58.90	43.29	50.70	68.59	67.77
	CPS [7]	82.98	76.91	79.21	29.91	52.53	92.59	41.37	86.59	74.73	49.53	18.96	45.13	24.06	58.04	56.64
	SS-Net [35]	85.53	73.32	80.70	30.85	52.44	91.67	17.67	86.28	72.39	47.40	22.95	37.38	39.01	56.74	54.38
	SLC-Net [16]	90.04	84.38	85.84	53.55	55.87	92.86	57.80	89.83	80.87	53.22	38.78	48.10	37.03	66.78	68.81
	MagicNet (ours)	**91.42**	**84.64**	**86.19**	**62.86**	**62.49**	**93.89**	**72.87**	**90.70**	**83.52**	**70.07**	**64.94**	**60.88**	**57.48**	**75.53**	**76.31**
40%	V-Net [20]	84.98	82.72	82.07	36.64	63.48	93.54	57.49	89.74	78.63	60.42	49.39	55.60	38.49	67.17	67.84
	MT [28]	85.70	78.93	79.08	42.80	61.09	93.45	57.57	89.70	80.30	63.95	41.14	50.46	29.69	65.68	65.98
	UA-MT [39]	88.74	75.88	78.91	54.25	58.55	93.46	58.90	89.23	76.15	62.30	47.91	51.53	44.92	67.75	68.87
	ICT [30]	90.31	84.41	86.96	49.22	65.65	94.29	65.95	90.23	81.44	69.56	66.61	57.35	56.01	73.69	74.98
	CPS [7]	87.56	72.99	77.59	53.31	54.08	92.41	54.58	87.75	74.32	58.68	48.02	50.39	43.86	65.81	65.34
	SS-Net [35]	84.74	76.37	74.19	43.42	57.05	92.90	14.37	83.14	69.77	52.45	27.08	54.29	27.66	58.26	57.75
	SLC-Net [16]	90.05	84.00	86.43	56.16	58.91	**94.68**	70.72	89.93	79.45	60.59	54.22	51.03	39.08	70.40	72.87
	MagicNet (ours)	**91.61**	**85.02**	**88.13**	**58.16**	**66.72**	94.07	**74.46**	**90.77**	**84.31**	**71.56**	**68.90**	**63.48**	**60.47**	**76.74**	**78.68**
100%	V-Net [20]	84.00	84.82	86.38	67.42	65.02	94.83	73.75	90.27	84.19	69.85	63.54	62.60	65.02	76.28	77.45

Table 1. Comparison results (DSCs of each organ, avg. DSC, and avg. NSD) between our method and existing semi-supervised medical image segmentation methods on the ***BTCV dataset***. **V-Net** means training with only labeled samples on the V-Net backbone. **DSC**: Dice Similarity Coefficient. **NSD**: Normalized Surface Dice. The standard deviations are not reported due to space limit. Note: Spl: spleen, R.Kid: right kidney, L.Kid: left kidney, Gall: gallbladder, Eso: esophagus, Liv: liver, Sto: stomach, Aor: aorta, IVC: inferior vena cava, Veins: portal and splenic veins, Pan: pancreas, LG/RG: left/right adrenal glands.

Labeled%	Methods	Spleen	L.kidney	Gallbladder	Esophagus	Liver	Stomach	Pancreas	Duodenum	Avg. DSC	Avg. NSD
10%	V-Net [20]	88.89 (16.31)	84.90 (20.11)	55.73 (34.27)	58.27 (20.05)	93.15 (7.11)	57.15 (30.62)	53.33 (22.44)	35.28 (19.83)	65.84	49.96
	MT [28]	89.22 (13.56)	88.04 (17.61)	61.40 (31.91)	58.89 (20.62)	93.60 (5.04)	74.00 (20.37)	64.70 (19.31)	44.53 (17.61)	71.80	53.54
	UA-MT [39]	89.94 (13.36)	89.26 (15.67)	59.19 (32.35)	59.43 (19.45)	93.77 (5.05)	75.43 (19.61)	65.86 (17.48)	44.57 (17.40)	72.18	54.04
	ICT [30]	90.27 (13.70)	89.89 (15.81)	61.47 (30.31)	58.70 (20.17)	93.47 (5.39)	72.45 (20.95)	65.81 (18.28)	45.00 (19.22)	72.13	54.07
	CPS [7]	91.73 (12.11)	90.59 (14.85)	60.54 (32.29)	55.82 (21.03)	94.19 (4.93)	78.80 (15.28)	62.93 (20.03)	32.58 (18.68)	70.90	52.80
	SS-Net [35]	83.52 (19.56)	78.28 (25.05)	53.08 (33.88)	56.24 (20.54)	91.84 (6.07)	47.65 (29.70)	43.23 (24.11)	25.70 (18.71)	59.94	42.68
	SLC-Net [16]	86.89 (17.54)	87.54 (18.00)	55.28 (34.75)	58.51 (20.54)	93.70 (5.33)	69.74 (22.89)	46.55 (26.21)	28.86 (18.87)	65.89	13.28
	MagicNet (ours)	**92.33** (11.12)	**91.19** (14.18)	**66.35** (29.29)	**70.95** (12.30)	**94.31** (4.64)	**81.56** (14.91)	**76.31** (10.56)	**61.50** (12.85)	**79.31**	**62.31**
20%	V-Net [20]	90.53 (15.33)	88.28 (19.76)	65.16 (30.03)	66.10 (16.98)	94.49 (4.18)	79.56 (16.56)	69.04 (17.43)	51.98 (18.53)	75.64	59.66
	MT [28]	91.76 (12.77)	91.40 (14.49)	63.83 (31.74)	64.15 (18.05)	94.44 (4.86)	81.19 (15.79)	72.24 (15.79)	57.01 (16.70)	77.00	60.32
	UA-MT [39]	92.39 (11.31)	91.44 (14.61)	63.63 (31.46)	64.33 (17.92)	94.50 (4.77)	83.21 (14.03)	71.93 (15.51)	57.21 (17.35)	77.33	60.64
	ICT [30]	92.96 (9.60)	91.67 (14.52)	65.90 (30.49)	64.35 (16.79)	94.46 (4.66)	82.60 (15.46)	74.18 (12.99)	57.67 (16.16)	77.97	61.24
	CPS [7]	92.66 (11.46)	91.87 (14.33)	64.75 (31.21)	57.27 (19.56)	**94.97** (4.51)	85.65 (9.50)	74.15 (12.49)	55.00 (16.58)	77.04	60.01
	SS-Net [35]	91.17 (14.77)	87.77 (19.37)	63.49 (30.83)	65.02 (16.65)	93.76 (4.82)	73.04 (22.16)	70.37 (17.70)	52.81 (20.32)	74.68	57.57
	SLC-Net [16]	92.60 (11.24)	91.38 (14.57)	62.46 (31.37)	62.84 (19.31)	94.54 (4.72)	80.30 (14.73)	69.79 (16.99)	53.48 (18.40)	75.92	59.93
	MagicNet (ours)	**93.52** (10.22)	**92.01** (14.30)	**71.04** (26.92)	**70.95** (12.58)	94.89 (4.42)	**85.77** (10.27)	**78.81** (8.07)	**66.21** (12.33)	**81.65**	**65.87**
100%	V-Net [20]	94.35 (5.81)	92.30 (14.22)	73.61 (28.80)	73.48 (12.43)	95.17 (3.10)	89.09 (7.74)	80.86 (7.38)	70.03 (11.60)	83.61	68.85

Table 2. Comparison results (DSCs of each organ, avg. DSC, and avg. NSD) between our method and existing semi-supervised medical image segmentation methods on ***MACT dataset*** under 4-fold cross-validation. The value in $(\cdot)$ is standard deviation.

MACT Dataset Multi-organ abdominal CT reference standard segmentations (MACT) is a public multi-organ segmentation dataset, containing 90 CT volumes with 8 organs annotation. It is manually re-annotated by [11] based on 43 cases of NIH pancreas [21] and 47 cases of Synapse [13]. We follow the setting of UMCT [36]. 1) In preprocessing, we set the soft tissue CT window range of $[-125, 275]$ HU and all the CT scans are re-sampled to an isotropic resolution of $1.0mm^3$. Then, image intensities are normalized to have zero mean and unit variance. 2) We randomly split the dataset into four folds, and perform 4-fold cross-validation. Labeled training set is then randomly selected, which are applied to all experiments for fair comparison.

4.2. Experimental Setup and Evaluation Metrics

In this work, we conduct all the experiments based on PyTorch with one NVIDIA 3090 GPU and use V-Net as our backbone [16, 35, 39]. For magic-cube location reasoning, we add a classification head composed of two fully-connected layers at the end of our encoder. Firstly, for BTCV dataset, the framework is trained by an SGD optimizer for 70K iterations, with an initial learning rate (lr) 0.01 with a warming-up strategy: $lr = base\_lr \times (1 - \frac{iterations}{max\_iterations})^{0.9}$. For MACT dataset, the framework is trained by an SGD optimizer for 30K iterations, with an initial learning rate (lr) 0.01 decayed by 0.1 every 12K iterations. Following [16, 18, 35, 36, 39], for both two datasets, batch size is set as 4, including 2 labeled images and 2 unlabeled images, and the hyper-parameter α is set using a time-dependent Gaussian warming-up function and β is empirically set as 0.1 [10]. We randomly crop volume of size 96^3, and then we partition the volume into N^3 small cubes, forming our magic-cubes, and the choice of N is also discussed in Sec. 4.4. In the testing phase, a sliding window strategy is adopted to obtain final results with a stride of 16^3. The final evaluation metrics for our method are DSC (%, Dice-Sørensen Coefficient) and NSD (%, Normalized Surface Dice) which are commonly used in the cur-

Methods	Cross	In	Loc	Bld	Avg. DSC	spleen	l.kidney	gallbladder	esophagus	liver	stomach	pancreas	duodenum
Baseline (MT [28])					71.88	86.84	88.92	54.50	53.40	93.01	70.86	71.77	55.76
Cross	✓				78.13	92.11	90.53	63.42	70.05	93.90	79.60	75.58	59.87
Cross + In	✓	✓			78.83	92.82	90.75	64.91	70.68	93.90	80.92	76.25	60.40
Cross + Loc	✓		✓		78.56	92.49	91.07	65.95	69.72	94.12	80.40	75.82	58.94
Cross + In + Loc	✓	✓	✓		78.87	92.55	91.18	65.72	70.77	94.42	80.05	75.81	60.43
Cross + In + Loc + Bld	✓	✓	✓	✓	79.31	92.33	91.19	66.35	70.95	94.31	81.56	76.31	61.50

Table 3. Ablation study (DSC %) for the effectiveness of each component of MagicNet. **Cross**: cross-image partition-and-recovery. **In**: Within-image partition-and-recovery. **Bld**: cube-wise pseudo-label blending. **Loc**: magic-cube location reasoning.

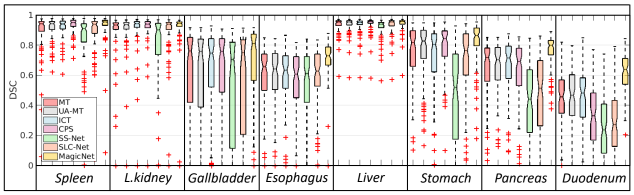

Figure 5. DSC comparison in box plots on MACT dataset with 10% labeled images (Best viewed electronically, zoom in).

rent multi-organ segmentation challenge [12, 19].

4.3. Comparison with the State-of-the-art Methods

We compare MagicNet with six state-of-the-art semi-supervised segmentation methods, including mean-teacher (MT) [28], uncertainty-aware mean-teacher (UA-MT) [39], interpolated consistency training (ICT) [30], cross pseudo-supervision (CPS) [7], smoothness and class-separation (SS-Net) [35], shape-awareness and local constraints (SLC-Net) [16]. All experiments of other methods are re-trained by their official code when transferred to multi-organ task. V-Net is adopted as our backbone which is the same as the above methods for fair comparison.

In Table 1, we summarize the results of BTCV dataset. Visualizations compared with state-of-the-arts are shown in the supplementary material. Noticeable improvements compared with state-of-the-arts can be seen for organs such as Stomach (↑ 8.68), Gallbladder (↑ 9.31), Veins (↑ 6.56), Right adrenal glands (↑ 6.72) and Left adrenal glands (↑ 6.78). It is interesting to observe that some semi-supervised methods even perform worse than the lower bound. This is because these methods are less able to learn common semantics from the labeled images to unlabeled ones, generating less accurate pseudo-labels, which even hamper the usage of unlabeled images (see the supplementary material).

We then evaluate MagicNet on MACT dataset. In Table 2, we can observe: (1) a more significant performance improvement between MagicNet and other state-of-the-arts with 10% labeled data than 20%, which demonstrates the effectiveness of MagicNet when training with fewer labeled CT images; (2) MagicNet successfully mitigates the inferior performance in tail classes, *e.g.*, DSC improvement with 10% labeled images on Gallbladder (↑ 4.88), Esophagus (↑ 11.52), Pancreas (↑ 10.45) and Duodenum (↑ 16.50). Fig. 5 shows a comparison in DSC of six methods for different organs. Note that UMCT [36] on MACT dataset achieves 75.33% in DSC. It uses AH-Net [17] as the backbone (3 are used), whose parameter number is much larger. Due to computation resources, we do not directly compare with it.

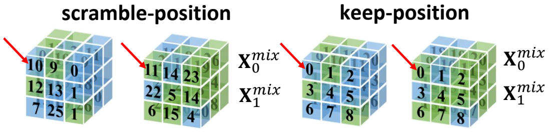

Figure 6. Scramble position and keep position when mixing small-cubes across images (from Fig. 3). Red arrow indicates the cube location in mixed images.

4.4. Ablation Study

We conduct ablation studies on MACT dataset [11] with 10% labeled data to show the effectiveness of each module.

The effectiveness of each component in MagicNet We conduct ablation studies to show the impact of each component in MagicNet. In Table 3, the first row indicates the mean-teacher baseline model [28], which our method is designed on. Compared to the baseline, our magic-cube partition and recovery can yield good segmentation performance. **Cross** and **In** represent our cross- and within-image partition-and-recovery, which increase the performance from 71.88% to 78.13% and 78.87%, respectively. This shows the powerful ability with our specially-designed data augmentation method. **Loc** represents magic-cube location reasoning for within-image branch. We can see from the table that based on our cross-image branch, adding relative locations of small-cubes can achieve 78.56% in DSC, adding only within-image partition-and-recovery branch achieves 78.83%, and adding both two branches leads to 78.87%. Finally, our proposed cube-wise pseudo label blending (short for **Bld** in Table 3) provides significant improvement to 79.31%, for which we will make a more complete analysis later in the current section.

Design choices of partition and recovery Here, we discuss the design for cross-image partition-and-recovery branch: ① Should we maintain or scramble the magic-cube relative locations when manipulating our cross-image branch, as shown in Fig. 6? The comparison results are shown in the last two rows in Table 4. **Scramble** and **Keep** represent the

①	②	DSC	NSD
scramble	U	69.53 ± 11.27	50.98 ± 11.00
keep	U	73.84 ± 9.96	55.80 ± 10.81
scramble	LU	71.91 ± 11.20	54.10 ± 11.28
keep	LU	78.13 ± 8.12	60.79 ± 9.90

Table 4. Ablation of design choices for cross-image partition and recovery (Question ① and ②, mean ± std of all cases). **Scramble/keep**: ignore/keep original positions when mixed. **U**: only for unlabeled data. **LU**: for both labeled and unlabeled data. The last row is ours.

	DSC	NSD
teacher sup.	78.47 ± 7.83	61.52 ± 9.21
mutual sup.	70.45 ± 11.47	53.71 ± 11.78
blending	79.31 ± 7.55	62.31 ± 9.08

Table 5. Comparison of different pseudo-label supervision/blending strategies for unlabeled data. **Sup.**: supervision. **Blending**: our cube-wise pseudo-label blending.

	DSC	NSD
CutMix [40]	73.80 ± 9.99	55.86 ± 10.54
CutOut [8]	72.47 ± 10.27	54.62 ± 10.67
MixUp [41]	72.13 ± 11.32	54.07 ± 12.01
Ours (2)	78.68 ± 7.45	60.78 ± 9.12
Ours (3)	79.31 ± 7.55	62.31 ± 9.08

Table 6. Comparison of different data augmentation methods, where we try different CutMix and CutOut sizes, and choose the best results. For MagicNet, we compare different N values in $(\cdot)$.

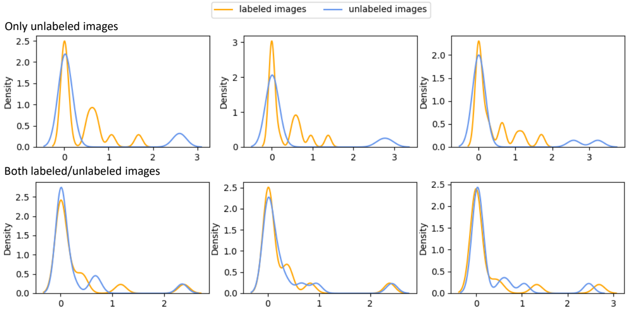

Figure 7. Kernel Density Estimation of labeled and unlabeled samples on the first three channels of the feature map of "only unlabeled images" and "both labeled and unlabeled images". Considerable distribution mismatch between labeled and unlabeled images can be observed (top), while our method can well improve the distribution mismatch (bottom). Best viewed zoom in.

partitioned small-cubes are randomly mixed while ignoring their original locations or kept when mixing them across images. Results show that the relative locations between multiple organs are important for multi-organ segmentation. ② Should our cross-image data augmentation be operated on only unlabeled images (see **U** in Table 4) or both labeled and unlabeled images (see **LU** in Table 4)? The latter obtains a much better performance compared to the former. Besides, it largely reduces the distribution mismatch between labeled and unlabeled data, as shown in Fig. 7.

Cube-wise pseudo-label blending As illustrated in Sec. 3.2, we blend the output of within-image branch and the output of teacher model to obtain the final pseudo-label for complementing local attributes. We compare our blending with two other methods, as shown in Table 5. Three supervision schemes, illustrated in Fig. 8, are compared for unlabeled images based on our framework. **Teacher sup.** means the outputs of cross- and within-image branch are both supervised by the pseudo-label from the teacher model. **Mutual sup.** means the outputs of cross- and within-image are mutually supervised. It can be seen that our blending strategy works favorably for unlabeled data.

Comparison with interpolated-based methods As shown in Table 6, our augmentation method outperforms other methods such as CutMix, CutOut and MixUp. For fair comparison, we try several CutMix and CutOut sizes, and choose the best results.

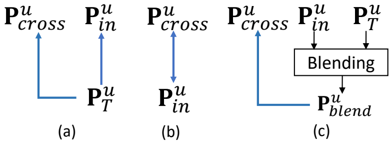

Figure 8. Different pseudo-label supervision/blending strategies corresponding to Table 5. **Blue line indicates supervision.** (a) teacher sup. (b) mutual sup. (c) our blending.

Different number (N) of small-cubes We study the impact of different numbers of small-cubes N, as shown in Table 6. Slightly better performance is achieved when $N = 3$. When $N = 4$, the size of the small cube does not match the condition for VNet. Thus, we only compare the results given $N = 2$ and $N = 3$.

5. Limitations and Conclusion

We have presented *MagicNet* for semi-supervised multi-organ segmentation. MagicNet encourages unlabeled data to learn organ semantics in relative locations from the labeled data, and enhances the learning ability for small organs. Two essential components are proposed: (1) partition-and-recovery N^3 small cubes cross- and within-labeled and unlabeled images, and (2) the quality of pseudo label refinement by blending the cube-wise representations. Experimental results on two public multi-organ segmentation datasets verify the superiority of our method.

Limitations: CT scans are all roughly aligned by the nature of the data. Magic-cube may not work well on domains *w/o* such prior, *e.g.*, microscopy images *w/o* similar prior.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 62101191), Shanghai Natural Science Foundation (Grant No. 21ZR1420800), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ2229004).

References

[1] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew P. King, Paul M. Matthews, and Daniel Rueckert. Semi-supervised learning for network-based cardiac MR image segmentation. In *Proc. MICCAI*, 2017. 2

[2] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *Proc. ICLR*, 2020. 3

[3] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, 2019. 3

[4] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen de Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Proc. MICCAI*, 2019. 1, 2

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 5

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 5

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proc. CVPR*, 2021. 6, 7

[8] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 8

[9] Kang Fang and Wu-Jun Li. Dmnet: Difference minimization network for semi-supervised segmentation in medical images. In *Proc. MICCAI*, 2020. 1, 2

[10] Shuhao Fu, Yongyi Lu, Yan Wang, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Domain adaptive relational reasoning for 3d multi-organ segmentation. In *Proc. MICCAI*, 2020. 5, 6

[11] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Trans. Medical Imaging*, 37(8):1822–1834, 2018. 2, 6, 7

[12] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 7

[13] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Workshop*, 2015. 2, 5, 6

[14] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *Proc. MICCAI*, 2020. 1, 2

[15] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, and Pheng-Ann Heng. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. *IEEE Trans. Neural Networks and Learning Systems*, 32(2):523–534, 2020. 1, 2

[16] Jinhua Liu, Christian Desrosiers, and Yuanfeng Zhou. Semi-supervised medical image segmentation using cross-model pseudo-supervision with shape awareness and local context constraints. In *Proc. MICCAI*, 2022. 6, 7

[17] Siqi Liu, Daguang Xu, S Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. 3d anisotropic hybrid network: Transferring convolutional features from 2d images to 3d anisotropic volumes. In *Proc. MICCAI*, 2018. 7

[18] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proc. AAAI*, 2021. 1, 2, 6

[19] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6695–6714, 2022. 7

[20] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proc. 3DV*, 2016. 6

[21] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Proc. MICCAI*, 2015. 6

[22] Holger R. Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Comput. Medical Imaging Graph.*, 66:90–99, 2018. 2

[23] Yinghuan Shi, Jian Zhang, Tong Ling, Jiwen Lu, Yefeng Zheng, Qian Yu, Lei Qi, and Yang Gao. Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation. *IEEE Trans. Medical Imaging*, 41(3):608–620, 2022. 1, 2

[24] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, 2020. 3

[25] Hao Tang, Xuming Chen, Yang Liu, Zhipeng Lu, Junhua You, Mingzhou Yang, Shengyu Yao, Guoqi Zhao, Yi Xu, Tingfeng Chen, Yong Liu, and Xiaohui Xie. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat. Mach. Intell.*, 1(10):480–491, 2019. 1

[26] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo

Bermudez, Michael R. Savona, Richard G. Abramson, Shunxing Bao, Ilwoo Lyu, Yuankai Huo, and Bennett A. Landman. High-resolution 3d abdominal segmentation with random patch network fusion. *Medical Image Anal.*, 69:101894, 2021. 2

[27] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proc. CVPR*, 2022. 5

[28] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, 2017. 2, 3, 6, 7

[29] Peng Tu, Yawen Huang, Feng Zheng, Zhenyu He, Liujuan Cao, and Ling Shao. Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In *Proc. AAAI*, 2022. 3

[30] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *Proc. IJCAI*, 2019. 3, 6, 7

[31] Kaiping Wang, Bo Zhan, Chen Zu, Xi Wu, Jiliu Zhou, Luping Zhou, and Yan Wang. Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. *Medical Image Anal.*, 79:102447, 2022. 1, 2

[32] Qin Wang, Wen Li, and Luc Van Gool. Semi-supervised learning by augmented distribution alignment. In *Proc. ICCV*, 2019. 2

[33] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K. Fishman, and Alan L. Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical Image Anal.*, 55:88–102, 2019. 1, 2

[34] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L. Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proc. CVPR*, 2019. 5

[35] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. In *Proc. MICCAI*, 2022. 1, 2, 6, 7

[36] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan L. Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Anal.*, 65:101766, 2020. 2, 6, 7

[37] Lingxi Xie, Qihang Yu, Yuyin Zhou, Yan Wang, Elliot K. Fishman, and Alan L. Yuille. Recurrent saliency transformation network for tiny target segmentation in abdominal ct scans. *IEEE Trans. Medical Imaging*, 39(2):514–525, 2020. 2

[38] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence H. Staib, and James S. Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Trans. Medical Imaging*, 41(9):2228–2237, 2022. 1, 2

[39] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Proc. MICCAI*, 2019. 1, 2, 6, 7

[40] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proc. ICCV*, 2019. 3, 8

[41] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 3, 8

[42] Xinkai Zhao, Chaowei Fang, De-Jun Fan, Xutao Lin, Feng Gao, and Guanbin Li. Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation. In *Proc. ISBI*, 2022. 4

[43] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot K. Fishman, and Alan L. Yuille. Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In *Proc. WACV*, 2019. 2