Supplementary Materials for "*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

A. Experimental Analysis

A.1. Ablation study for BTCV dataset

In the main paper, ablation studies are conducted on MACT dataset [6]. In the supplementary material, we further conduct ablation studies on BTCV dataset [7] with 30% labeled images to show the effectiveness of each module in our method. Similar results are obtained in MACT and BTCV datasets. Next, we follow the main paper to report the results in BTCV dataset.

A.1.1 Effectiveness of each component

In this subsection, we conduct ablation studies to show the effectiveness of each component of MagicNet on BTCV dataset. In Table 1, the first row indicates the mean-teacher baseline model [10], which our method is designed on. Compared to the baseline, our MagicNet can yield good segmentation performance. **Cross** and **In** represent our cross- and within-image partition-and-recovery, which increase the performance from 61.42% to 71.09% and 74.29%, 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 73.72% in DSC, adding only within-image partition-and-recovery branch achieves 74.29%, and adding both two branches leads to 74.80%. Finally, our proposed cube-wise pseudo label blending (short for **Bld** in Table 1) provides significant improvement to 75.53%.

A.1.2 Design choices of partition and recovery

In this subsection, 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 magic-cube partition and recovery? The comparison results are shown in the last two rows in Table 2. **Scramble** and **Keep** represent the 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 CT multi-organ segmentation. ② Should our cross-image data augmentation be operated on only unlabeled images (see **U** in Table 2) or both labeled and unlabeled images (see **LU** in Table 2)? The latter obtains a much better performance compared to the former.

A.1.3 Design of cube-wise pseudo-label blending strategy

To obtain pseudo-labels for unlabeled data, 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 3. Three supervision schemes are compared for unlabeled images based on our framework. **Teacher supervision** means the outputs of cross-image and within-image branch are both supervised by the pseudo-label from the teacher model. **Mutual supervision** means the outputs of cross-image and within-image are mutually supervised. It can be seen that our blending strategy works favorably for unlabeled data.

Methods	Cross	In	Loc	Bld	Avg. DSC	Spl	R.kid	L.kid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	RG	LG
Baseline (MT [10])					61.42	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
Cross	✓				71.09	89.13	80.83	81.24	51.95	55.09	93.97	64.94	90.16	84.71	69.95	60.58	58.29	43.28
Cross + In	✓	✓			74.29	89.80	84.74	85.88	56.69	62.19	93.85	66.16	90.42	84.51	71.33	64.14	58.80	57.24
Cross + Loc	✓		✓		73.72	88.05	83.14	84.15	62.67	63.34	93.09	66.55	90.63	82.58	68.89	60.79	55.81	58.70
Cross + In + Loc	✓	✓	✓		74.80	90.64	85.76	86.49	61.36	63.02	94.93	67.38	90.51	82.46	70.90	63.95	56.80	58.15
Cross + In + Loc + Bld	✓	✓	✓	✓	75.53	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

Table 1. Ablation study (DSC, %) for the effectiveness of each component of MagicNet on BTCV dataset. **Cross**: cross-image partition-and-recovery. **In**: Within-image partition-and-recovery. **Bld**: cube-wise pseudo-label blending. **Loc**: magic-cube location reasoning. 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.

①	②	DSC	NSD
scramble	U	62.90 ± 4.63	63.03 ± 5.95
keep	U	64.92 ± 5.53	64.61 ± 6.96
scramble	LU	67.09 ± 4.74	67.56 ± 5.62
keep	LU	71.09 ± 4.54	71.20 ± 5.91

Table 2. Ablation of design choices for cross-image partition and recovery on BTCV dataset (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.	69.81 ± 4.27	70.56 ± 4.69
mutual sup.	66.46 ± 5.30	67.97 ± 6.94
blending	75.53 ± 4.90	76.31 ± 6.30

Table 3. Comparison of different pseudo-label supervision/blending strategies for unlabeled data on BTCV dataset. **Sup.**: supervision. **Blending**: our cube-wise pseudo-label blending.

	DSC	NSD
CutMix [14]	68.83 ± 5.10	66.87 ± 7.50
CutOut [4]	69.54 ± 4.34	68.12 ± 5.87
MixUp [15]	68.59 ± 5.68	67.77 ± 7.44
Ours (2)	74.80 ± 4.84	75.09 ± 5.98
Ours (3)	75.53 ± 4.90	76.31 ± 6.30

Table 4. Comparison of different data augmentation methods on BTCV dataset, where we try different CutMix and CutOut sizes, and choose the best results. For MagicNet, we compare different N values in $(\cdot)$.

A.1.4 Comparison with interpolated-based methods

As shown in Table 4, our augmentation method outperforms other methods such as CutMix, CutOut and MixUp.

A.1.5 Different number (N) of small-cubes

We study the impact of different numbers of small-cubes N, as shown in Table 4. 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 $N = 3$.

B. Other Analysis

B.1. Visualization Results

We visualize some qualitative results from BTCV dataset [7] in Fig. 1 for different semi-supervised medical image segmentation methods, including MT [10], UA-MT [13], CPS [3], SS-Net [12], our MagicNet and ground-truth. We can observe that our method achieves more accurate segmentation results compared with previous methods.

B.2. Pseudo-label Quality

In this subsection, we validate the effectiveness of MagicNet for refining the quality of pseudo-label. To this end, we calculate DSC scores between pseudo-masks and ground-truth of unlabeled training images for our method and other methods. In Fig. 2, we show DSCs of each organ and Avg. DSC of unlabeled training set of BTCV dataset. It is observed that our method can well improve DSC performance for each organ-class, especially for the small organ-class, *e.g.*, esophagus, veins, right/left adrenal glands, pancreas and gallbladder.

B.3. Comparison with Rubik-Cube Style Self-Supervised Learning

To show that our MagicNet has a better ability in leveraging unlabeled data than state-of-the-art self-supervised learning methods, we compare MagicNet with Swin UNETR [8] and a rubik-cube based method Rubik++ [9]. We use the TotalSegmentator dataset [11] (with 1,204 CT scans) as unlabeled training set, and pick 18 images in BTCV dataset [7] as labeled training set and treat the remaining 12 images as the testing set (18 and 12 images are chosen following [1,2,5]). In the above

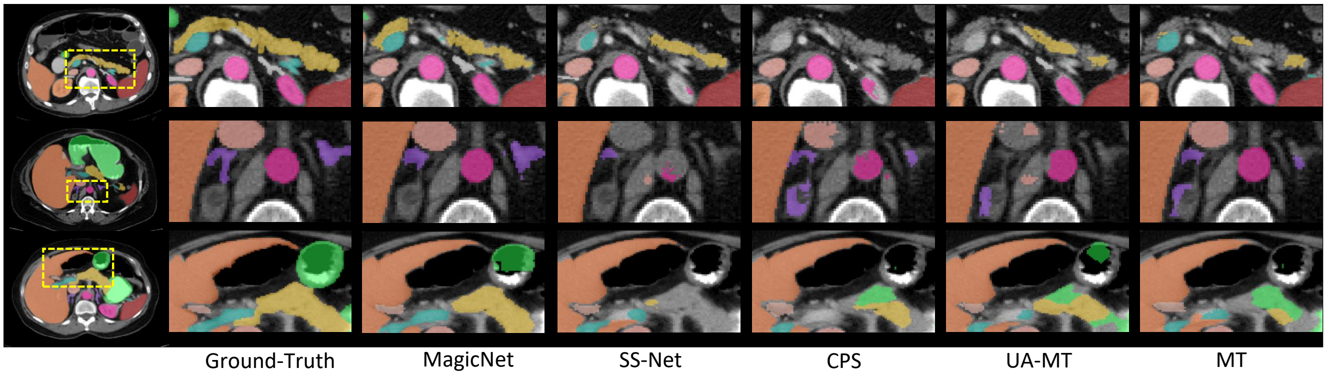

Figure 1. Qualitative visualizations of the proposed MagicNet and state-of-the-art methods on BTCV dataset.

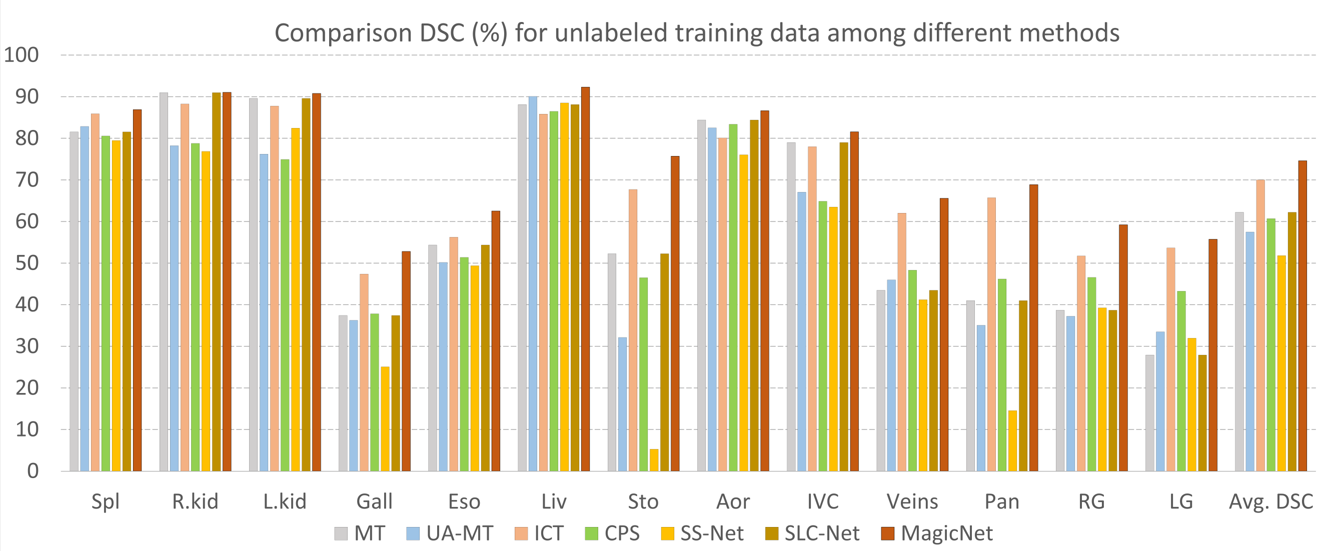

Figure 2. DSC comparison for multi-organ categories of unlabeled data among different approaches on BTCV dataset.

experimental setting, MagicNet, Swin UNETR [8], Rubik++ [9] achieve 79.42, 76.91, 75.90 in DSC with the backbone VNet, respectively. This shows the superiority of our method.

References

[1] 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. 2

[2] 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. 2

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proc. CVPR*, 2021. 2

[4] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[5] 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. 2

[6] 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. 1

[7] 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. 1, 2

[8] 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. 2, 3

[9] Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: Self-supervised learning with volume-wise transformation for 3d medical image segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Proc. MICCAI*, volume 12264 of *Lecture Notes in Computer Science*, pages 238–248. Springer, 2020. 2, 3

[10] 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. 1, 2

[11] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images, 2022. 2

[12] 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. 2

[13] 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. 2

[14] 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. 2

[15] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018. 2