Supplementary Material for "Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation"

Yunhao Bai[1] Duowen Chen[1] Qingli Li[1] Wei Shen[2] 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
{yhbai@stu., duowen_chen@stu., qlli@cs., ywang@cee.}ecnu.edu.cn, wei.shen@sjtu.edu.cn

A. Ablation studies on ACDC dataset

Following SSNet [2], 2D U-Net is used as the backbone and the 2D slices from the original 3D volume data are used as inputs.

A.1. Size of Zero-value Region in the Mask

We set $\beta = \{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}\}$ for the zero-value region $\beta H \times \beta W$ in the mask $\mathcal{M}$ on ACDC dataset to see how the performance changes. Results are shown in Table 1. Although our method outperforms SOTA when β is set as $\frac{2}{3}$, we can get a better result when $\beta = \frac{1}{2}$. For LA dataset, which is trained with 3D volume as input, the best results are achieved when $\beta = \frac{2}{3}$. Results are not very sensitive when $\beta \in [\frac{1}{3}, \frac{2}{3}]$. The best β is slightly different for these two datasets. From the perspective of zero-value region ratio γ of mask $\mathcal{M}$, the two datasets achieved the best performance at similar γ (*i.e.*, $\gamma = \frac{8}{27}$ for LA dataset when $\beta = \frac{2}{3}$ and $\gamma = \frac{1}{4}$ for ACDC dataset when $\beta = \frac{1}{2}$). It is worth mentioning that for $\beta = \frac{1}{2}$, our method also outperforms SOTA methods on LA dataset.

β	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
1/3	3(5%)	67(95%)	87.79	78.98	4.75	1.29
1/2			**89.37**	**81.40**	**1.40**	**0.43**
2/3			87.59	78.67	1.90	0.67
5/6			81.63	70.29	9.37	2.91
1/3	7(10%)	63(90%)	89.28	81.23	2.27	0.81
1/2			**89.83**	**82.13**	1.84	**0.52**
2/3			88.84	80.62	3.98	1.17
5/6			87.95	79.23	**1.68**	0.65

Table 1. Ablation study of β on ACDC dataset. We use $\beta = 2/3$ as the default value for all experiments.

A.2. Teacher Network Initialization Strategy

Like LA dataset, we also perform ablation study of teacher network initialization strategies on ACDC dataset. As shown in Table 2, the performance is not sensitive when the teacher network is initialized from a pre-trained model, no matter whether the labeled data is copy-pasted. When the labeled data is scarce and the teacher network is initialized randomly, the performance drops a lot (*i.e.*, 82.33% Dice score for random initialization on ACDC dataset with 5% labeled data).

*Corresponding Author.

Strategy	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
random	3(5%)	67(95%)	82.33	72.76	9.78	4.74
w/o CP			**87.60**	78.60	**1.74**	0.71
w/ CP			87.59	**78.67**	1.90	**0.67**
random	7(10%)	63(90%)	88.66	80.33	3.75	1.23
w/o CP			**89.19**	**81.15**	6.11	1.64
w/ CP			88.84	80.62	**3.98**	**1.17**

Table 2. Ablation study of pre-training strategy on ACDC dataset. random: Initialized randomly. w/o CP: Initialized from a pre-trained model trained on labeled data without copy-paste. w/ CP: Initialized from a pre-trained model trained on labeled data with copy-paste. We use w/ CP as the default strategy for all experiments.

A.3. Design Choices of Masking Strategies

Table 3 shows the results of different masking strategies' impact on ACDC dataset. Note that for *Random*, we randomly sample 36 small $\beta H \times \beta W$ zero-value cubes in an all-one mask $\mathcal{M}$, where β is set as 1/9. And for *Contact*, the shape of zero-value region in mask $\mathcal{M}$ is $\beta H \times W$, where β is 4/9. The different settings of β in *Random* and *Contact* are to maintain the same number of zero-value voxels. Due to the complex differences between LA dataset and ACDC dataset, *e.g.*, different data dimension used in training (*i.e.*, 3D volume and 2D slice), varying target organ and the number of target organs (*i.e.*, atrial vs. right ventricle, left ventricle and myocardium), contact mask strategy performs the best on ACDC datset.

Mode	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
Random	3(5%)	67(95%)	86.29	76.84	4.11	1.27
Contact			**88.31**	**79.70**	2.56	0.84
Context			87.59	78.67	**1.90**	**0.67**
Random	7(10%)	63(90%)	87.53	78.64	4.30	1.34
Contact			**89.22**	**81.14**	**3.64**	**1.10**
Context			88.84	80.62	3.98	1.17

Table 3. Results with three masking strategies on ACDC dataset. We use context as the default strategy for all experiments.

Strategy	Scans used		Metrics			
	Labeled	Unlabeled	Dice↑	Jaccard↑	95HD↓	ASD↓
Mixup	3(5%)	67(95%)	67.27	56.26	10.71	3.05
FG-CutMix			84.02	74.22	6.66	1.92
Ours			**87.59**	**78.67**	**1.90**	**0.67**
Mixup	7(10%)	63(90%)	81.68	70.44	6.24	2.24
FG-CutMix			86.48	76.98	**3.76**	**1.12**
Ours			**88.84**	**80.62**	3.98	1.17

Table 4. Ablation study of interpolation strategies on ACDC dataset.

B. Interpolation Strategies

To demonstrate the advantages of BCP, we compare it with two other interpolation strategies on LA dataset in Sec 4.5 in the main paper. Here we show the results on ACDC dataset in Table 4, and give more detailed descriptions.

B.1. Mixup Interpolation

Inspired by GuidedMix-Net [1], we follow the design of its framework under our semi-supervised medical image segmentation setting. The designed pipeline is shown in Fig 1. In real implementation, we set batch size as four for fair comparison. For simplicity, let's take batch size as two for example, which contains one labeled image $\mathbf{X}^l$ and one unlabeled image $\mathbf{X}^u$. We combine the two images in a Mixup [3] manner and obtain a new image: $\mathbf{X}^{mix} = \lambda \mathbf{X}^l + (1-\lambda)\mathbf{X}^u$, where $\lambda \in (0, 1)$

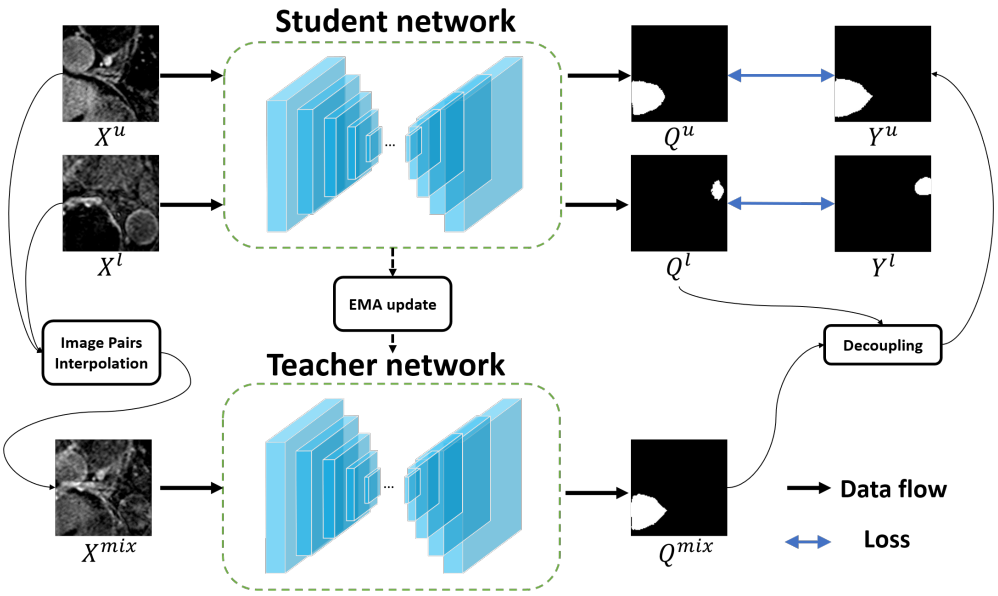

Figure 1. Overview of our Mixup-interpolation ablation study. Teacher network takes a labeled-unlabeled mixed image as input, and its prediction is used as pseudo-label after decoupling [1].

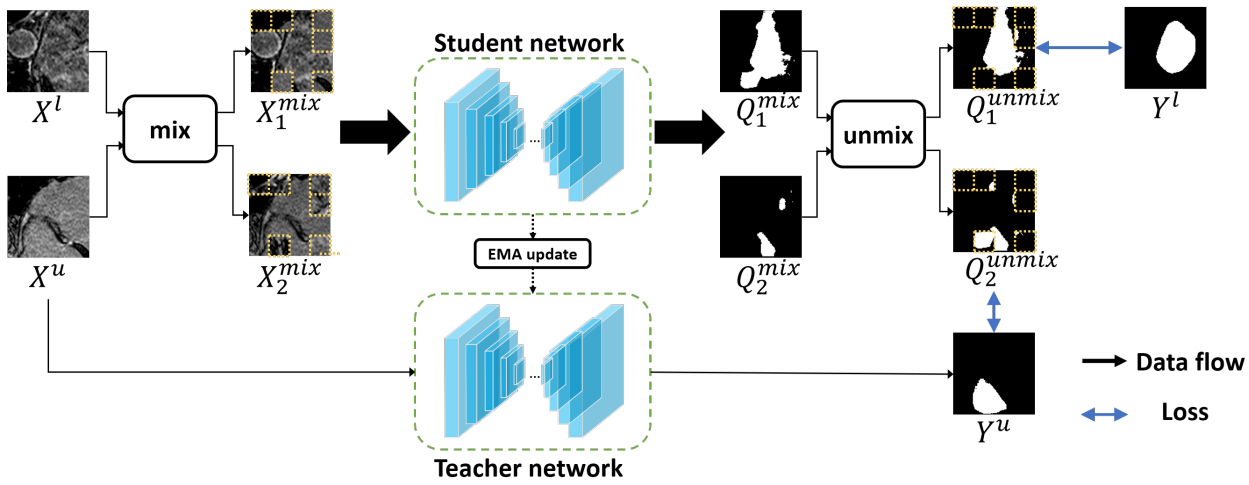

Figure 2. Overview of our FG-CutMix ablation study. Each image in a batch is cropped into 4×4 patches. Then they are copy-pasted to generate new images while keeping relative position. The predictions of the mixed images are reconstructed according to their original coordinates and supervised by ground-truth and pseudo-label. Teacher network is used for generating pseudo-labels.

is a hyper-parameter sampled from the *Beta* (α, α). $\mathbf{X}^l$ and $\mathbf{X}^u$ are fed into the Student network to get their corresponding predictions $\mathbf{Q}^l$ and $\mathbf{Q}^u$. $\mathbf{X}^{mix}$ is sent to Teacher network and whose prediction is defined as $\mathbf{Q}^{mix}$. $\mathbf{Q}^l$ is directly supervised by ground-truth $\mathbf{Y}^l$. To acquire the supervision of unlabeled prediction, we use $\mathbf{Q}^l$ to decouple $\mathbf{Q}^{mix}$ which ends up with $\mathbf{Y}^u$: $\mathbf{Y}^u = \mathbf{Q}^{mix} - \lambda \mathbf{Q}^l$. $\mathbf{Y}^u$ is adopted as the supervision of $\mathbf{Q}^u$. We utilize the same loss functions as used in BCP for fair comparison.

In medical images, target regions usually have lower contrast compared with targets in natural images. Mixup operations make the pixels to be superimposed over one another. Thus, target regions with low contrast will be further hidden behind, which makes the network hard to distinguish the targets from the background in mixed images. As an example shown in Fig. 1, $\mathbf{Q}^{mix}$ fails to predict the targets in the labeled data (note that the visualization is based on the well-trained model).

B.2. FG-CutMix Interpolation

We design this ablation study to explore whether a more fine-grained shuffling between labeled data and unlabeled data is better than BCP.

Let's take the batch size as two for example, containing one labeled image $\mathbf{X}^l$ and one unlabeled image $\mathbf{X}^u$. First we crop each image into 4×4 patches and randomly shuffle them to generate new images $\mathbf{X}_1^{mix}$ and $\mathbf{X}_2^{mix}$, meanwhile we keep the relative position of the patches. This step is called *mix* as shown in Fig 2. $\mathbf{X}_1^{mix}$ and $\mathbf{X}_2^{mix}$ are fed into Student network whose predictions are $\mathbf{Q}_1^{mix}$ and $\mathbf{Q}_2^{mix}$, respectively. Then we reconstructed $\mathbf{Q}_1^{mix}$ and $\mathbf{Q}_2^{mix}$ into $\mathbf{Q}_1^{unmix}$ and $\mathbf{Q}_2^{unmix}$ according to their coordinates in original images, as *unmix* shows in the figure. For visualization purpose, we denote the patches which are from other image/prediction after *mix* and *unmix* by yellow dot boxes in the figure. The unmixed predictions are supervised by the ground-truth and the pseudo-label.

$\mathbf{X}_1^{mix}$ and $\mathbf{X}_2^{mix}$ introduces more noises due to a more fine-grained copy-paste operation than our BCP, which may sabotage the predicted results. As an example shown in Fig. 2, $\mathbf{Q}_2^{unmix}$ suffers from severe noise compared with $\mathbf{Y}^u$). This further highlights the advantages of BCP.

B.3. Summary of Interpolation Strategies

The interpolation strategies used in semi-supervised medical image segmentation can be summarized as follows:

- Interpolation strategies such as Mixup superimposed two images, which may not be suitable for medical images. The weighted sum in pixel-level could further make the contrast of the target lower and thus increasing difficulty of training the network.
- A fine-grained copy-paste method brings more non-negligible noise in medical images, which influence the accuracy of the prediction.
- Our proposed BCP serves as a strong data augmentation tool to avoid aforementioned problems while mitigating the empirical distribution mismatch.

Note that the shortcomings of other interpolation strategies may be remedied by combining with other modules. Since BCP is simple and clean, we only explore their frameworks with the same degree of conciseness of BCP.

C. Reproducibility

Our code is modified from CoraNet [†] for Pancreas-NIH dataset, SSNet [‡] for LA dataset and ACDC dataset. Our code will be published after the paper is accepted.

References

[1] 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. 2, 3

[2] Yicheng Wu, Zhonghua Wu, Qianyi Wu, Zongyuan Ge, and Jianfei Cai. Exploring smoothness and class-separation for semi-supervised medical image segmentation. *CoRR*, abs/2203.01324, 2022. 1

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

[†] https://github.com/koncle/CoraNet

[‡] https://github.com/ycwu1997/SS-Net