# M$^3$HL: Mutual Mask Mix with High-Low Level Feature Consistency for Semi-Supervised Medical Image Segmentation

Yajun Liu[1], Zenghui Zhang[1*], Jiang Yue[2], Weiwei Guo[3], and Dongying Li[1]

[1] Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai, China
{liuyajun,zenghui.zhang,dongying.li}@sjtu.edu.cn

[2] Department of Endocrinology and Metabolism, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China
rjnfm3083@163.com

[3] Center for Digital Innovation, Tongji University, Shanghai, China
weiweiguo@tongji.edu.cn

**Abstract.** Data augmentation methods inspired by CutMix have demonstrated significant potential in recent semi-supervised medical image segmentation tasks. However, these approaches often apply CutMix operations in a rigid and inflexible manner, while paying insufficient attention to feature-level consistency constraints. In this paper, we propose a novel method called **M**utual **M**ask **M**ix with **H**igh-**L**ow level feature consistency (**M$^3$HL**) to address the aforementioned challenges, which consists of two key components: 1) M$^3$: An enhanced data augmentation operation inspired by the masking strategy from Masked Image Modeling (MIM), which advances conventional CutMix through dynamically adjustable masks to generate spatially complementary image pairs for collaborative training, thereby enabling effective information fusion between labeled and unlabeled images. 2) HL: A hierarchical consistency regularization framework that enforces high-level and low-level feature consistency between unlabeled and mixed images, enabling the model to better capture discriminative feature representations. Our method achieves state-of-the-art performance on widely adopted medical image segmentation benchmarks including the ACDC and LA datasets. Source code is available at https://github.com/PHPJava666/M3HL.

**Keywords:** Semi-supervise learning · Medical image segmentation · Mutual mask mix · Feature consistency constraints.

## 1 Introduction

Semi-supervised medical image segmentation (SSMIS) aims to achieve performance comparable to fully supervised methods while utilizing only limited annotated data, effectively alleviating the challenges of scarce labeled data and labor-

[*] Corresponding author

intensive annotation processes in medical imaging, which holds significant implications for computer-aided diagnosis and clinical applications. Currently, one effective category of SSMIS methods is based on consistency regularization[8,15,16,19]. These methods enforce consistency constraints to ensure performance stability across different input views, which are typically generated through diverse data augmentation strategies[1,5] or different network initialization approaches[7,11].

Among data augmentation based SSMIS methods, BCP[1] breaks the paradigm of training labeled and unlabeled data separately, inspired by CutMix[20] in semi-supervised learning, by generating new training samples through bidirectional copy-pasting of co-located image patches. PSC[5] extends BCP's approach by splitting paired labeled/unlabeled images into equal-sized patches and randomly shuffling them to create mixed samples. OMF[9] crops foreground and background regions along segmentation edges using label guidance and swaps them across images to synthesize hybrid samples. ABD[4] enhances segmentation in low-confidence regions via confidence-guided bidirectional replacement of image patches between strongly and weakly augmented inputs. These methods, by locally mixing or perturbing labeled and unlabeled data, break the conventional paradigm of independent training, promoting cross-distribution information interaction. However, these methods rely on rigid and inflexible data mixing strategies (e.g., fixed patch sizes, predefined replacement rules), limiting their adaptability to complex anatomical variations. Moreover, they do not incorporate feature-level consistency constraints, which can hinder the effective capture of high-level semantic information, potentially causing error propagation due to local noise, and ultimately restricting the model's ability to capture subtle pathological features.

In this work, inspired by the Masked Image Modeling (MIM)[6] paradigm in visual representation learning, we propose a dynamic mutual mask mix strategy to refine existing data augmentation frameworks, incorporating high-low level consistency constraints that enable simultaneous attention to global contextual patterns and localized structural details, thereby improving semantic alignment and feature robustness. Specifically, we first devise a random mask generator that parametrically controls mask patch sizes and mask ratios, enabling dynamic and random mutual mask mixing between labeled and unlabeled data. This dynamic mixing mechanism systematically explores the impact of diverse spatial-contextual combinations on feature learning, compelling the model to develop a more comprehensive understanding of anatomical structures through alternating occlusion and recombination strategies. Furthermore, we introduce high-low level feature consistency constraints: at the low-level feature space, we enforce geometric consistency of local edge features by constructing multi-view L1 norm constraints between the mixed samples and the unlabeled samples; at the high-level feature space, we design a symmetric cosine similarity metric, constraining the directional consistency of mixed and unlabeled samples' features in the semantic space from multiple dimensions. This design effectively filters out outlier noise in pseudo-labels through hierarchical feature calibration, enhancing the model's feature discriminability in complex scenarios, such as occlusion and boundary blurring.

In summary, the main contributions of this work are as follows: (1) We introduce a novel dynamic mutual mask mixing (M$^3$) strategy, that enhances semi-supervised medical image segmentation through a random mask generator with adjustable mask patch sizes and ratios, addressing the limitations of rigid data mixing in existing methods. (2) We propose a hierarchical high-low level feature consistency framework (HL), significantly improving the model's ability to capture both global contextual patterns and localized structural details while mitigating pseudo-label noise. (3) We achieve state-of-the-art performance on the ACDC[2] and LA[18] datasets, demonstrating the efficacy of our M$^3$HL method in handling scarce labeled data and complex anatomical variations.

# 2 Method

## 2.1 Problem Setting and Overall Architecture

In our semi-supervised segmentation task, we use a labeled dataset $\mathcal{D}_l$ with $N$ labeled samples and an unlabeled dataset $\mathcal{D}_u$ with $M$ unlabeled samples, where $X_l$ and $Y_l$ represent the labeled image and its corresponding labels, respectively. Notably, the number of unlabeled samples $M$ significantly exceeds the number of labeled samples $N$.

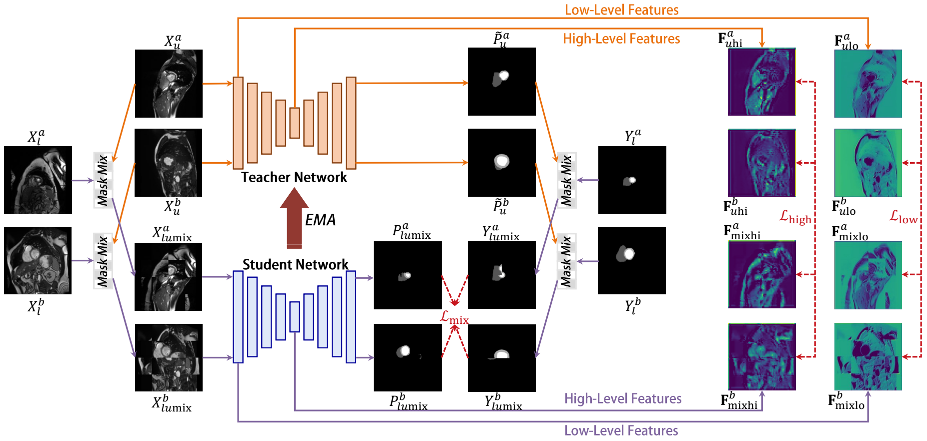

Fig. 1: Overview of our mutual mask mix with high-low level feature consistency method.

The architecture follows a teacher-student paradigm, as shown in Fig. 1. Each batch of labeled images is divided into two parts, $X^a_l$ and $X^b_l$, and unlabeled images are similarly split into $X^a_u$ and $X^b_u$. After passing $X^a_l$ and $X^a_u$ through our mutual mask mix module, we obtain mixed images $X^a_{l\text{umix}}$, and similarly for $X^b_l$ and $X^b_u$ to get $X^b_{l\text{umix}}$. These mixed images are input into the student network, while the unlabeled images go into the teacher network. The teacher network's parameters are updated using an exponential moving average (EMA) of the student network.

The student network generates predictions $P^a_{l\text{umix}}$ and $P^b_{l\text{umix}}$ for the mixed images, while the teacher network generates pseudo-labels $\tilde{P}^a_u$ and $\tilde{P}^b_u$ for the unlabeled images. These pseudo-labels are mixed with labeled data to generate the mixed outputs $Y^a_{l\text{umix}}$ and $Y^b_{l\text{umix}}$, which are used in the mutual mask mix loss function $\mathcal{L}_{\text{mix}}(P^a_{l\text{umix}}, P^b_{l\text{umix}}, Y^a_{l\text{umix}}, Y^b_{l\text{umix}})$.

The teacher network extracts high-level and low-level features from $X^a_u$ and $X^b_u$, denoted as $\mathbf{F}^a_{\text{uhi}}$, $\mathbf{F}^a_{\text{ulo}}$, $\mathbf{F}^b_{\text{uhi}}$, and $\mathbf{F}^b_{\text{ulo}}$, respectively. Similarly, the student network extracts high-level and low-level features from $X^a_{l\text{umix}}$ and $X^b_{l\text{umix}}$, denoted as $\mathbf{F}^a_{\text{uhi}}$, $\mathbf{F}^a_{\text{ulo}}$ , $\mathbf{F}^b_{\text{uhi}}$, and $\mathbf{F}^b_{\text{ulo}}$, respectively. The high-level and low-level feature consistency losses are defined as $\mathcal{L}_{\text{high}}(\mathbf{F}^a_{\text{uhi}}, \mathbf{F}^b_{\text{uhi}}, \mathbf{F}^a_{\text{mixhi}}, \mathbf{F}^b_{\text{mixhi}})$ and $\mathcal{L}_{\text{low}}(\mathbf{F}^a_{\text{ulo}}, \mathbf{F}^b_{\text{ulo}}, \mathbf{F}^a_{\text{mixlo}}, \mathbf{F}^b_{\text{mixlo}})$, respectively, the high-low level feature consistency loss is the sum of both, expressed as $\mathcal{L}_{\text{HL}} = \mathcal{L}_{\text{high}} + \mathcal{L}_{\text{low}}$.

The overall loss function for the training process is composed of the mask mix loss and the weighted high-low level feature consistency loss, expressed as:

$$\mathcal{L} = \mathcal{L}_{\text{mix}} + \lambda \mathcal{L}_{\text{HL}} \tag{1}$$

where the hyperparameter $\lambda$ controls the strength of the high-low level feature consistency constraint. The following section will describe loss functions $\mathcal{L}_{\text{mix}}$ and $\mathcal{L}_{\text{HL}}$ in detail. It is worth noting that our method does not require a separate supervised loss $\mathcal{L}_{\text{sup}}(X_l, Y_l)$ based on labeled data to train the student network, nor does it require pretraining. We argue that pretraining on very few labeled samples may cause the model to exhibit confirmation bias, which will be verified in our results section.

### 2.2 Mutual Mask Mix

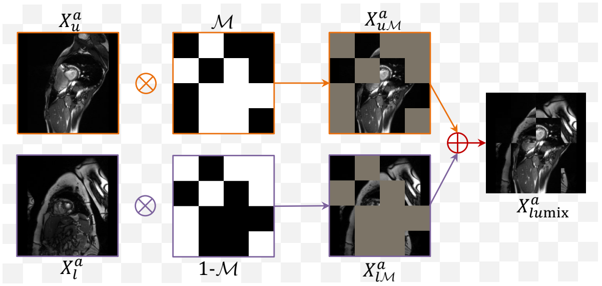

Fig. 2: Illustration of our mutual mask mix operation.

We illustrate the detailed process of the mutual mask mixing operation using the example of generating $X^a_{l u\mathrm{mix}}$ from $X^a_u$ and $X^a_l$. As shown in Fig. 2, we first generate a mask of the same size as $X^a_u$, for instance, $256 \times 256$. The size of the mask patch and the mask ratio are adjustable. In this case, we set the mask patch size to $64 \times 64$ and the mask ratio to 50%. Random mask patches are generated within the mask, denoted as $\mathcal{M}$. By multiplying $X^a_u$ with $\mathcal{M}$, we obtain the masked image $X^a_{u\mathcal{M}}$. Next, we apply the inverse mask $(1 - \mathcal{M})$ and multiply it with $X^a_l$, producing the masked image $X^a_{l\mathcal{M}}$. Finally, by adding $X^a_{u\mathcal{M}}$ and $X^a_{l\mathcal{M}}$, we obtain the mutually mask-mixed image $X^a_{l u\mathrm{mix}}$. This process can be mathematically expressed as:

$$X^a_{l u\mathrm{mix}} = X^a_{u\mathcal{M}} + X^a_{l\mathcal{M}} = X^a_u \odot \mathcal{M} + X^a_l \odot (1 - \mathcal{M}) \tag{2}$$

We adopt a similar procedure to obtain $X^b_{l u\mathrm{mix}}$, $Y^a_{l u\mathrm{mix}}$ and $Y^b_{l u\mathrm{mix}}$, construct the mutual mask mix loss as follows:

$$\mathcal{L}_{\mathrm{mix}} = \sum_{s\in\{a,b\}} \big(\mathcal{L}_{\mathrm{ce}}(P^s_{l u\mathrm{mix}}, Y^s_{l u\mathrm{mix}}) + \mathcal{L}_{\mathrm{dice}}(P^s_{l u\mathrm{mix}}, Y^s_{l u\mathrm{mix}})\big) \odot \big(\mathcal{M} + \alpha(1 - \mathcal{M})\big) \tag{3}$$

The design of $\mathcal{L}_{\mathrm{mix}}$, integrating cross-entropy loss $\mathcal{L}_{\mathrm{ce}}$ and Dice loss $\mathcal{L}_{\mathrm{dice}}$, and leveraging the dynamic mask $\mathcal{M}$ and its weighted complement $(1 - \mathcal{M})$ (controlled by the parameter $\alpha$), optimizes the student model's performance on mixed samples. This loss function enhances collaborative training between labeled and unlabeled data, strengthens the model's robustness against occlusions, noise, and incomplete data, and improves SSMIS performance through dynamic spatial-contextual exploration, particularly for handling complex anatomical structures and scarce labeled data.

### 2.3 High-Low Level Feature Consistency

The low-level features of the mixed samples, $\mathbf{F}^a_{\mathrm{mixlo}}$ and $\mathbf{F}^b_{\mathrm{mixlo}}$, obtained after the first downsampling layer in the segmentation network encoder, are constrained by quadruple L1-distance losses with the unlabeled sample's low-level features, $\mathbf{F}^a_{\mathrm{ulo}}$ and $\mathbf{F}^b_{\mathrm{ulo}}$, as shown in Eq.(4). Similarly, the high-level features of the mixed samples, $\mathbf{F}^a_{\mathrm{mixhi}}$ and $\mathbf{F}^b_{\mathrm{mixhi}}$, extracted after the bottleneck layer of the segmentation network, are constrained with the unlabeled samples' high-level features, $\mathbf{F}^a_{u\mathrm{hi}}$ and $\mathbf{F}^b_{u\mathrm{hi}}$through cosine similarity computation between each pair, as shown in Eq.(5).

$$\mathcal{L}_{\mathrm{low}} = \frac{1}{4} \sum_{s\in\{a,b\}} \sum_{t\in\{a,b\}} \|\mathbf{F}^s_{\mathrm{mixlo}} - \mathbf{F}^t_{u\mathrm{lo}}\|_1 \tag{4}$$

$$\mathcal{L}_{\mathrm{high}} = \frac{1}{4} \sum_{s\in\{a,b\}} \sum_{t\in\{a,b\}} \left[1 - \cos\left(\mathbf{F}^s_{\mathrm{mixhi}}, \mathbf{F}^t_{u\mathrm{hi}}\right)\right] \tag{5}$$

The high-low level feature consistency loss $\mathcal{L}_{\mathrm{HL}}$ is designed to enforce alignment between the mixed and unlabeled samples at both low and high feature levels. Specifically, $\mathcal{L}_{\text{low}}$ minimizes the L1-distance across all pairs of low-level features, ensuring geometric consistency of local edge details. Conversely, $\mathcal{L}_{\text{high}}$ computes the cosine similarity between high-level features to enforce semantic alignment and directional consistency in the semantic space, averaged over all pairs to enhance robustness. Together, $\mathcal{L}_{\text{HL}}$ mitigate noise and improve feature discriminability, particularly in handling complex anatomical structures and incomplete data.

## 3 Experiments and Results

### 3.1 Datasets

**ACDC dataset** - The ACDC dataset is a multi-class segmentation dataset that includes the myocardium, left and right ventricles. It consists of 100 cardiac MR imaging samples from 100 patients. We follow the data split in [12], dividing the dataset into training, validation, and test sets with a 70/10/20 ratio.
**LA dataset** - The LA dataset is a binary segmentation dataset consisting of 100 gadolinium-enhanced MR scans. For consistency, we adopt the data split strategy from [11], using 80 samples for training and 20 for validation.

### 3.2 Implementation Details and Evaluation Metrics

In our experiments, we set the parameters $\lambda$ and $\alpha$ to 0.5. We used an NVIDIA Quadro RTX 6000 GPU (24GB) with a fixed random seed. The SGD optimizer was used with a learning rate of $10^{-3}$ and a weight decay of $10^{-4}$. For the LA dataset, we employed a 3D V-Net[13] as the backbone network, with input patches randomly cropped to $112 \times 112 \times 80$, a mask patch size of $28 \times 28 \times 20$, a mask ratio of 50%, a batch size of 8, and a total of 15K training iterations. For the ACDC dataset, we used a 2D U-Net[14] for segmentation, with input patch sizes of $256 \times 256$, a mask patch size of $64 \times 64$, a mask ratio of 50%, a batch size of 24, and a total of 30K training iterations. The evaluation metrics included Dice score, Jaccard score, average surface distance (ASD), and 95% Hausdorff distance (95HD).

### 3.3 Comparison with State-of-the-Art Methods

Table 1 presents a comprehensive comparison of our proposed M$^3$HL method with eight state-of-the-art semi-supervised approaches on the ACDC (10% labeled) and LA (10% labeled) datasets. Our method consistently achieves the highest performance across all metrics. Specifically, on the ACDC dataset, M$^3$HL outperforms the latest ABD method by 0.66% in Dice score, 1.28% in Jaccard, and reduces 95HD and ASD to 1.43 and 0.34, respectively. On the LA dataset, it surpasses the top-performing OMF and AD-MT methods, achieving a Dice score of 91.01% (0.78% and 0.46% improvements over OMF and AD-MT, respectively) and an ASD of 1.59. Notably, our method demonstrates robust performance across both datasets without requiring pretraining, unlike BCP and

OMF, as this eliminates potential confirmation bias from limited labeled data. The qualitative segmentation results in Fig. 3 show that our method effectively suppresses regions of missegmentation observed in other approaches, producing segmentations closer to the ground-truth.

Table 1: Segmentation performance comparison on ACDC (10% labeled) and LA (10% labeled) datasets. * indicates that the method needs pretraining. – indicates unreported results in original papers.

| Method | ACDC (10%/7 labeled) | | | | LA (10%/8 labeled) | | | |
|---|---|---|---|---|---|---|---|---|
| | Dice↑ | Jaccard↑ | 95HD↓ | ASD↓ | Dice↑ | Jaccard↑ | 95HD↓ | ASD↓ |
| U-Net/VNet (SupOnly) | 79.41 | 68.11 | 9.35 | 2.70 | 82.74 | 71.72 | 13.35 | 3.26 |
| UA-MT [19] (MICCAI'19) | 81.65 | 70.64 | 6.88 | 2.02 | 86.28 | 76.11 | 18.71 | 4.63 |
| SASSNet [8] (MICCAI'20) | 84.50 | 74.34 | 5.42 | 1.86 | 85.22 | 75.09 | 11.18 | 2.89 |
| CPS [3] (CVPR'21) | 86.91 | 78.11 | 5.72 | 1.92 | – | – | – | – |
| DTC [11] (AAAI'21) | – | – | – | – | 87.51 | 78.17 | 8.23 | 2.36 |
| SS-Net [17] (MICCAI'22) | 86.78 | 77.67 | 6.07 | 1.40 | 88.55 | 79.62 | 7.49 | 1.90 |
| PS-MT [10] (CVPR'22) | 88.91 | 80.79 | 4.96 | 1.83 | 89.72 | 81.48 | 6.94 | 1.92 |
| BCP* [1] (CVPR'23) | 88.84 | 80.62 | 3.98 | 1.17 | 89.62 | 81.31 | 6.81 | 1.76 |
| OMF* [9](MICCAI'24) | – | – | – | – | 90.23 | 82.34 | 5.95 | 1.63 |
| AD-MT [21] (ECCV'24) | 89.46 | 81.47 | 1.51 | 0.44 | 90.55 | 82.79 | 5.81 | 1.70 |
| ABD [4](CVPR'24) | 89.81 | 81.95 | 1.46 | 0.49 | – | – | – | – |
| **M$^3$HL** (Ours) | **90.47** | **83.23** | **1.43** | **0.34** | **91.01** | **83.43** | **5.72** | **1.59** |

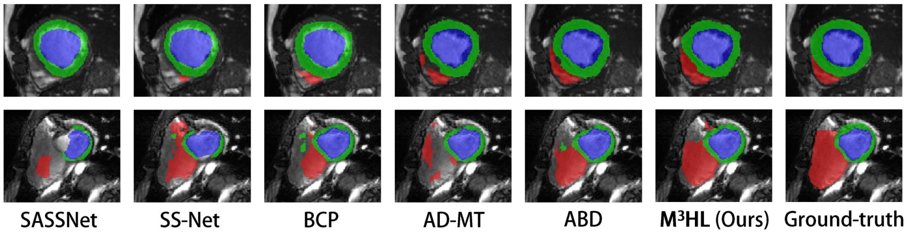

Fig. 3: Visualization of segmentation results on ACDC dataset with 10% labeled data.

### 3.4 Ablation Studies

**Effectiveness of the Proposed Losses $\mathcal{L}_{\text{mix}}$ and $\mathcal{L}_{\text{HL}}$:** As shown in Table 2, we systematically validate the effectiveness of the proposed losses $\mathcal{L}_{\text{mix}}$ and $\mathcal{L}_{\text{HL}}$ by incrementally integrating them with/without the supervised loss $\mathcal{L}_{\text{sup}}$ on the LA dataset (10% labeled data). The results reveal that either individual or combined use of $\mathcal{L}_{\text{mix}}$ and $\mathcal{L}_{\text{HL}}$ without $\mathcal{L}_{\text{sup}}$ consistently outperforms the baselines with supervised training. Specifically, introducing $\mathcal{L}_{\text{mix}}$ alone achieves significant performance gains (7.02% and 7.31% Dice score gains over VNet with/without $\mathcal{L}_{\text{sup}}$, respectively), highlighting the efficacy of our mutual mask mix strategy in fusing semantic information from labeled and unlabeled data through collaborative training. Furthermore, incorporating $\mathcal{L}_{\text{HL}}$ on top of $\mathcal{L}_{\text{mix}}$ yields additional improvements, verifying that hierarchical feature utilization enables the model to capture both global contextual and local detailed information for enhanced segmentation.

Table 2: Effectiveness of the proposed losses $\mathcal{L}_{\text{mix}}$ and $\mathcal{L}_{\text{HL}}$.

| Method | $\mathcal{L}_{\text{sup}}$ | $\mathcal{L}_{\text{mix}}$ | $\mathcal{L}_{\text{HL}}$ | Metrics | | | |
|---|---|---|---|---|---|---|---|
| | | | | Dice↑ | Jaccard↑ | ASD↓ | 95HD↓ |
| VNet | ✓ | | | 82.74 | 71.72 | 13.35 | 1.51 |
| VNet + $M^3$ | ✓ | ✓ | | 89.76 | 81.51 | 6.95 | 1.93 |
| VNet + HL | ✓ | | ✓ | 88.69 | 80.20 | 7.16 | 2.03 |
| VNet + HL + $M^3$ | ✓ | ✓ | ✓ | 90.32 | 82.48 | 7.06 | 1.68 |
| VNet + $M^3$ | | ✓ | | 90.05 | 82.39 | 7.10 | 1.82 |
| VNet + HL | | | ✓ | 89.40 | 81.43 | 7.28 | 2.02 |
| **M$^3$HL** (Ours) | | ✓ | ✓ | **91.01** | **83.43** | **5.72** | **1.59** |

**Selection of Mask Patch Size and Mask Ratio:** Fig. 4(a) and Fig. 4(b) present the heatmaps of Dice scores and ASD values under varying mask patch sizes and mask ratios on the ACDC dataset (10% labeled data). The optimal performance is achieved with a patch size of 64 and a mask ratio of 50%. This configuration equally masks identical regions in labeled and unlabeled data before mixing, allowing the model to balance feature learning from both data types and achieve optimal representation learning.

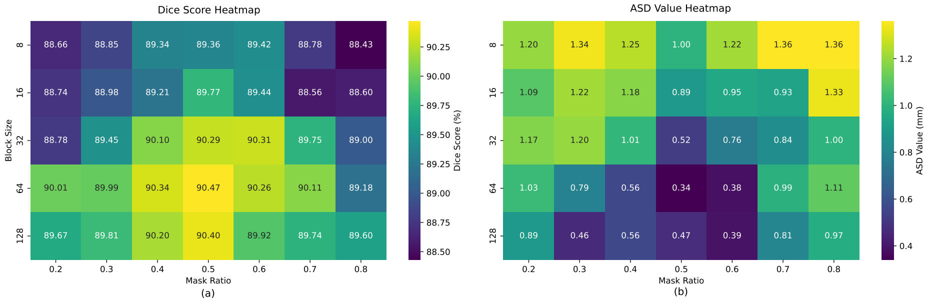

Fig. 4: Heatmaps of Dice scores and ASD values under varying mask patch sizes and mask ratios.

## 4 Conclusion

In this paper, we propose a semi-supervised medical image segmentation method based on mutual mask mix strategy and high-low level feature consistency constraints. The core idea is to enhance data by randomly masking and mutually mixing labeled and unlabeled data, generating mixed data that integrates semantic information from both sources for training. Additionally, by enforcing high-low level feature consistency constraints between mixed samples and unlabeled samples, the method more effectively captures global and local features, thereby improving segmentation performance. In future work, we plan to design more adaptive masking strategies and further explore other feature consistency approaches to address more complex scenarios.

## 5 Disclosure of Interests

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

## References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11514–11524 (June 2023)
2. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11), 2514–2525 (2018)
3. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2613–2622 (2021)
4. Chi, H., Pang, J., Zhang, B., Liu, W.: Adaptive bidirectional displacement for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4070–4080 (2024)
5. He, J., Cai, C., Li, Q., Ma, A.J.: Pair shuffle consistency for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 489–499. Springer (2024)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
7. Lei, T., Zhang, D., Du, X., Wang, X., Wan, Y., Nandi, A.K.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE transactions on medical imaging **42**(5), 1265–1277 (2022)
8. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 552–561. Springer (2020)

9. Liu, J., Qian, W., Cao, J., Liu, P.: Overlay mantle-free for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 589–598. Springer (2024)
10. Liu, Y., Tian, Y., Chen, Y., Liu, F., Belagiannis, V., Carneiro, G.: Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4258–4267 (2022)
11. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
12. Luo, X., Hu, M., Song, T., Wang, G., Zhang, S.: Semi-supervised medical image segmentation via cross teaching between cnn and transformer. In: Medical Imaging with Deep Learning (2021)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
15. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
16. Wu, Y., Ge, Z., Zhang, D., Xu, M., Zhang, L., Xia, Y., Cai, J.: Mutual consistency learning for semi-supervised medical image segmentation. Medical Image Analysis **81**, 102530 (2022)
17. Wu, Y., Wu, Z., Wu, Q., Ge, Z., Cai, J.: Exploring smoothness and class-separation for semi-supervised medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 34–43. Springer (2022)
18. Xiong, Z., Xia, Q., Hu, Z., Huang, N., Bian, C., Zheng, Y., Vesal, S., Ravikumar, N., Maier, A., Yang, X., et al.: A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. Medical image analysis **67**, 101832 (2021)
19. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2019: 22nd international conference, Shenzhen, China, October 13–17, 2019, proceedings, part II 22. pp. 605–613. Springer (2019)
20. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
21. Zhao, Z., Wang, Z., Wang, L., Yu, D., Yuan, Y., Zhou, L.: Alternate diverse teaching for semi-supervised medical image segmentation. In: European Conference on Computer Vision. pp. 227–243. Springer (2024)