# ReSeg-UNet: A Reconstruction-Guided Optimization Framework for Enhanced Medical Image Segmentation

Lin Li[1], Dong Tang[1⋆], Xiaowen Chu[2], Xiaofei Yang[1], and Fei Yu[3⋆⋆]

[1] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China
tangdong@gzhu.edu.cn

[2] The Hongkong University of Science and Technology (Guangzhou), Duxue Road 1, Guangzhou, China

[3] School of Electronics and Information Engineering, Liaoning University of Technology, Jinzhou 121001, China
yufei_hitcs@163.com

**Abstract.** Medical image segmentation is critical for accurate diagnosis; however, the task remains challenging due to the inherent ambiguities in low-contrast anatomical boundaries and the presence of extensive redundant features in the skip connections of segmentation models. To address these limitations, we propose ReSeg-UNet, a novel two-stage framework that synergizes image reconstruction with segmentation optimization. In the first stage, a composite reconstruction loss—combining Mean Squared Error (MSE) and L1 regularization—is applied to a standard segmentation network, generating stable reconstruction weights that encode multi-scale feature representations. These weights explicitly capture both global anatomical context and local boundary details. In the second stage, a three-level cross-feature alignment mechanism is introduced: the encoder of the reconstruction model is aligned with the decoder of the segmentation model, the decoder of the former is aligned with the encoder of the latter, and the intermediate features of both models are also aligned. This strategy ensures multi-level feature consistency during downsampling, intermediate layers, and upsampling, effectively mitigating information loss in blurred regions. Extensive experiments on the Synapse (abdominal CT) and ACDC (cardiac MRI) datasets demonstrate significant improvements. Our code is available at https://github.com/Li-gzhu/ReSeg-UNet.git.

**Keywords:** Reconstruction-guided Segmentation · Medical Image Segmentation · Three-Level Cross Feature Alignment.

⋆ Corresponding author: tangdong@gzhu.edu.cn
⋆⋆ Corresponding author: yufei_hitcs@163.com

## 1 Introduction

Medical image segmentation is essential for modern computer-aided diagnosis (CAD), providing pixel-level localization of anatomical structures and pathologies [2, 25, 30]. Deep learning methods, like U-Net [23], drove significant advancements in this field due to their powerful feature extraction ability. However, most approaches are built with CNN, which may struggle to model long-range dependencies because of their fixed receptive fields [22, 14]. This is a significant drawback in complex scenarios such as multi-organ CT segmentation or diffuse lesion delineation in MRI, where global anatomical relationships are crucial for accurate results. To address limitation, some researchers introduced Vision Transformers (ViTs) [7] to the medical image segmentation task, using self-attention mechanisms to capture global representation [10, 12]. For instance, TransUNet [5] combines CNNs and ViTs for joint local-global feature learning in medical image classification. While Transformers excel at modeling distant voxel interactions, their quadratic computational complexity hinders deployment on high-resolution 3D medical volumes, particularly in resource-constrained clinical environments.

Recent advancements in state space models (SSMs) [9], such as Mamba [8], leverage the core design principle of utilizing SSMs to capture global dependencies while maintaining linear computational complexity [19, 17, 32, 29, 34], making them highly effective for processing large-scale data. However, in the field of medical image segmentation, where extremely high segmentation accuracy is required, their ability to capture fine-grained local features remains somewhat limited [28, 20]. Additionally, numerous studies [27, 26, 21, 33] have demonstrated that the skip connections in U-shaped architectures merely concatenate features from corresponding encoder and decoder layers, often introducing significant feature redundancy, which subsequently adversely affects segmentation performance.

In current medical image segmentation tasks, the inherent ambiguities in low-contrast anatomical boundaries and imaging artifacts (e.g., metal artifacts in CT or motion artifacts in MRI) often lead to inadequate capture of local features by traditional methods, resulting in semantic information loss. In the domain of natural image processing, existing studies have addressed this issue by introducing cross-level knowledge transfer mechanisms (e.g., feature distillation and spatial alignment) to mitigate semantic feature loss [4, 6, 16].

Inspired by advancements in natural image processing and the aforementioned limitations in medical image segmentation, we propose a two-stage optimization framework for medical image segmentation, termed ReSeg-UNet. Specifically, in the first stage, the framework performs an image reconstruction task by inputting the ground truth into a traditional segmentation network, generating reconstruction weights that encapsulate rich semantic information. In the second stage, the segmentation task leverages these stable weights from the first stage and employs a three-level cross-feature alignment mechanism (see Fig. 1) to guide and optimize the learning of local features in the segmentation model. Through this approach, the segmentation model is able to learn finer-grained semantic features while alleviating the issue of feature redundancy introduced

by skip connections, thereby significantly improving segmentation performance. Moreover, the additional computational overhead introduced by our proposed optimization method in the segmentation task is minimal, as it is confined solely to the calculation of the feature alignment loss. This results from our decoupled training strategy, where the first and second stages are trained independently, and the reconstruction weights in the second stage are kept frozen (i.e., no parameter updates are performed).

**Contribution**: 1) We innovatively propose a two-stage network framework that leverages image reconstruction to optimize medical image segmentation. 2) We design a composite reconstruction loss function to generate stable multi-scale feature weights, and furthermore, we propose a three-level cross-feature alignment mechanism to optimize local feature learning for the segmentation task. 3)The proposed method can be seamlessly integrated into existing U-shaped frameworks (e.g., U-Net, TransUNet, Swin-UNet, VM-UNet) with minimal additional cost, delivering performance gains effectively.

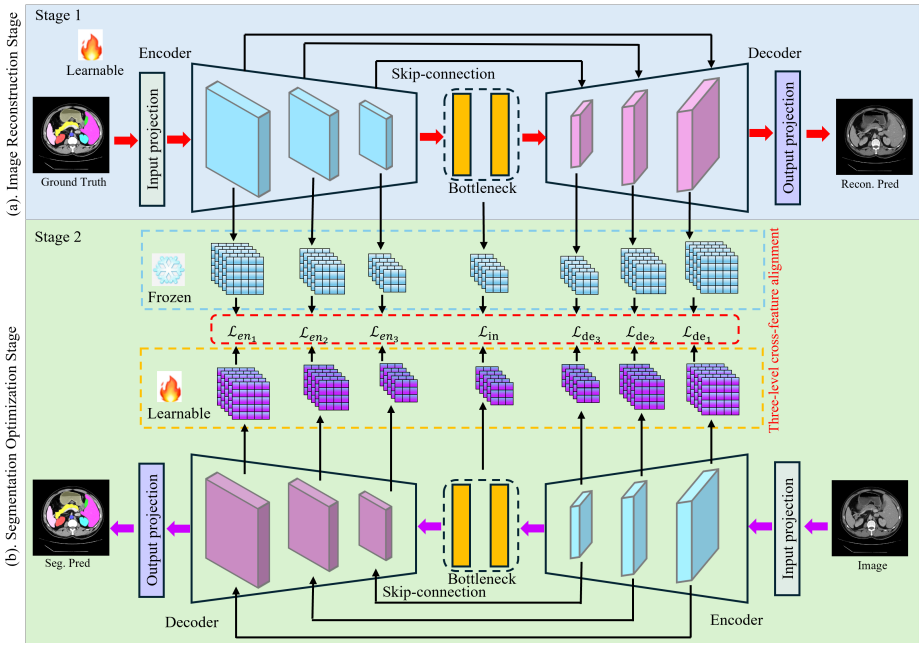


**Fig. 1.** Overview of ReSeg-UNet. (a) The ground truth is fed into a standard UNet segmentation network to perform image reconstruction, thereby obtaining stable reconstruction weights. (b) By utilizing the frozen reconstruction weights and adopting a three-level cross-feature alignment mechanism, the segmentation task is guided to learn more refined local features.

## 2 Method

### 2.1 Overview of ReSeg-UNet

ReSeg-UNet is a dual-stage optimization network tailored for medical image segmentation, as shown in Fig. 1. Current mainstream models, based on CNNs, Transformers, and Mamba, typically adopt a U-shaped structure comprising an encoder, bottleneck layer, and decoder. In the first stage, a standard segmentation network performs a reconstruction task using ground truth to generate stable, semantically rich weights. In the second stage, these weights guide the segmentation task through a three-level cross-feature alignment mechanism, enhancing local feature learning for robust and accurate segmentation. Below, we detail the ReSeg-UNet framework.

### 2.2 Stage 1: Image Reconstruction

Fig. 1(a) illustrates the detailed architecture of the image reconstruction process. First, while keeping the model architecture unchanged, we perform the image reconstruction task by using the ground truth $\hat{x} \in \mathbb{R}^{H\times W\times C}$ (where $H$, $W$, and $C$ denote the height, width, and number of channels of the image, respectively) as input, thereby obtaining stable reconstruction weights. To this end, we designed a composite loss function comprising the MSE loss and L1 regularization loss, which are employed to evaluate the global structural consistency and local detail restoration capabilities of the reconstructed image, respectively. The MSE loss is employed to quantify the pixel-level differences between the reconstructed image and the ground truth image, while the L1 loss is utilized to enhance the local detail restoration capability of the reconstructed image, particularly in preserving edge and texture features. The detailed formulas are as follows:

$$\mathcal{L}_{\mathrm{MSE}}(x,\hat{x}) = \frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}(x_{i,j,k}-\hat{x}_{i,j,k})^2 \tag{1}$$

$$\mathcal{L}_{\mathrm{L1}}(x,\hat{x}) = \frac{1}{H\times W\times C}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{C}|x_{i,j,k}-\hat{x}_{i,j,k}| \tag{2}$$

Where $x$ represents the reconstructed predicted image and $\hat{x}$ denotes the ground truth image, the final reconstruction loss is formulated as the weighted sum of the MSE loss and the L1 loss, as shown in the following equation:

$$\mathcal{L}_{\mathrm{Recon}}(x,\hat{x}) = \alpha\cdot\mathcal{L}_{\mathrm{MSE}}(x,\hat{x}) + \beta\cdot\mathcal{L}_{\mathrm{L1}}(x,\hat{x}) \tag{3}$$

Here, $\alpha$ and $\beta$ are weighting coefficients used to balance the contributions of the two loss terms, with $\alpha = 0.6$ and $\beta = 0.4$.

### 2.3 Stage 2: Segmentation Optimization

Fig. 1(b) illustrates the detailed architecture of the three-level cross-feature alignment mechanism in the second stage. This stage leverages the image reconstruction weights trained in the first stage to guide the optimization of the segmentation task. The core objective of the second stage is to achieve feature alignment between the reconstruction task and the segmentation task, ensuring that the segmentation network can fully reuse the anatomical features learned in the first stage. Specifically, the three-level cross-feature alignment mechanism consists of the following three levels: Feature alignment between the reconstruction encoder and the segmentation decoder (Recon-E $\rightarrow$ Seg-D): For each level $l$, the alignment loss is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{Recon-E}\rightarrow\text{Seg-D}}(l) = \frac{1}{H_l \times W_l \times C_l} \left\| F_{\text{Recon-E}}^{l} - F_{\text{Seg-D}}^{l} \right\|^2 \tag{4}$$

Here, $F_{\text{Recon-E}}^{l}$ denotes the feature at the $l$-th layer of the reconstruction task encoder, and $F_{\text{Seg-D}}^{l}$ represents the feature at the $l$-th layer of the segmentation task decoder. $H_l$, $W_l$, and $C_l$ denote the spatial dimensions and the number of channels of the feature maps, respectively. The alignment loss between Recon-D and Seg-E is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{Recon-D}\rightarrow\text{Seg-E}}(l) = \frac{1}{H_l \times W_l \times C_l} \left\| F_{\text{Recon-D}}^{l} - F_{\text{Seg-E}}^{l} \right\|^2 \tag{5}$$

Here, $F_{\text{Recon-D}}^{l}$ denotes the feature at the $l$-th layer of the reconstruction task decoder, and $F_{\text{Seg-E}}^{l}$ represents the feature at the $l$-th layer of the segmentation task encoder. For the bottleneck layer (intermediate features), the alignment loss is defined as:

$$\mathcal{L}_{\text{Align}}^{\text{in}} = \frac{1}{H_{\text{in}} \times W_{\text{in}} \times C_{\text{in}}} \left\| F_{\text{Recon-in}} - F_{\text{Seg-in}} \right\|^2 \tag{6}$$

Here, $F_{\text{Recon-in}}$ denotes the feature of the bottleneck layer in the reconstruction task, and $F_{\text{Seg-in}}$ represents the feature of the bottleneck layer in the segmentation task. The total feature alignment loss is the average of the aforementioned three-level alignment losses, formulated as:

$$\mathcal{L}_{\text{Align}} = \frac{1}{3} \left( \sum_{l=1}^{3} \left( \mathcal{L}_{\text{Align}}^{\text{Recon-E}\rightarrow\text{Seg-D}}(l) + \mathcal{L}_{\text{Align}}^{\text{Recon-D}\rightarrow\text{Seg-E}}(l) \right) + \mathcal{L}_{\text{Align}}^{\text{in}} \right) \tag{7}$$

**Optimization Objective Function:** The total loss function combines $\mathcal{L}_{\text{Align}}$ with the original segmentation loss function $\mathcal{L}_{\text{Seg}}$, which employs both cross-entropy and Dice loss. The formula is as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Seg}} + \lambda \cdot \mathcal{L}_{\text{Align}} \tag{8}$$

Here, $\lambda$ is the weighting coefficient for the feature alignment loss, which controls the contribution of feature alignment to the segmentation task. The value of $\lambda$ is set to 0.035.

**Table 1.** Comparison with SOTA methods on Synapse dataset. $\Delta$ denotes the improvement gain (%) by comparing with the original method.

| Model | DSC(%)↑ | HD↓ | Aorta | Gallbladder | Kidney(L) | Kidney(R) | Liver | Pancreas | Spleen | Stomach |
|---|---|---|---|---|---|---|---|---|---|---|
| U-Net [23](MICCAI'15) | 76.85 | 39.70 | 89.07 | 69.72 | 77.77 | 68.60 | 93.43 | 53.98 | 86.67 | 75.58 |
| U-Net++ [31] (MICCAI'18) | 76.91 | 36.93 | 88.19 | 68.89 | 81.76 | 75.27 | 93.01 | 58.20 | 83.44 | 70.52 |
| TransUNet [5](arxiv'21) | 77.48 | 31.69 | 87.23 | 63.13 | 81.87 | 77.02 | 94.08 | 55.86 | 85.08 | 75.62 |
| UCTransNet [27](AAAI'22) | 78.23 | 26.75 | 84.25 | 64.65 | 82.35 | 77.65 | 94.36 | 58.18 | 84.74 | 79.66 |
| Swin-UNet [3](ECCV'22) | 79.13 | 21.55 | 85.47 | 66.53 | 83.28 | 79.61 | 94.29 | 56.58 | 90.66 | 76.60 |
| HiFormer [12](WACV'23) | 80.69 | 19.14 | 87.03 | 68.61 | 84.23 | 78.37 | 94.07 | 60.77 | 90.44 | 82.03 |
| VM-UNet [24](arxiv'24) | 81.08 | 19.21 | 86.40 | 69.41 | 86.16 | 82.76 | 94.17 | 58.80 | 89.51 | 81.40 |
| H2Former [11](TMI'23) | 82.16 | 18.79 | 87.03 | 69.78 | 86.38 | 83.49 | 94.71 | 64.49 | **90.69** | 80.73 |
| MISSFormer [13](TMI'23) | 81.96 | 18.20 | 87.71 | 63.86 | 87.97 | 82.80 | 94.68 | 60.90 | 90.60 | 82.27 |
| Swin-UMamba [18] (MICCAI'24) | 82.57 | 16.21 | 88.40 | 70.41 | 87.16 | **83.76** | 95.17 | **64.80** | 89.51 | 81.40 |
| U-Net (our) | 78.86 | 27.68 | 88.60 | 71.02 | 82.24 | 70.21 | 94.78 | 62.23 | 88.60 | 73.20 |
| $\Delta$ | +2.01 | -12.02 | -0.47 | +1.30 | +4.47 | +1.61 | +1.35 | +8.25 | +1.93 | -2.38 |
| TransUNet (our) | 79.65 | 26.59 | 89.20 | 64.21 | 81.23 | 76.58 | 94.60 | 62.38 | 87.33 | 81.67 |
| $\Delta$ | +2.17 | -5.10 | +1.97 | +1.08 | -0.64 | -0.44 | +0.52 | +6.52 | +2.25 | +6.05 |
| Swin-UNet (our) | 80.58 | 18.09 | 86.36 | 67.51 | 87.96 | 79.26 | 95.26 | 60.35 | 89.66 | 78.28 |
| $\Delta$ | +1.45 | -3.46 | +0.89 | +0.98 | +4.68 | -0.35 | +0.97 | +3.77 | -1.00 | +1.68 |
| VM-UNet (our) | **83.22** | **15.73** | **89.62** | **71.21** | **88.02** | 81.36 | **95.76** | 63.36 | 90.61 | **85.82** |
| $\Delta$ | +2.14 | -3.48 | +3.22 | +1.80 | +1.86 | -1.40 | +1.59 | +4.56 | +1.10 | +4.42 |

**Table 2.** Comparison of different methods in ACDC dataset.

| Model | DSC(%)↑ | RV | Myo | LV |
|---|---|---|---|---|
| U-Net [23](MICCAI'15) | 89.57 | 85.81 | 87.47 | 95.42 |
| TransUNet [5](arxiv'21) | 89.71 | 86.67 | 87.27 | 95.18 |
| Swin-UNet [3](ECCV'22) | 90.00 | 88.55 | 85.62 | 95.73 |
| HiFormer [12](TMI'23) | 90.82 | 88.55 | 88.44 | 94.47 |
| MISSFormer [13](TMI'23) | 91.19 | 89.85 | 88.38 | 95.34 |
| VM-UNet [24](arxiv'24) | 90.56 | 88.77 | 87.89 | 95.02 |
| Swin-UMamba [18](MICCAI'24) | 91.53 | 90.01 | 88.98 | 95.60 |
| U-Net (our) | 91.21 | 88.61 | **89.73** | 95.29 |
| $\Delta$ | +1.64 | +2.80 | +2.26 | -0.13 |
| TransUNet (our) | 91.83 | **90.60** | 89.25 | 95.64 |
| $\Delta$ | +2.12 | +3.93 | +1.98 | +0.54 |
| Swin-UNet (our) | 91.24 | 89.46 | 88.59 | **95.74** |
| $\Delta$ | +1.24 | +0.91 | +2.97 | +0.01 |
| VM-UNet (our) | **91.85** | 90.30 | 89.58 | 95.67 |
| $\Delta$ | +1.29 | +1.53 | +1.69 | +0.65 |

## 3 Experiments and Results

### 3.1 Dataset

**Synapse Dataset [15]:** The Synapse dataset comprises 30 scans of eight abdominal organs: left and right kidneys, aorta, spleen, gallbladder, liver, stomach, and pancreas. It contains 3,779 clinically enhanced abdominal CT images in axial view. The dataset is divided into 18 samples for training and 12 for testing. We report Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance (HD95) as evaluation metrics for this dataset.

**ACDC Dataset [1]:** The ACDC dataset contains 100 cardiac MRI scans acquired from diverse clinical patients, with pixel-level annotations for three cardiac substructures: left ventricle (LV), right ventricle (RV), and myocardium (MYO). Following the standard experimental protocol of baseline models, the dataset was partitioned into 70 cases (1,930 axial slices) for training, 10 cases for validation, and 20 cases for testing. We evaluate our method using the DSC as the evaluation metric.

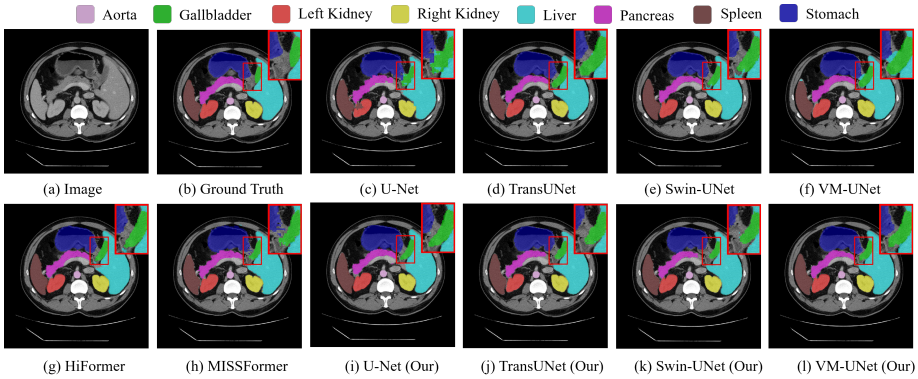


**Fig. 2.** Result visualization on Synapse dataset.

### 3.2 Experiment settings

We evaluated the effectiveness of our proposed loss function on U-Net [23], TransUNet [5], Swin-UNet [3], and VM-UNet [24] using both the Synapse and ACDC datasets. The training configurations (i.e., batch size, optimizer, learning rate, etc.) were kept consistent with those of the baseline models. For all experiments, the input image size was set to $224 \times 224$, and the data augmentation and preprocessing steps were identical to those used in the baseline models. Training was conducted on an Nvidia GTX 4090 GPU with 24GB of memory. In line with the literature, TransUNet, Swin-UNet, and VM-UNet utilized pre-trained weights from ImageNet, while U-Net was trained from scratch.

### 3.3 Results

The experimental results are presented in Table 1 for the Synapse dataset and Table 2 for the ACDC dataset. From these results, it is evident that our proposed optimization method for medical image segmentation baseline models is highly effective, delivering significant performance improvements. Specifically, on the Synapse dataset, U-Net, TransUNet, Swin-UNet, and VM-UNet achieved average Dice Similarity Coefficient (DSC) improvements of 2.01%, 2.17%, 1.45%, and 2.14%, respectively. Moreover, the Hausdorff Distance (HD) was reduced by 12.02mm, 5.10mm, 3.46mm, and 3.48mm, respectively. For the segmentation accuracy of the eight organs, most organs showed significant improvements, especially the pancreas, where the four models achieved improvements of 8.25%, 6.25%, 3.77%, and 4.65%, respectively. On the ACDC dataset, the average DSC improved by 1.64%, 2.02%, 1.24%, and 1.29%, respectively. Further comparisons reveal that the optimized VM-UNet outperformed Swin-UMamba on the Synapse dataset, with an average DSC increase of 0.65% and an HD reduction of 0.48mm. On the ACDC dataset, the optimized VM-UNet also achieved a 0.32% higher average DSC than Swin-UMamba.

The optimization effects are more evident in the segmentation visualizations. In the Synapse dataset (see Fig.2), the optimized model achieves higher segmentation accuracy than the baseline models, with outputs closer to the ground

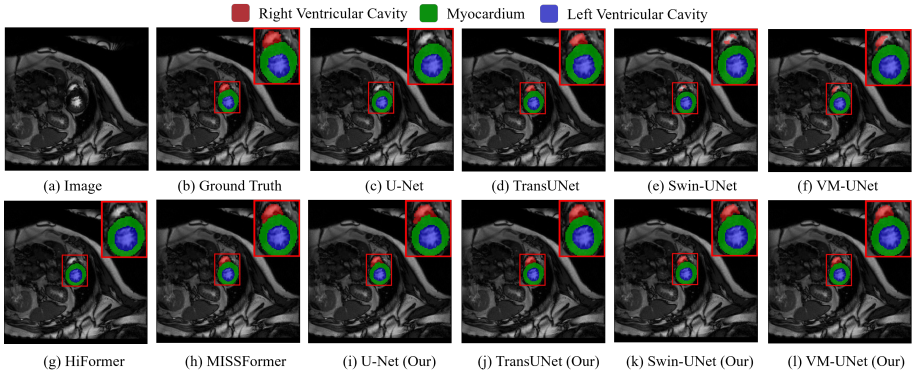


**Fig. 3.** Result visualization on ACDC dataset.

truth. In the ACDC dataset (see Fig.3), the baseline model performs poorly in right ventricle segmentation, while the optimized model significantly improves edge precision. This validates the effectiveness of our method in enhancing local feature extraction for baseline models.

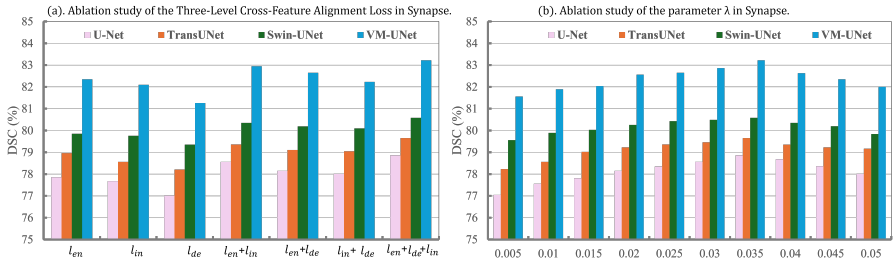


**Fig. 4.** Ablation study on the number of feature alignment loss functions and $\lambda$ is conducted based on the Synapse dataset.

### 3.4 Ablation studies

To further investigate our proposed method, we conducted a series of ablation studies on four optimized baseline models using the Synapse dataset. The results and analysis are as follows: In the ablation experiments on different loss functions (see Fig. 4(a)), the segmentation performance of the baseline models significantly improved after incorporating the three-level cross-feature alignment, achieving optimal performance when all three losses were utilized. In the ablation experiment on the feature alignment balancing factor $\lambda$ (see Fig. 4(b)), we observed that the model reached its optimal performance when $\lambda = 0.035$. This demonstrates that proper feature guidance can effectively enhance segmentation performance.

## 4 Conclusion

In this paper, we introduce ReSeg-UNet, a dual-stage optimization framework for medical image segmentation that integrates image reconstruction with segmentation. By incorporating a composite reconstruction loss and a three-level cross-feature alignment mechanism, our approach effectively enhances the performance of current mainstream 2D medical image segmentation models based on CNNs, Transformers, and Mamba architectures. Specifically, it improves the segmentation capabilities for low-contrast boundaries and small anatomical structures, such as the pancreas. Extensive experiments on the Synapse (abdominal CT) and ACDC (cardiac MRI) datasets demonstrate that ReSeg-UNet achieves significant performance improvements over baseline models and outperforms existing state-of-the-art methods. Future work will focus on extending this framework to 3D and multi-modal segmentation tasks.

**Acknowledgments.** This work was supported in part by the Doctoral Research Start-up Foundation of Liaoning University of Technology (Grant No. XB2025019)

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

## References

1. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Medical Imaging **37**(11), 2514–2525 (2018)
2. Bongratz, F., Rickmann, A., Wachinger, C.: Neural deformation fields for template-based reconstruction of cortical surfaces from MRI. Medical Image Anal. **93**, 103093 (2024)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV Workshops (3). Lecture Notes in Computer Science, vol. 13803, pp. 205–218. Springer (2022)
4. Chen, D., Mei, J., Zhang, Y., Wang, C., Wang, Z., Feng, Y., Chen, C.: Cross-layer distillation with semantic calibration. In: AAAI. pp. 7028–7036. AAAI Press (2021)
5. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. CoRR **abs/2102.04306** (2021)
6. Chung, I., Park, S., Kim, J., Kwak, N.: Feature-map-level online adversarial knowledge distillation. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 2006–2015. PMLR (2020)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR. OpenReview.net (2021)
8. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. CoRR **abs/2312.00752** (2023)
9. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. In: ICLR. OpenReview.net (2022)

10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B.A., Roth, H.R., Xu, D.: UNETR: transformers for 3d medical image segmentation. In: WACV. pp. 1748–1758. IEEE (2022)
11. He, A., Wang, K., Li, T., Du, C., Xia, S., Fu, H.: H2former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Trans. Medical Imaging **42**(9), 2763–2775 (2023)
12. Heidari, M., Kazerouni, A., Kadarvish, M.S., Azad, R., Aghdam, E.K., Cohen-Adad, J., Merhof, D.: Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: WACV. pp. 6191–6201. IEEE (2023)
13. Huang, X., Deng, Z., Li, D., Yuan, X., Fu, Y.: Missformer: An effective transformer for 2d medical image segmentation. IEEE Trans. Medical Imaging **42**(5), 1484–1494 (2023)
14. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021)
15. Landman, B., Xu, Z., Iglesias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proc. MICCAI. p. 12 (2015)
16. Li, L.: Self-regulated feature learning via teacher-free feature distillation. In: ECCV (26). Lecture Notes in Computer Science, vol. 13686, pp. 347–363. Springer (2022)
17. Liao, W., Zhu, Y., Wang, X., Pan, C., Wang, Y., Ma, L.: Lightm-unet: Mamba assists in lightweight unet for medical image segmentation. CoRR **abs/2403.05246** (2024)
18. Liu, J., Yang, H., Zhou, H., Xi, Y., Yu, L., Li, C., Liang, Y., Shi, G., Yu, Y., Zhang, S., Zheng, H., Wang, S.: Swin-umamba: Mamba-based unet with imagenet-based pretraining. In: MICCAI (9). Lecture Notes in Computer Science, vol. 15009, pp. 615–625. Springer (2024)
19. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. CoRR **abs/2401.10166** (2024)
20. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. CoRR **abs/2401.04722** (2024)
21. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. CoRR **abs/1804.03999** (2018)
22. Rahman, M.M., Munir, M., Marculescu, R.: EMCAD: efficient multi-scale convolutional attention decoding for medical image segmentation. In: CVPR. pp. 11769–11779. IEEE (2024)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (3). Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer (2015)
24. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. CoRR **abs/2402.02491** (2024)
25. Shaker, A.M., Maaz, M., Rasheed, H.A., Khan, S.H., Yang, M., Khan, F.S.: UNETR++: delving into efficient and accurate 3d medical image segmentation. IEEE Trans. Medical Imaging **43**(9), 3377–3390 (2024)
26. Sun, G., Pan, Y., Kong, W., Xu, Z., Ma, J., Racharak, T., Nguyen, L., Xin, J.: Datransunet: Integrating spatial and channel dual attention with transformer u-net for medical image segmentation. CoRR **abs/2310.12570** (2023)

27. Wang, H., Cao, P., Wang, J., Zaïane, O.R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In: AAAI. pp. 2441–2449. AAAI Press (2022)
28. Wang, Z., Zheng, J., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. CoRR **abs/2402.05079** (2024)
29. Wu, R., Liu, Y., Liang, P., Chang, Q.: Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. CoRR **abs/2403.20035** (2024)
30. Xie, Y., He, X., Yang, B., Lyu, F., Liu, S.: Cam-guided translation for unpaired weakly-supervised medical image segmentation. In: ICME. pp. 1–6. IEEE (2024)
31. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: DLMIA/ML-CDS@MICCAI. Lecture Notes in Computer Science, vol. 11045, pp. 3–11. Springer (2018)
32. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: ICML. OpenReview.net (2024)
33. Zhu, W., Chen, X., Qiu, P., Farazi, M., Sotiras, A., Razi, A., Wang, Y.: Selfreg-unet: Self-regularized unet for medical image segmentation. In: MICCAI (8). Lecture Notes in Computer Science, vol. 15008, pp. 601–611. Springer (2024)
34. Zhu, Y., Zhang, D., Lin, Y., Feng, Y., Tang, J.: Merging context clustering with visual state space models for medical image segmentation. IEEE Transactions on Medical Imaging pp. 1–1 (2025). https://doi.org/10.1109/TMI.2025.3525673