2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

SegMAE: A Dual Decoder Framework with Patch Wise Constraint for Skin Lesion Segmentation

Jiacheng Huang+, Haozhe Li+, Gexian Liu+, Gao Wang, and Keming Mao(✉)

Software College, Northeastern University, Shenyang, China[i]
maokm@mail.neu.edu.cn

Abstract. Skin lesion segmentation remains a challenging task in medical image analysis. Although Transformer-based segmentation models have achieved. notable progress in recent years, they still suffer from limitations such as the imbalance between local and global modeling, single-task architectural design, and insufficient attention to critical regions. These issues hinder their segmentation performance on complex skin lesion images. To address these challenges, we propose SegMAE, a dual-decoder segmentation framework that integrates image reconstruction and segmentation tasks to jointly enhance the model's understanding of both global context and local details. The model adopts a CNN-Transformer hybrid encoder, with a MAE decoder for reconstruction and a Cascaded Upsampler for segmentation. To enhance the model's performance and generalization, we design a two-stage training strategy that first involves. pretraining and then proceeds to hybrid multi-task training. In addition, we introduce a Patch-wise Loss function that adaptively emphasizes training on critical regions, thereby improving segmentation accuracy and robustness. Experimental results on ISIC2017, ISIC2018 and PH2 demonstrate that SegMAE consistently outperforms existing mainstream methods across multiple evaluation metrics, showcasing superior segmentation performance and strong generalization capability.

Keywords: Skin Lesion Segmentation, Patch-wise Loss, Hybrid Training Strategy, Dual Decoder Architecture.

1 Introduction

In recent years, medical image segmentation has been increasingly applied in computer-aided diagnosis, particularly in the context of skin lesion detection. Accurate segmentation of lesion regions is of vital importance for the early identification of abnormalities, as well as for improving diagnostic efficiency and accuracy. However, skin lesion images typically exhibit substantial visual complexity, characterized by intricate lesion structures, diverse morphological patterns, indistinct boundaries, and low contrast between lesions and surrounding tissues. These factors pose significant challenges to the generalization and robustness of conventional segmentation methods. Especially under

+ These authors contributed equally.

conditions of limited sample size or complex lesion distributions, a key research challenge is how to design a segmentation model that effectively combines local feature perception with global semantic understanding.

In the domain of medical image segmentation, convolutional neural networks (CNNs) have long served as the mainstream solution. Models like U-Net [1], UNet++ [2], and Att U-Net [3] adopt encoder-decoder structures with skip connections to preserve spatial information and deliver reliable performance in segmenting lesions. However, due to their reliance on local receptive fields, CNNs often struggle to capture global semantic relationships, which limits their effectiveness in complex medical imaging scenarios. In recent years, Vision Transformers (ViT) [4] have demonstrated remarkable potential in computer vision by effectively modeling long-range dependencies and capturing global semantic context via self-attention mechanisms. Building on this, models such as SegFormer [5], Swin-UNet [6], and TransUNet [7] have extended ViT to medical image segmentation. For instance, TransUNet embeds ViT into the. bottleneck of a U-Net structure to capture global features, while Swin-UNet introduces local window-based attention to balance global modeling and spatial detail preservation. These methods significantly broaden the capacity of segmentation networks to handle complex medical imaging tasks.

Additionally, researchers have begun exploring multi-task learning frameworks by introducing auxiliary tasks to enhance the expression ability and robustness of segmentation models. Current methods have attempted to integrate tasks such as deep supervision, boundary regression, feature contrastive learning, and image classification into segmentation pipelines. By increasing training signals and diversifying optimization objectives, these frameworks guide networks to learn richer and more discriminative feature representations. Overall, this direction is gaining increasing attention, providing new avenues for improving medical image segmentation performance.

Despite the advances made by existing methods, there remain significant limitations in the segmentation of skin lesion images:

Imbalanced local-global modeling: Current models predominantly focus either on local textures or on global context modeling, struggling to integrate multi-scale information effectively. This imbalance leads to insufficient accuracy in handling lesions characterized by indistinct boundaries and varying scales.

Single-task structural design: Mainstream segmentation methods typically. concentrate on the segmentation task, lacking structural designs that incorporate auxiliary tasks to enhance semantic perception. This deficiency prevents effective task synergy and limits deep understanding and expressive capability regarding skin lesion regions.

Insufficient weighting in loss functions: Conventional segmentation losses (e.g., Dice and BCE) apply uniform weights to all pixels, neglecting focused training on challenging areas such as boundaries and low-contrast regions. This oversight reduces the model's discriminative power and overall segmentation accuracy in critical regions.

To address the aforementioned challenges, we propose SegMAE, a novel framework tailored specifically for skin lesion segmentation tasks. Firstly, to enhance the representation of local textures and structural details, we design a hybrid CNN-Transformer encoder by introducing a CNN module preceding the ViT encoder, which extracts mid-to-high-resolution local features. Unlike ViT, which excels at global context modeling,

CNN effectively captures subtle structural information inherent in skin lesion images. By integrating the strengths of both local and global perception, our proposed architecture significantly improves segmentation performance, especially for complex lesion shapes. Secondly, we introduce a dual-decoder architecture to overcome the limitations of traditional single-task segmentation models. The segmentation branch utilizes a Cascaded Upsampler to gradually restore spatial resolution, generating. accurate segmentation masks. Concurrently, the reconstruction branch employs an MAE decoder, responsible for reconstructing masked image patches. We adopt a pretraining-guided hybrid training strategy, where the reconstruction branch first undergoes an initial pretraining phase to enhance the model's understanding of skin lesion structures and semantics. Subsequently, both segmentation and reconstruction tasks are trained jointly, facilitating cross-task knowledge transfer and mutual optimization through backpropagation, thereby boosting segmentation performance and generalization capabilities. Finally, to address the insufficient focus on challenging regions during training, we propose a Patch-wise Loss function. Built upon pixel-level losses, this function computes error distributions at the patch level and employs a Softmax weighting mechanism to emphasize patches with higher prediction errors, thereby directing the model's attention toward areas that are indistinct or irregularly shaped. This effectively mitigates the inherent limitation of ViT in modeling fine-grained details, significantly enhancing its capability to accurately identify subtle lesion areas. The main contributions of our study are summarized as follows:

- We propose a hybrid encoder architecture that integrates a CNN module before the ViT encoder, enabling enhanced extraction of local textures and structural details while maintaining the global context modeling strengths of Transformers
- We introduce a dual-decoder-based hybrid training framework that jointly performs image segmentation and reconstruction. The training strategy enables the reconstruction task to enrich semantic representations, while task collaboration improves segmentation accuracy and generalization.
- We design a Patch-wise Loss function to enhance the model's focus on detailed regions. By incorporating a temperature parameter and a Softmax-based weighting strategy, the loss adaptively emphasizes high-error patches, improving segmentation accuracy in blurred boundaries and complex textures.
- We conduct comprehensive evaluations on ISIC2017, ISIC2018, and PH2 datasets. Compared with mainstream segmentation models, SegMAE achieves superior. performance across various metrics, demonstrating high accuracy and robustness in fine-grained lesion delineation.

2 Related Works

2.1 Vision Transformer

Transformers were first introduced by Vaswani et al. [8] for machine translation, and later extended to vision by Dosovitskiy et al. [4], who proposed Vision Transformer

(ViT), by dividing the image into patches and modeling them as sequential tokens. Subsequently, hierarchical Vision Transformers such as Swin Transformer and PVT [9,10] were developed, introducing pyramidal structures and local attention to balance global modeling with efficiency.

In recent years, Transformers have been widely explored in image segmentation. For instance, Zheng et al. [11] proposed SETR, which leverages ViT as an encoder and combines it with various CNN-based decoders to improve semantic segmentation performance. Chen et al. [7] introduced a Transformer module into the bottleneck of a U-Net architecture to achieve multi-organ segmentation. Valanarasu et al. [12] designed a Transformer-based attention guidance mechanism to improve the accuracy of 2D medical image segmentation.

However, despite their global modeling advantages, these approaches still struggle with skin lesions' unique challenges like blurred boundaries, diverse lesion shapes, and complex textures. To address this, we propose a hybrid encoder that integrates CNN and Transformer modules, aiming to bridge local detail perception and global semantic understanding for more robust skin lesion segmentation.

2.2 Masked Autoencoder

Masked Autoencoders (MAE), introduced by He et al. [13], are an efficient self-supervised learning paradigm. By randomly masking a large portion of the input and reconstructing it from the visible patches, MAE enables efficient pretraining and has become a popular framework for visual representation learning.

Following its success in natural image domains, researchers have extended MAE to medical imaging tasks, especially for enhancing segmentation performance. Gupta et al. proposed MedMAE [14], which employs a multi-scale patch input strategy for 3D medical images, enabling better modeling of anatomical structures. Li et al. introduced UM-MAE [15], where MAE is incorporated as a pretraining module in a multimodal segmentation network, enhancing the understanding of fine-grained structures.

However, existing work typically uses MAE as a pretraining tool, underutilizing its structural modeling potential. In contrast, we incorporate MAE as a dedicated decoding branch in our framework, allowing it to learn structural and contextual cues and effectively support the main segmentation task.

3 Methodology

This research introduces a dual-branch skin lesion segmentation model that integrates both image segmentation and reconstruction tasks (see Fig. 1). The model takes raw skin lesion images as input, first extracting local spatial features through a CNN module, and then feeding the processed features into a ViT encoder to construct global semantic representations of the image. To enhance the computational efficiency of the encoder and improve its representation learning ability, a masking mechanism is introduced, whereby only a subset of patches is retained as input to the ViT encoder.

The output of the ViT encoder is then directed into two task-specific branches. For the segmentation task, a Cascaded Upsampler decoder progressively upsamples the encoded features back to the original resolution to generate the segmentation mask; for the reconstruction task, the token sequence with mask is passed into MAE decoder to recover the occluded image regions and reconstruct the complete image. In terms of training strategy, the model adopts a pretraining-driven hybrid training strategy. Specifically, the ViT encoder is first optimized via the reconstruction task, followed by a joint training phase that alternates between segmentation and reconstruction tasks. The optimization of both tasks is guided by a combination of loss functions, including Dice Loss, Binary Cross-Entropy Loss, and Patch-wise Loss, which emphasizes local reconstruction errors. These losses collaboratively strengthen the model's attention to critical regions during backpropagation, thereby enhancing overall segmentation performance.

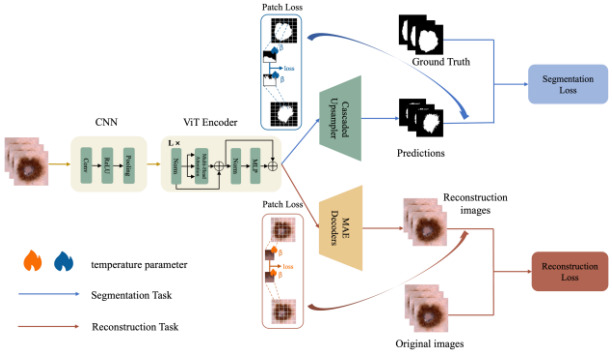

Fig. 1. A general overview of our SegMAE framework.

3.1 Hybrid Encoder

In SegMAE, we adopt a hybrid encoder architecture that integrates CNN with Transformers. Specifically, the raw input skin lesion image is first processed by a CNN block to extract intermediate feature maps. Unlike traditional approaches that directly divide the raw image into patches, we extract patch from the output feature maps of the CNN and feed them into the Transformer encoder. This design not only leverages the CNN's strength in modeling local textures, but also allows the decoder pathway to incorporate medium-and-high resolution CNN representations, thereby enhancing the model's capacity to capture structural details and local texture patterns in the image.

To further improve the encoder's ability to extract meaningful representations, we introduce a Masking mechanism that randomly samples a subset of visible patches from the input patch token sequence for encoding, discarding the remaining patches without using additional mask tokens, thus avoiding redundant computation.

Formally, let the input image of size be partitioned into non-overlapping patches of size $P \times P$, resulting in a total of patches. The resulting patch sequence can be denoted as:

$$\mathcal{Z} \in R^{N \times D} \tag{1}$$

where is the dimension of each token.

During the Masking stage, we apply random sampling over to obtain a visible patch subset while discarding the remaining invisible patch subset. This process can be formalized as:

$$\mathcal{Z}_{\mathcal{V}} \cup \mathcal{Z}_{\mathcal{M}} = \mathcal{Z},\ \mathcal{Z}_{\mathcal{V}} \cap \mathcal{Z}_{\mathcal{M}} = \emptyset \tag{2}$$

where and represent the sets of visible and masked patch tokens, respectively.

To increase the learning difficulty and improve generalization, a high masking ratio is applied during encoding, retaining only a small portion of patches for computation. This strategy effectively breaks the spatial redundancy among patches, preventing the model from relying on local neighborhood extrapolation. Instead, it compels the encoder to capture richer and more meaningful global context, which significantly benefits the downstream segmentation performance.

3.2 Dual Decoder Architecture

To simultaneously perform image segmentation and image reconstruction, we design a dual-decoder architecture. In this architecture, the output of the ViT encoder is fed into two parallel decoder branches: one for the segmentation task, implemented via a Cascaded Upsampler, and the other for the reconstruction task, realized through an MAE Decoder. This parallel multi-task design enables mutual enhancement between the two tasks during training.

Segmentation Decoder. In the multi-task architecture of SegMAE, the MAE branch focuses on the image reconstruction task. The input to this branch consists of two components: (i) the visible patch obtained from the ViT encoder, and (ii) the learnable mask tokens corresponding to the masked positions. These two types of tokens are concatenated and combined with positional embeddings to form the decoder input:

$$Z_{\text{dec}} = Concat(\mathcal{Z}_{\mathcal{V}}, \mathcal{Z}_{\mathcal{M}}) + \boldsymbol{E}\text{pos} \tag{3}$$

where denotes the positional embeddings that encode spatial information. This combined input is then passed through a series of lightweight Transformer modules, producing the reconstructed output for the masked patches:

$$\widehat{x_j} = f_{\text{dec}}(Z_{\text{dec}})_j \tag{4}$$

where represents the decoder network and is the predicted value of the j-th masked patch. Owing to the shallower and narrower architecture of the decoder compared to the encoder, the per-token computation is significantly reduced, thereby greatly improving pretraining efficiency and lowering computational overhead.

The reconstruction objective is to minimize the pixel-level error between the predicted and original image regions within the masked areas. A weighted MSE loss is

applied, calculated only over the masked patches, guiding the model to focus on structural restoration. The detailed formulation of this loss is provided in the subsequent Loss Function section.

Notably, we adopt a random sampling masking strategy during training, where visible patches are uniformly sampled from the full patch set, and the rest are masked. This approach disrupts the spatial redundancy between patches, forcing the model to rely more on global contextual reasoning for image restoration, thereby enhancing its structural understanding and reconstruction quality.

Segmentation Decoder. For the segmentation branch, we employ a Cascaded Upsampler (CUP) as the decoding module to progressively recover the low-resolution feature maps produced by the ViT encoder. CUP utilizes a multi-stage upsampling structure to gradually restore the feature dimensions from back to the original input resolution $H \times W$. Each upsampling stage consists of a upsampling operation, a 3×3 convolutional layer, and a ReLU activation function. Skip connections are used to fuse features from different stages of the encoder, enabling multi-scale feature aggregation and facilitating the generation of high-quality segmentation predictions with enhanced spatial detail and semantic consistency.

3.3 Training Strategy

To achieve collaborative optimization of the segmentation and reconstruction tasks, we adopt a pretraining-driven hybrid training strategy in this research. Under this strategy, the model training process is divided into two phases:

In the first phase, namely the pretraining stage, we exclusively train the reconstruction branch (i.e., the MAE Decoder). At this point, the model loads the official pretrained weights of ViT and performs image reconstruction training on the skin lesion dataset. The segmentation branch remains frozen during this stage. The number of training epochs is set to 50, which, based on empirical results, is sufficient for the model to effectively capture structural and textural patterns in skin images. With the high-ratio masking mechanism, the model learns to reconstruct complex image content, thereby laying a solid foundation for representation learning in subsequent segmentation tasks.

In the second phase, namely the hybrid training stage, we alternately optimize the segmentation and reconstruction tasks. Specifically, in each training cycle, the model first performs segmentation and computes a combined loss (comprising Binary Cross-Entropy Loss, Dice Loss, and Patch-wise Loss), followed by backpropagation to update parameters. It then carries out an image reconstruction step, using the Patch-wise weighted MSE loss to further refine the feature extraction capability of the encoder.

This training strategy establishes a complementary relationship between the two tasks. The segmentation task guides the model to focus on semantic region recognition and localization, while the reconstruction task enhances its ability to model fine-grained structures and spatial details within the image.

3.4 Loss Function

Segmentation Loss. The segmentation task for skin lesion images often suffers from severe class imbalance, where the number of pixels belonging to the lesion region is significantly smaller than that of the background. This imbalance can cause the model to overlook small lesion areas during training. To address this issue, we design a composite loss function composed of three components: Patch-wise Loss, Binary Cross-Entropy Loss (BCE), and Dice Loss. Among them, the Patch-wise Loss is a locality-aware enhancement loss function specifically designed for the ViT architecture, which guides the model to focus more on regions with larger prediction errors. The construction process is as follows:

First, we compute the pixel-wise squared error:

$$L = (pred - gt)^2 \tag{5}$$

where denotes the predicted pixel value, and is the ground truth label. represents the MSE at the pixel level.

Then, Next, we compute the average loss within each patch:

$$L'_i = \frac{1}{l}\sum_{j=1}^{l} L_{i,j} \tag{6}$$

where is the mean error of the i-th patch, is the number of pixels in each patch, and is the squared error of the j-th pixel in the i-th patch.

Then, we introduce a temperature factor to scale the patch-wise losses and apply a Softmax function to obtain attention weights:

$$w_i = Softmax\left(\frac{L'_i}{\beta}\right) \tag{7}$$

where is a hyperparameter controlling the sharpness of the Softmax distribution, and is the weight assigned to the i-th patch, giving higher priority to patches with larger reconstruction errors.

Finally, the Patch-wise Loss is calculated as:

$$\mathcal{L}_{patch} = \frac{1}{N}\sum_{i=1}^{N} w_i \, L'_i \tag{8}$$

where is the number of patches. This formulation ensures that the patch-wise weighted loss is converted into an average pixel-level loss, making it comparable across images of varying sizes or patch configurations. Additionally, it prevents isolated large patch errors from dominating the training, thus enhancing stability and generalization.

In addition, we incorporate two commonly used loss functions: Binary Cross- Entropy (BCE) Loss and Dice Loss. The BCE Loss is defined as:

$$\mathcal{L}_{BCE} = -\sum_{i=0}^{N}[(1-\hat{y}_i)\ln(1-y_i) + \hat{y}_i \ln(y_i)] \tag{9}$$

where denotes the ground truth label, is the predicted probability, and is the total number of pixels.

The Dice Loss is given by:

$$\mathcal{L}_{Dice} = 1 - 2 \times \frac{2\sum_{i=0}^{N} y_i \widehat{y_i}}{\sum_{i=0}^{N} (y_i + \widehat{y_i})} \tag{10}$$

where the numerator denotes the intersection of the predicted and ground truth masks, while the denominator measures their combined extent, capturing the overlap quality.

Finally, the overall segmentation loss is formulated as a weighted combination of the three components:

$$\mathcal{L}_{Seg} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \mathcal{L}_{Dice} + \lambda_3 \mathcal{L}_{patch} \tag{11}$$

where λ_1, λ_2, are weighting coefficients that determine the contribution of each loss component to the total loss.

Reconstruction Loss. We introduce two key modifications to the reconstruction loss originally proposed by MAE in this study: (1) the use of a Designated Masking strategy, where a deterministic mask replaces the original random masking mechanism; and (2) the introduction of a temperature coefficient β, which amplifies the training weight of high-error regions, thereby accelerating the convergence of the MAE branch parameters and improving reconstruction quality in critical areas.

In the reconstruction branch, we adopt a weighted mechanism similar to the Patch-wise Loss. First, the reconstruction error for each pixel is computed, followed by patch-level normalization, and then a Softmax operation is applied to generate a weight distribution. This mechanism enables the model to automatically focus on regions that are more difficult to reconstruct during training, enhancing its ability to model fine details and improve perceptual understanding of images. The basic formulation of the loss function is defined as:

$$\mathcal{L}_{MAE} = \frac{1}{N} \sum_{i=0}^{N} \left(y_i^{true} - y_i^{pred} \right)^2 \tag{12}$$

where denotes the ground truth pixel value, denotes the predicted reconstruction value, and represents the number of pixels involved in the reconstruction. This loss is computed only over the masked patches, avoiding overfitting on the visible regions and improving the encoder's ability to model the global structure of the image.

Additionally, to improve training efficiency, we employ a temperature-adjusted Softmax weighting mechanism to emphasize patch-level reconstruction errors. The corresponding mathematical formulation is consistent with the Patch-wise Loss described earlier and is thus omitted here.

4 Experiments

4.1 Datasets

This paper utilizes three publicly available datasets for performance evaluation:

- ISIC2017 [26] dataset includes 2000 training images, 150 validation images, and 600 test images. All images have been manually annotated by professional dermatologists for segmentation tasks. Following Cheng et al., the colors of the images are normalized using the gray world algorithm.
- ISIC2018 [27] dataset contains 2594 RGB skin lesion images with various types of skin lesions at different resolutions. Adopting the partitioning approach by Wu et al., the dataset is split into 1815 training images, 259 validation images, and 520 test images, maintaining a 7:1:2 ratio.
- PH2 dataset [28] consists of 200 RGB skin lesion images. Following a similar partitioning strategy (7:1:2 ratio), we use 140 images for training, 20 for validation, and 40 for testing.

4.2 Evaluation Metrics

For comprehensive comparison, we employ several evaluation metrics to assess the performance of the proposed model, including Accuracy (ACC), Dice coefficient (Dice), Intersection over Union (IoU), Sensitivity (SE), and Specificity (SP).

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{13}$$

$$Dice = \frac{2\times TP}{2\times TP+FP+FN} \tag{14}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{15}$$

$$SE = \frac{TP}{TP+FN} \tag{16}$$

$$SP = \frac{TN}{TN+FP} \tag{17}$$

where TP refers to the number of correctly segmented skin lesion pixels, TN refers to the number of correctly identified background pixels, FP refers to the number of background pixels incorrectly labeled as skin lesion pixels, and FN refers to the number of skin lesion pixels incorrectly predicted as background pixels.

4.3 Implementation Details

The proposed model is developed using the PyTorch framework, with all experiments conducted on an Nvidia 4090 RTX 24G GPU. To maintain uniformity in model input and enhance computational efficiency, we resized all training, validation, and test images to a standard resolution of 224×224. Additionally, to improve model initialization, we adopted ViT-B_16 as a pre-trained model.

For training, we utilized a stochastic gradient descent optimizer with a weight decay of 0.001. The initial learning rate was configured at 0.008, and the ReduceLROnPlateau algorithm was applied for dynamic learning rate adjustment. The model was trained with a batch size of 24 over 100 epochs. The weight parameters λ_1, and were assigned values of 1.2, 0.4, and 0.8, respectively.

To reduce the risk of overfitting, we followed the strategy outlined by Wu et al. and incorporated multiple data augmentation techniques. These included random horizontal and vertical flips, random rotations within a range of -15 to 15 degrees, and random modifications to brightness and contrast within predefined limits.

4.4 Evaluations and Analyses

ISIC2017 dataset. Experimental results on ISIC 2017 dataset, as presented in Table 1, demonstrate that the proposed SegMAE model achieves outstanding performance across multiple evaluation metrics. Specifically, SegMAE attains the highest scores in Dice coefficient (0.864), SE (0.858), and ACC (0.938), highlighting its strong capability in accurately identifying lesion regions, particularly excelling in the detection of small or indistinct lesions.

Table 1. Quantitative comparison on the ISIC 2017 dataset.

Method	Dice	SE	SP	ACC	IoU
U-Net [1]	0.783	0.806	0.954	0.933	0.696
UNet++ [2]	0.832	0.830	0.965	0.925	0.743
Att U-Net [3]	0.808	0.800	0.978	0.915	0.717
FocusNet [16]	0.832	0.767	**0.990**	0.921	0.756
DoubleU-Net [17]	0.845	0.841	0.967	0.933	0.760
DAGAN [18]	0.859	0.835	0.976	0.935	**0.771**
TransUNet [7]	0.841	0.807	0.979	0.932	0.755
FAT-Net [19]	0.850	0.840	0.973	0.933	0.765
ResGANet-MsASPP[20]	0.862	0.842	0.950	0.936	0.764
SegMAE	**0.864**	**0.858**	0.961	**0.938**	0.757

Compared with traditional models such as U-Net, UNet++, and Att U-Net, SegMAE significantly improves foreground sensitivity while maintaining high overall accuracy. This improvement is largely attributed to the integration of the reconstruction decoder and the Patch-wise Loss, which enables the model to more precisely localize blurred or small lesion regions. Moreover, when compared with more advanced architectures like TransUNet and FAT-Net, which aim to fuse local and global features, SegMAE still shows clear advantages in both Dice and SE metrics, reflecting its superior ability in capturing fine-grained structures while preserving global consistency.

Although the IoU score (0.757) of SegMAE is slightly lower than that of certain methods (e.g., DAGAN at 0.771), it still achieves overall superior performance in terms of ACC, Dice, and SE. This demonstrates that SegMAE effectively balances accuracy and sensitivity, making it both practical and generalizable. In summary, the results on ISIC2017 validate the effectiveness of SegMAE's multi-task collaborative optimization strategy.

ISIC2018 dataset. Experimental results on ISIC 2018 dataset, as shown in Table 2, further validate the superior performance of SegMAE. Specifically, SegMAE achieves the highest scores in two key metrics: Dice coefficient (0.896) and SE (0.957), and also performs exceptionally well in SP (0.979) and ACC (0.947), outperforming several advanced models, including FAT-Net, CKDNet, and TransUNet.

Table 2. Quantitative comparison on the ISIC 2018 dataset.

Method	Dice	SE	SP	ACC	IoU
U-Net [1]	0.855	0.880	0.970	0.940	0.773
Att U-Net [3]	0.857	0.867	0.984	0.938	0.776
CPFNet [21]	0.877	0.895	0.966	0.950	0.799
TransUNet [7]	0.850	0.858	**0.986**	0.945	0.809
CKDNet [22]	0.878	0.906	0.970	0.949	0.804
FAT-Net [19]	0.890	0.910	0.970	**0.958**	**0.820**
SegMAE	**0.896**	**0.957**	0.979	0.947	**0.820**

Compared with the strong-performing FAT-Net, SegMAE improves Dice, SE, and SP by 0.6%, 4.7%, and 0.9%, respectively, indicating its superior capability in both lesion localization and boundary discrimination. The notably high SE score demonstrates the model's enhanced sensitivity to lesion detection, which is particularly valuable for. real-world clinical applications where missed detections can be critical.

In addition to its accuracy, SegMAE also exhibits strong stability and robustness. It maintains consistent performance across images with varying resolutions and lesion complexities. This robustness is largely attributed to the multi-task training strategy and the Patch-wise Loss mechanism, which effectively guide the model to focus on challenging regions, thereby improving the overall segmentation quality.

In summary, the experimental results on ISIC2018 not only reflect the model's stable performance on a large-scale dataset but also confirm the effectiveness and generalizability of its architectural design and training strategy.

PH2 dataset. The experimental results on the PH2 dataset, as presented in Table 3, further validate the generalization ability of SegMAE in small-sample medical image segmentation scenarios. Unlike the ISIC series datasets, which typically contain thousands of skin lesion images, PH2 consists of only 200 high-resolution images with relatively regular lesion shapes. Therefore, this dataset not only evaluates the model's learning ability under data-scarce conditions but also serves as a benchmark for its performance in relatively simple segmentation tasks.

As shown in Table 3, SegMAE outperforms all competing methods across key metrics, including Dice, IoU, ACC, and SP. Specifically, SegMAE achieves a Dice coefficient of 0.941, surpassing the strong-performing MB-DCNN (0.933) and iFCN (0.932), indicating improved precision in reconstructing the overall lesion regions. In terms of IoU, SegMAE reaches 0.881, outperforming all existing methods, reflecting superior

capability in differentiating fine-grained foreground and background areas. The model also attains ACC of 0.961 and SP of 0.982, demonstrating high reliability in background exclusion.

Table 3. Quantitative comparison on the PH2 dataset.

Method	Dice	SE	SP	ACC	IoU
U-Net [1]	0.894	0.913	0.959	0.923	0.841
Att U-Net [3]	0.900	0.921	0.964	0.928	0.858
DSNet [23]	0.920	0.960	0.961	0.948	0.872
iFCN [24]	0.932	**0.961**	0.959	**0.961**	0.876
MB-DCNN [25]	0.933	0.954	0.953	0.959	0.871
SegMAE	**0.941**	0.954	**0.982**	**0.961**	**0.881**

While models such as DSNet and iFCN have also shown competitive performance on PH2, they often rely on additional complex modules—such as color enhancement or lesion-centered attention—to compensate for representational limitations. In contrast, SegMAE leverages the synergy between the MAE-guided reconstruction branch and the main segmentation branch, maintaining a structurally concise design while achieving significant performance gains. This multi-task collaborative training strategy enables the model to effectively extract structural information even with limited data, resulting in high-quality segmentation outcomes.

In conclusion, SegMAE demonstrates not only strong performance on large-scale datasets but also excellent robustness and generalizability on small-sample tasks like PH2, highlighting its potential and adaptability as a universal medical image segmentation framework.

Ablation studies. To evaluate the effectiveness of each component in SegMAE, we conducted a series of ablation studies. We began with a single-branch U-shaped architecture consisting of a pure ViT encoder and a Cascaded Upsampler decoder as the 1) baseline model (①Baseline). We then progressively introduced key modules to form multiple comparative configurations: 2) adding the MAE decoder without pre-trained weights (②Baseline + MAE decoder); 3) adding the MAE decoder with pre-trained weights (③Baseline + MAE decoder + finetuned weights); 4) introducing the MAE decoder and Patch-wise Loss (L1) without pre-training (④Baseline + MAE decoder + L1); 5) introducing the MAE decoder and reconstruction loss L2 (⑤Baseline + MAE decoder + L2); 6) introducing both L1 and L2 without pre-training (⑥Baseline + MAE decoder + L1 + L2); and finally, 7) introducing the MAE decoder with pre-trained weights along with both L1 and L2 losses (⑦Baseline + MAE decoder + finetuned weights + L1 + L2).

Experimental results (see Table 4) demonstrate that the inclusion of the MAE decoder significantly enhances segmentation performance. For example, on the ISIC2017 dataset, the Dice score improved from 0.841 (baseline) to 0.860 after adding the MAE decoder, and further increased to 0.864 after loading the pre-trained weights. On the

PH2 dataset, the Dice score reached 0.941 with the full configuration. These incremental improvements confirm the effectiveness of both the MAE module and the designed loss functions. Notably, the integration of the MAE decoder did not compromise the reconstruction task performance but instead enhanced segmentation results through collaborative optimization. This highlights the strong synergy and enhancement effect of the multi-task mechanism and well-designed loss structure within the SegMAE framework for skin lesion segmentation.

Table 4. Ablation study results on ISIC2017, ISIC2018, and PH2 datasets.

Method	Acc(17)	IoU(17)	Dice(17)	Acc(18)	IoU(18)	Dice(18)	Acc(PH2)	IoU(PH2)	Dice(PH2)
①	0.932	0.755	0.841	0.942	0.810	0.868	0.952	0.860	0.926
②	0.935	0.752	0.860	0.924	0.774	0.873	0.947	0.852	0.924
③	0.937	0.756	0.864	0.932	0.801	0.889	0.950	0.871	0.931
④	0.935	0.747	0.857	0.938	0.811	0.892	0.954	0.867	0.931
⑤	0.936	0.750	0.858	0.943	0.813	0.893	0.957	0.877	0.937
⑥	0.938	0.754	0.862	0.947	0.815	0.889	0.955	0.872	0.937
⑦	0.939	0.757	0.864	0.947	0.820	0.896	0.961	0.881	0.941

5 Conclusion

This paper proposes SegMAE, a dual-decoder framework for skin lesion image segmentation, which jointly optimizes segmentation and reconstruction tasks to fully exploit the potential of ViT in global context modeling and local detail capturing. By introducing a two-branch architecture for segmentation and reconstruction, the model achieves clear task decoupling and effective feature complementarity. We design a "pretraining + hybrid training" strategy that significantly improves both generalization and training efficiency. Furthermore, the proposed Patch-wise Loss adaptively guides the model to focus on challenging regions such as blurry boundaries, thereby enhancing segmentation accuracy and robustness. Extensive experiments demonstrate that SegMAE consistently outperforms existing mainstream methods across multiple public skin lesion segmentation datasets, showcasing superior segmentation performance and strong generalization ability.

References

1. Ronneberger, O., Fischer, P., & Brox, T. (2015). 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 international publishing.
2. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop,

DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4 (pp. 3-11). Springer International Publishing.

3. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
5. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems, 34, 12077-12090.
6. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022, October). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.
7. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
10. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... & Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 568-578).
11. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6881-6890).
12. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24 (pp. 36-46). Springer International Publishing.
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).
14. Gupta, A., Osman, I., Shehata, M. S., & Braun, J. W. (2024). MedMAE: A Self-Supervised Backbone for Medical Imaging Tasks. arXiv preprint arXiv:2407.14784.
15. Li, X., Wang, W., Yang, L., & Yang, J. (2022). Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. arXiv preprint arXiv:2205.10063.
16. Gao, Y., Huang, R., Chen, M., Wang, Z., Deng, J., Chen, Y., ... & Li, H. (2019). FocusNet: imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck CT images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22 (pp. 829-838). Springer International Publishing.

17. Jha, D., Riegler, M. A., Johansen, D., Halvorsen, P., & Johansen, H. D. (2020, July). Doubleu-net: A deep convolutional neural network for medical image segmentation. In 2020 IEEE 33rd International symposium on computer-based medical systems (CBMS) (pp. 558-564). IEEE.
18. Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., ... & Firmin, D. (2017). DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE transactions on medical imaging, 37(6), 1310-1321.
19. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., & Wen, Z. (2022). FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis, 76, 102327.
20. Cheng, J., Tian, S., Yu, L., Gao, C., Kang, X., Ma, X., ... & Lu, H. (2022). ResGANet: Residual group attention network for medical image classification and segmentation. Medical Image Analysis, 76, 102313.
21. Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., ... & Chen, X. (2020). CPFNet: Context pyramid fusion network for medical image segmentation. IEEE transactions on medical imaging, 39(10), 3008-3018.
22. Jin, Q., Cui, H., Sun, C., Meng, Z., & Su, R. (2021). Cascade knowledge diffusion network for skin lesion diagnosis and segmentation. Applied soft computing, 99, 106881.
23. Hasan, M. K., Dahal, L., Samarakoon, P. N., Tushar, F. I., & Martí, R. (2020). DSNet: Automatic dermoscopic skin lesion segmentation. Computers in biology and medicine, 120, 103738.
24. Shuai, B., Liu, T., & Wang, G. (2016). Improving fully convolution network for semantic segmentation. arXiv preprint arXiv:1611.08986.
25. Xie, Y., Zhang, J., Xia, Y., & Shen, C. (2020). A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE transactions on medical imaging, 39(7), 2482-2493.
26. Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., ... & Halpern, A. (2018, April). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018) (pp. 168-172). IEEE.
27. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368.
28. Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013, July). PH 2-A dermoscopic image database for research and benchmarking. In 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 5437-5440). IEEE.