LUT-SLS: A Lightweight Transformer Network Based on U-Net for Skin Lesion Segmentation

Ming Zhao[1,2], Bingxue Zhou[1], Ling-Ju Hung[3]*

[1] School of Computer Science, Yangtze University, China
[2] School of Network and Information Security, Wuxi University, China
[3] Department of Creative Technologies and Product Design, National Taipei University of Business, Taiwan
hitmzhao@gmail.com, 010105zzing@gmail.com, ljhung@ntub.edu.tw

Abstract

In recent years, deep learning has achieved significant advancements in medical image segmentation, primarily focusing on CNN structures and Transformer-based architectures. However, these network architectures often suffer from high computational complexity and a large number of parameters. To address these challenges, this paper proposes a lightweight, novel structured network called LUT-SLS, based on U-Net and Transformer. Firstly, the overall structure integrates U-Net and Transformer, which effectively captures long-range dependencies in image relationships and contextual information, thereby improving segmentation accuracy. Secondly, a novel PLTS module is designed, which replaces the traditional self-attention mechanism with average pooling operations to extract global features and local details. Additionally, a novel MMLP structure is introduced, incorporating residual depth-separable operations into the traditional fully-connected framework. This enhances the processing of pooled features and further improves feature expression capability. Finally, the encoder and decoder parts are connected by the MSBN module, which facilitates the extraction of deep features while fusing encoder features. Experimental results demonstrate that the proposed model achieves competitive advantages in balancing the number of parameters, computational complexity, and performance compared to current leading models on multiple public datasets. This solution enables model deployment on IoT terminals, assisting doctors in making more accurate clinical decisions.

Keywords: Medical image segmentation, U-Net, Transformer, Lightweight network, Skin lesions

1 Introduction

Medical image segmentation is a crucial step in medical image processing, significantly enhancing the accuracy of early disease detection and diagnosis. Researchers have developed various image processing techniques [1-4], such as the Otsu algorithm [4] and Canny edge detection [2], and widely applied them to image segmentation tasks. However, these methods rely on hand-designed features and thresholds, which are insufficient for complex scenes, necessitating more accurate and efficient segmentation techniques. In recent years, deep learning technology has rapidly advanced, with deep convolutional neural networks becoming the mainstream methods for medical image segmentation [5-7]. Among these, U-Net [7] stands out as a classical deep learning model. Its effective structure, along with its variants, has achieved remarkable results in medical image segmentation tasks [8-9]. For example, U-Net++ [9] improves model performance by introducing a nested U-Net structure, while Attention U-Net [8] enhances the model's focus on important image regions by incorporating an attention mechanism. Additionally, methods like [10-11] have been proposed to achieve more efficient segmentation through deeply separable convolution. Despite their success, U-Net networks have limitations, such as low processing efficiency for large-size images and limited ability to model global context. To address these issues, researchers have proposed dynamically adjusting convolution techniques [12-14], resulting in more efficient inference and better performance in handling large and complex images.

Meanwhile, researchers have also proposed a series of novel segmentation networks based on Transformers [15-19] to address the limitations of U-Net from a different perspective. Among these, the Swin Transformer [18], introduced by Liu et al. in 2021, is an innovative segmentation network that adopts a hierarchical visual Transformer structure with a shift window to capture the global information of an image. Additionally, Chen et al. proposed the SwinUNet model [15] in 2022, which combines the U-Net structure with the Swin Transformer and applies it to multi-organ segmentation, effectively improving segmentation performance and efficiency. There are also other Transformer-based models designed for specific tasks [20-23], such as TransBTS [23] and Swin UNETR [21] for brain tumor segmentation, and TransDeepLab [20] for medical image segmentation. Despite their strong performance, Transformer-based models have some disadvantages. One major issue is their high computational complexity. The self-attention mechanism of Transformer models requires substantial computational resources to process each position in the image, which can result in an excessive computational burden when dealing with large medical images. To mitigate

*Corresponding Author: Ling-Ju Hung; E-mail: ljhung@ntub.edu.tw
DOI: https://doi.org/10.70003/160792642024092505015

this, some researchers have proposed pruning techniques [24-25] to reduce the computational complexity of the models. While these methods have made significant progress in enhancing model performance and achieving lightweight structures, they may still face challenges such as network incompatibility or limited generalizability to other task domains.

To address the computational efficiency and performance challenges in medical image segmentation, this paper proposes a new lightweight network structure, LUT-SLS, specifically designed for skin lesion segmentation. LUT-SLS strikes a balance between computational efficiency and segmentation accuracy by incorporating both U-Net and Transformer elements, capturing long-range dependencies and contextual information in the images. The proposed structure includes several innovative components: a PLTS module (see Figure 1(a)) designed to reduce the number of parameters and computational complexity while maintaining high performance by using a lower number of channels and replacing the traditional self-attention mechanism with an average pooling operation; an MMLP structure that enhances the processing of pooled features through a residual depth-separable operation built on top of the traditional fully-connected layer, improving feature expression capabilities; and an MSBN module that connects the encoder and decoder parts, enabling further extraction of deep features while effectively fusing encoder features.

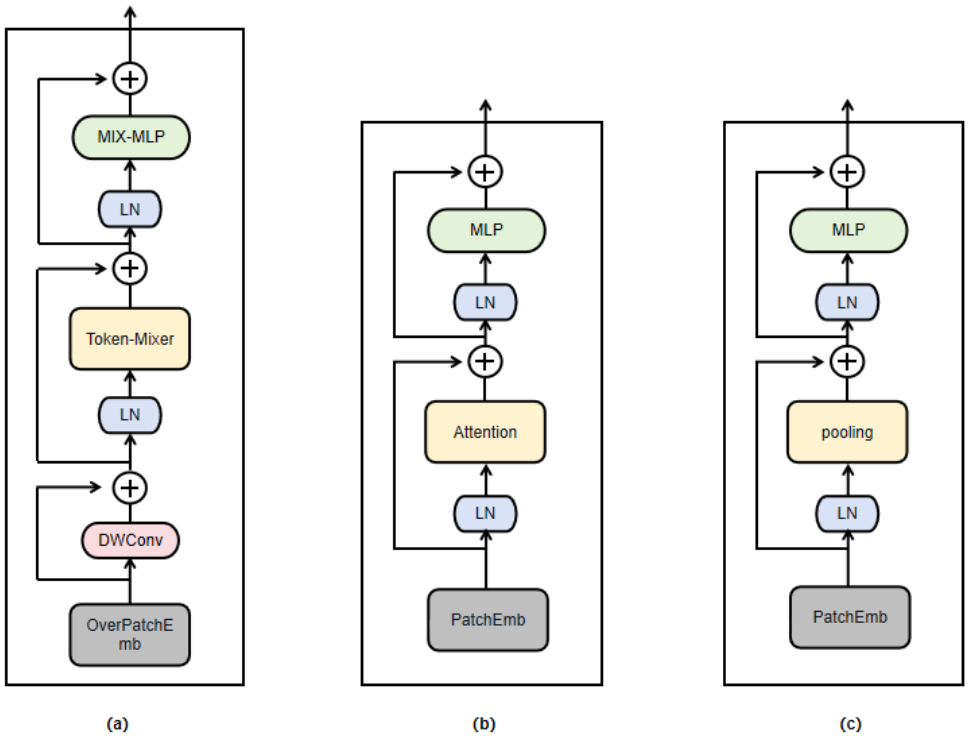
The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 provides a mathematical explanation of the model and describes the proposed method, Section 4 presents experimental results and specific details, Section 5 conducts ablation and comparative experiments with multi-channel analysis, and Section 6 summarizes the findings.

2 Related Works

2.1 U-Net Network Architecture

The U-Net model is a classical convolutional neural network widely used in medical image segmentation tasks. It features an encoder-decoder structure, with skip connections that link the bottom features to the top features to preserve richer spatial information. However, multiple down-sampling operations in the U-Net model can lead to resolution loss and boundary blurring. To further improve segmentation performance, dense skip connections [26] and multi-scale skip connections [27] have been employed to enhance the model's expressive power. Additionally, incorporating various attention mechanisms—such as channel attention [28] and spatial attention [29]—enables the model to selectively focus on the most relevant or information-rich parts of the input data, thereby improving the overall segmentation effect.

Recent studies have explored combining the U-Net architecture with Transformer networks [30-31] and multilayer perceptron to apply these fused architectures across various domains. Building on this concept, we have designed a lightweight network model that leverages the symmetric encoder-decoder architecture of U-Net.



(a) PLTS structure (b) Transformer structure (c) Meta-Former structure

Figure 1. U-Net network architecture

2.2 Transformer Structure

Due to Transformer's powerful modeling capabilities in natural language processing, it has sparked a revolution in computer vision. Traditional convolutional neural networks are limited by local induction bias and can only capture feature information within a local range. In contrast, the VIT (Vision Transformer) structure, a variant of Transformer, can capture global information through the self-attention mechanism (see Figure 1(b)). However, this global sensing capability often comes with significant computational and complexity overhead. To address this issue, researchers have proposed various improvement strategies. These include the introduction of local windows and cross-window communication to achieve sparse connectivity and adaptively learning model parameters [32]. Additionally, methods such as axial attention [33], separable attention [34], and fully convolutional structures [35] have been developed. These strategies enhance the model's computational efficiency, memory efficiency, and adaptability. To further achieve lightweight models, Meta-Former [36] posited that Transformer's success does not entirely depend on the self-attention mechanism. They demonstrated that the Token Mixer module could be replaced by a pure MLP or other structures and verified this assumption with a simple pooling operation (see Figure 1(c)).

Building on this concept, this paper presents a lightweight network model based on the Pool-Former architecture. This model significantly reduces the number of parameters and computational complexity while maintaining accuracy.

2.3 Bottleneck

The design of the bottleneck section typically employs a Transformer structure and a multi-scale spatial pyramid [37], which aids in extracting more informative feature representations and enhances the model's focus on critical features. To lighten the overall structure, we propose a bottleneck section based on a multi-scale spatial pyramid structure in this paper. This design captures contextual and detailed information more comprehensively by extracting and fusing information from different layers of feature maps. In the context of medical imaging for dermatology segmentation, this bottleneck structure adapts better to dermatology regions of varying sizes and shapes, improving the perception of details and boundaries. Consequently, it helps the network better understand the complexity of dermatologic diseases, thereby enhancing segmentation performance.

3 Methods

Building on these studies and considering the diversity of skin lesions along with the complex background and noise in the images, we propose a new network structure (see Figure 2) designed to accurately capture the features of skin lesions and improve segmentation accuracy. In this section, we provide a mathematical explanation of the model and detail the overall network structure, along with the key technical aspects of each module.

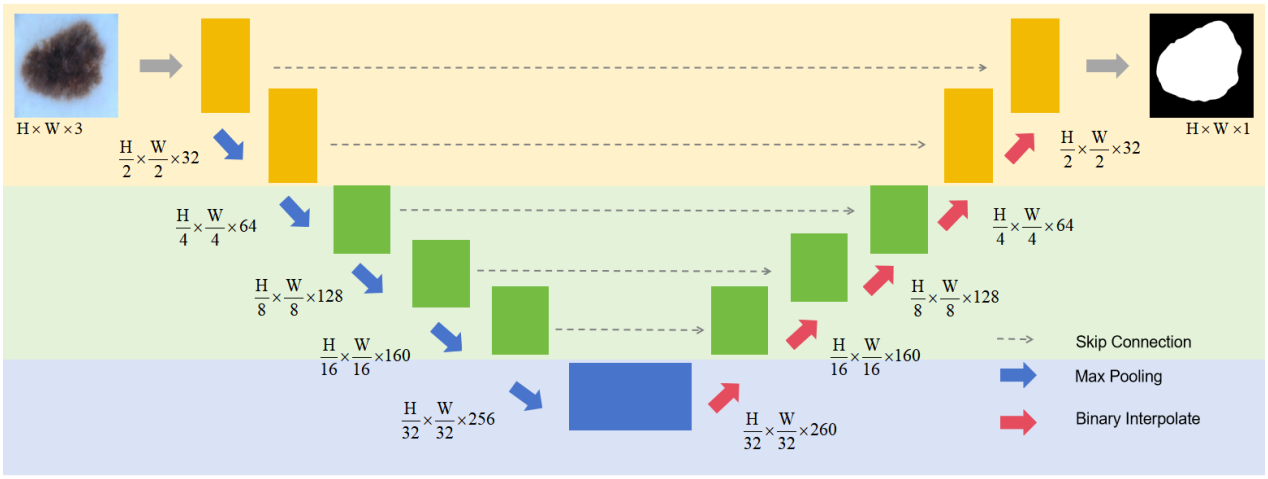


Figure 2. Overall structure of the model

3.1 Problem and Mathematical Description

For the skin lesion segmentation optimization problem, our goal is to minimize the objective function by optimizing the model parameters within the given parameter range constraints. We define the optimization problem as follows:

$$\text{Minimize } \theta \sum_i L(y_i, f(x_i;\theta)) + \lambda R(\theta),$$
$$\text{s.t. } \theta_{\min} \le \theta \le \theta_{\max} \tag{1}$$

where θ denotes the model parameters, x_i denotes the ith input image, y_i denotes the label corresponding to the ith input image, $f(x_i, \theta)$ is the image segmentation model defined by the model parameter which maps the input image to the predicted segmentation result, $L(\cdot,\cdot)$ is the loss function used to measure the difference between the predicted result and the true label, $R(\theta)$ is the regularization term used to constrain the model parameters and to control the model's complexity, and λ is the regularization coefficient used to balance the weights between the loss function and regularization.

To satisfy the parameter range constraints, we introduce upper and lower bounds for the parameters. These bounds ensure that the parameters stay within an acceptable range. Specifically, $\theta_{\min}$ represents the lower bound, while $\theta_{\max}$ represents the upper bound of the parameter.

The optimization objective is to achieve optimal image segmentation by minimizing the objective function through adjustments to the model parameters. During this process, the gradient descent algorithm is employed to update the model parameters, ensuring that constraints are considered at each iteration.

3.2 Network Design

LUT-SLS is designed based on the U-Net structure, as illustrated in Figure 2. The entire model consists of three modules: (i) the convolution module, (ii) the PLTS module, and (iii) the MSBN module. The convolution module encompasses the structure of the first two layers, while the last three layers comprise the PLTS structure. The MSBN module is situated between the last layer of the encoder and the decoder. Similar to U-Net, LUT-SLS employs skip connections between each layer to integrate low-level image details with high-level semantic information, enhancing feature fusion and improving segmentation accuracy.

3.3 Convolution Module

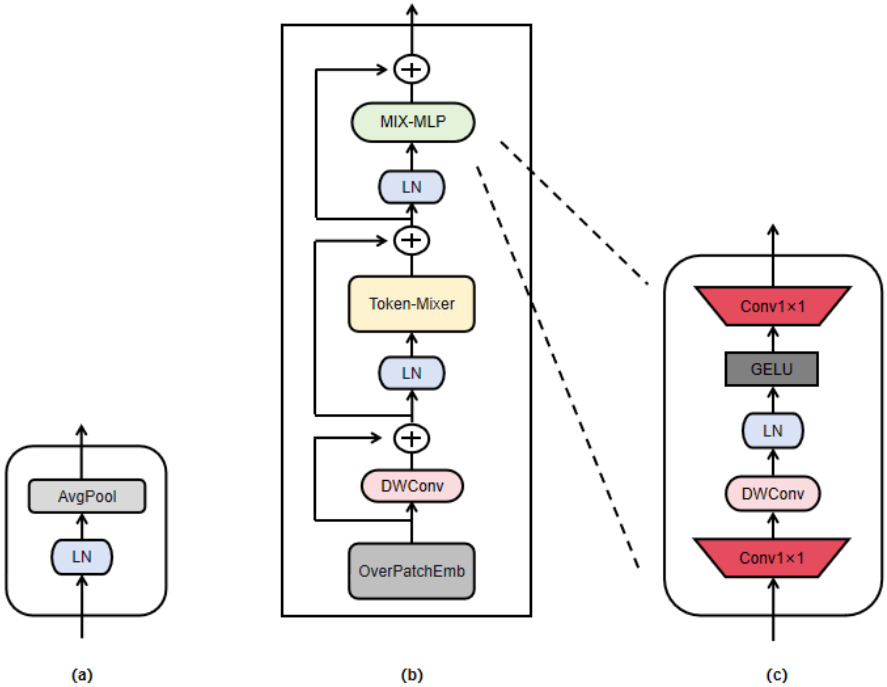
The input feature map is first passed through a 1×1 extended convolutional layer, which increases the number of channels and enhances the network's expressive power. The expansion factor determines the ratio of output channels to input channels. The output from the extended convolutional layer is then fed into a depth-separable convolutional layer, which includes the SE module to adaptively learn channel weights. These weights are applied to the output of the depth-separable convolution to adjust the feature map in a weighted manner. Skip connections are used to perform element-wise summation of the inputs and outputs of the convolutional module. Additionally, Drop-Connect operations are employed to reduce overfitting and lower the network's complexity.

3.4 PLTS Module

In deep learning, balancing high computational complexity and model performance is a common challenge. To address this issue, we propose the PLTS module, which includes residual depth-separable convolution, the TPLM structure, and the MMLP structure (see Figure 3).

First, the transformation and preservation of features are accomplished using residual depth-separable convolution structures. Depth-separable convolution is an efficient operation that significantly reduces computational load and the number of parameters by separating spatial convolution from channel convolution. This design enables the network to maintain high performance while minimizing computational complexity. Additionally, residual concatenation preserves a portion of the input feature information by summing it with the output of the convolutional layer, helping to prevent gradient vanishing during the training process.

$$Y = X + Conv_{ds}(X), X \in R^{n \times d \times h} \tag{2}$$



(a) The residual depth separable convolution structure (b) The TPLM structure (c) The MMLP structure

Figure 3. Schematic diagram of the PLTS module

Second, the TPLM module down-samples the input features through average pooling operations to capture global contextual information. Average pooling is a simple yet effective down-sampling method that reduces the spatial dimensions of the features, thereby lowering computational complexity. By calculating the average value of each feature channel and using it as the output value, this method extracts more representative feature representations. Compared to the self-attention mechanism in traditional Transformer models, average pooling has a significant advantage in terms of computational efficiency while still capturing global contextual information effectively.

$$Z = Y + \mathrm{AvgPool}(Y) \tag{3}$$

Finally, the MMLP module learns representations of the input features through a series of transformations and nonlinear operations. This structure begins with a 1×1 convolution operation, followed by a depth-separable convolutional layer utilizing the GELU activation function to introduce nonlinearity, and concludes with another 1×1 convolution to complete the feature transformations. This sequence of operations enables the network to learn higher-level feature representations. Compared to traditional MLP, the MMLP reduces computational complexity while maintaining strong performance.

$$O = Z + \mathrm{Conv}_{1\times1}\left(\mathrm{GELU}\left(\mathrm{Conv}_{\mathrm{ds}}\left(\mathrm{Conv}_{1\times1}\left(Z\right)\right)\right)\right) \tag{4}$$

where X represents the input matrix; n denotes the batch size; d denotes the product of the height and width of the feature map; h denotes the dimension of each feature; $\mathrm{Conv}_{\mathrm{ds}}$ denotes the depth separable convolution operation; $\mathrm{AvgPool}(\cdot)$ denotes average pooling operation. $\mathrm{Conv}_{1\times1}(\cdot)$ denotes 1×1 convolution operation; and $\mathrm{GELU}(\cdot)$ denotes the activation function.

3.5 Attention Module

The attention method employed in this paper is SE channel attention, which enhances the network's response to different channels by adaptively adjusting the importance of channel features. This approach improves the model's expressive and generalization capabilities. The SE channel attention operates through two key steps: compression and excitation. During the compression phase, each channel's feature map is converted into a scalar using a global average pooling operation. In the excitation phase, a small fully connected network is introduced to weight each channel by learning the excitation weights. These weights indicate the importance of each channel for extracting useful information, allowing for adaptive tuning of channel features.

$$\begin{aligned} F_{out} &= Reshape \\ &\left(\sigma\left(MLP\left(Reshape\left(GlogalAvgPool\left(F_{in}\right)\right)\right)\right)\right) \\ &= Reshape\left(\sigma\left(MLP\left(Reshape\left(f_{gap}\right)\right)\right)\right) \\ &= Reshape\left(\sigma\left(MLP\left(f'_{gap}\right)\right)\right) \\ &= Reshape\left(\sigma\left(\alpha_2\left(\alpha_{1(f'_{gap})}\right)\right)\right) \\ &= Reshape\left(f_{mgap}\right) \\ &= F_{out} \end{aligned} \tag{5}$$

$$F = F_{in} * F_{out} \tag{6}$$

Algorithm 1. Channel attention

Input: Characterization data F_{in}
Output: channel attention weights F_{out}

Function SEAttention(F_{in})

01: $f_{gap} \leftarrow$ The input features of B × H × W × C are subjected to a global average pooling operation
02: $f'_{gap} \leftarrow$ For the f_{gap} Perform a Reshape operation to convert the feature to a B×C form
03: $f_{mgap} \leftarrow$ sends the f'_{gap} Feed into the MLP to calculate
04: $F_{out} \leftarrow$ For the f_{mgap} Reshape operation is performed to transform the features into B×C×1×1 form to get the channel attention weights
05: $F \leftarrow$ Combine the original input features F_{in} and channel attention weights F_{out} and multiply them together to get the new feature
06: Return F

end Fnction

3.6 MSBN Module

Skin images often exhibit significant variations in lesion size and shape, making it challenging for traditional segmentation methods to efficiently handle lesion information at different scales. Additionally, different channels may contain varying key information crucial for accurate segmentation results. To address these issues, this paper introduces the MSBN module, designed to tackle the problems of multi-scale and channel importance modeling in skin image segmentation. This module enables the network to better understand and characterize the details and structures of skin lesions, resulting in more accurate skin lesion segmentation results.

The module captures the multi-scale information of the input features through a multi-level spatial pyramid pooling strategy (see Figure 4). The input features are first passed through four maximum pooling levels, with the kernel size of each pooling level divided into 2, 3, 4, and 6 to obtain different levels of feature representations, and furthermore, the SE channel attention mechanism is used in each pooling, which weights the channel attention of each pooled feature map, and adaptively adjusts the importance of each channel to increase the model's focus on specific features. Second, after each pooling level, the dimensionality of the feature map is reduced to 1/C of the original dimensionality by applying deep convolution to each input channel independently and combining the outputs of the deep convolution by point-by-point convolution, where C denotes the number of channels

in the original feature map. In order to enhance the model's ability to perceive targets at different scales for skin lesions of different sizes, the pooled feature maps are restored to the original size by interpolation to ensure that the feature maps have the same spatial resolution. Finally, the processed multi-scale feature maps are pooled with the original feature maps to generate the final output feature maps. The operation of the MSBN module can be represented by the following mathematical expression:

$$P_i = MaxPool(F, i),\ i \in \{2,3,4,6\} \quad \textbf{(7)}$$

$$P_i = SE(P_i),\ i \in \{2,3,4,6\} \quad \textbf{(8)}$$

$$P_i' = PointwiseConv\left(DepthwiseConv(P_i)\right),\ i \in \{2,3,4,6\} \quad \textbf{(9)}$$

$$P_i'' = Interpolate\left(P_i', H, W\right),\ i \in \{2,3,4,6\} \quad \textbf{(10)}$$

$$F' = \left[F, P_2'', P_3'', P_4'', P_6''\right] \quad \textbf{(11)}$$

where i denotes the size of the pooling kernel, *MaxPool* denotes the maximum pooling operation, *DepthwiseConv* denotes the depth convolution operation, *PointwiseConv* denotes point-by-point convolution operation, *Interpolate* denotes interpolation operation, and "[]" denotes splicing operation.

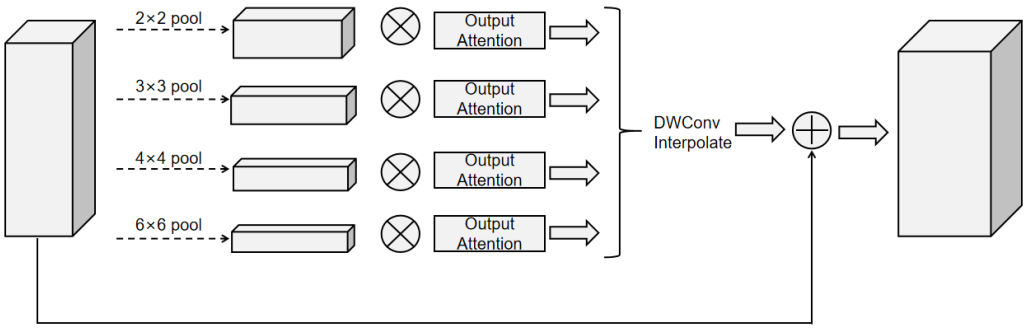


Figure 4. Schematic diagram of the bottleneck module

4 Experimental Results

In order to verify the effectiveness and superiority of our proposed method in the task of dermatologic image segmentation, we conducted a series of experiments and evaluated and compared its performance in several aspects. In this section, we present in turn the dataset used, the experimental setup, the loss function, the evaluation metrics, and the experimental performance comparisons.

4.1 Datasets

In order to comprehensively evaluate the performance of our proposed dermatological image segmentation method, we selected two representative datasets as experimental benchmarks, namely the ISIC2018 Task1 dataset and the ph2 dataset.

The ISIC2018 Task1 dataset is a competition dataset organized by the International Conference on Imaging of the Skin to advance research in the field of dermatology diagnosis and image analysis. The dataset contains dermatologic images from all over the world, covering many different types of skin lesions. Each image is equipped with a pixel-level segmentation mask, which is used as a reference standard for our model performance evaluation. Due to its large size and rich and diverse samples, the ISIC2018 Task1 dataset is able to provide sufficient data support to evaluate the generalization and robustness of our approach.

Another dataset is the ph2 dataset, which is a dataset provided by the Spanish Institute of Dermatology. It contains dermatological images from actual clinical situations and focuses on the task of image segmentation for melanoma. Since the ph2 dataset provides high quality images with accurate segmentation labels and has a small size, it makes the model proposed in this paper more focused on specific skin lesion types.

By using two datasets, ISIC2018 Task1 and ph2, we are able to perform a comprehensive evaluation of our method in this paper in different data contexts. In the next sections, we will describe the experimental setup and methodology in detail, and verify the effectiveness and superiority of our method on these datasets through comparative experiments and performance evaluation.

4.2 Implementations Setting

The experiment was conducted on v100 server using Pytorch framework and python 3.9. For data preprocessing, random rotation, flipping, scaling and normalization were used to resize the images all to 512×512. The optimizer is Adam with an initial learning rate of 1e-4 and CosineAnnealingLR scheduler with a minimum learning rate of 1e-5 and momentum of 0.9. The batch size is 8 and a total of 100 batches were trained. The loss function is a combination of binary cross-entropy loss and Dice loss.

4.3 Loss Function

In order to fully utilize the advantages of deep learning models and address the challenges in the task of dermatological image segmentation, we employ a fusion loss function that combines binary cross-entropy loss and Dice loss. The binary cross-entropy loss helps the model to better learn the distinction between target and background, while the Dice loss is able to emphasize the spatial consistency of the segmentation results. This combined loss function is designed to enhance the model's accurate localization of skin lesion regions and further improve the quality of segmentation results. In the next section, we will introduce the specific calculation and usage of this combined loss function in detail.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \quad (12)$$

$$DiceLoss = 1 - \frac{2 * \sum_{i=1}^{n}(input_i * target_i) + smooth}{\sum_{i=1}^{n} input_i + \sum_{i=1}^{n} target_i + smooth} \quad (13)$$

$$Loss = \gamma_1 * BCE + \gamma_2 DiceLoss \quad (14)$$

where N is the number of samples, y_i is the target value for the ith sample, and $\hat{y}_i$ is the predicted value for the ith sample. input and target denote the predicted and target values of the model, respectively, and smooth is a smoothing term to prevent the denominator from being zero. The final loss function is the weighted sum of the BCE loss and the Dice loss, γ_1 and γ_2 are 0.5 and 1, respectively.

4.4 Evaluation Indicators

In order to comprehensively evaluate the performance of the dermatological image segmentation method proposed in this paper, we introduce several commonly used evaluation metrics. These include IOU, Dice coefficient, accuracy, precision and recall.

First, IOU measures the proportion of overlapping parts between the segmentation result and the real label, which can quantify the consistency between the region segmented by the model and the real region, and the Dice coefficient predicts the similarity between the segmentation mask and the real label by calculating the similarity between the segmentation mask and the real label, and the closer the value of the calculation is to 1 means the more accurate the segmentation result is. Second, the accuracy rate is a metric to evaluate the correctness of the model's segmentation for the whole image, which defines the proportion of correctly segmented pixels to the total pixels, and can reflect the overall performance of the model. Precision and recall, on the other hand, focus on evaluating the model's performance in detecting skin lesion areas; precision measures the proportion of true positive samples among those predicted as positive by the model, while recall measures the model's ability to successfully detect true positive samples.

These five evaluation metrics consider the accuracy, consistency and comprehensiveness of the segmentation results and can provide a comprehensive assessment of the segmentation performance. In the subsequent experimental results, we will analyze and discuss the performance of the model on these metrics in detail to verify the effectiveness and superiority of the method proposed in this paper in the task of dermatological image segmentation.

$$\text{intersection} = \sum_{i=1}^{N}(\text{output}_i \wedge \text{target}_i) \quad (15)$$

$$\text{union} = \sum_{i=1}^{N}(\text{output}_i \vee \text{target}_i) \quad (16)$$

where N is the number of samples, output_i and target_i denote the predicted value and target value of the ith sample, respectively, and $\wedge$ and $\vee$ denote the bitwise-and and bitwise-or operations, respectively. intersection and union are used to compute the intersection and union of two binary images. Intersection is the number of pixels in both images that are 1, and union is the number of pixels in both images that are at least one of 1. TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

$$precision = \frac{TP}{TP + FP + smooth} \quad (17)$$

$$recall = \frac{TP}{TP + FN + smooth} \quad (18)$$

$$accuracy = \frac{TP + TN + smooth}{N + smooth} \quad (19)$$

$$IoU = \frac{intersection + smooth}{union + smooth} \quad (20)$$

$$Dice = \frac{2 * IoU}{IoU + 1} \quad (21)$$

4.5 Performance Comparison

In this paper, we compare several deep learning methods on two public datasets respectively, and it can be seen from Table 1 and Table 2 that the model proposed in this paper is better than or close to the other models in terms of the number of parameters, computation, and segmentation metrics. In order to further analyze the advantages of LUT-SLS, we discuss the following aspects in detail:

Table 1. Performance comparison on ISIC2018 dataset

Model	Params	GMac	IOU	Dice	Precision	Recall	Acc
U-Net [7]	31.04	218.92	0.7568	0.8580	0.9100	0.8203	0.9455
U-Net++ [6]	9.16	138.60	0.7537	0.8518	0.8850	0.8419	0.9439
Attention U-Net [5]	34.88	266.58	0.7430	0.8427	0.8768	0.8350	0.9414
UNeXt [30]	1.47	2.28	0.7905	0.8806	0.8591	0.9111	0.9504
MALUNet [38]	1.93	1.92	0.7608	0.8622	0.7880	0.9607	0.9411
Ours	2.06	2.96	0.8059	0.8894	0.9079	0.8799	0.9563

Table 2. Performance comparison on PH2 dataset

Model	Params	GMac	IOU	Dice	Precision	Recall	Acc
U-Net [7]	31.04	218.92	0.8703	0.9298	0.9297	0.9361	0.9602
U-Net++ [6]	9.16	138.60	0.8425	0.9125	0.8565	0.9489	0.9489
Attention U-Net [5]	34.88	266.58	0.8667	0.9276	0.9290	0.9310	0.9586
UNeXt_S [30]	0.25	0.41	0.8342	0.9083	0.9440	0.8808	0.9485
UNeXt [30]	1.47	2.28	0.8575	0.9223	0.8966	0.9544	0.9533
UNeXt_L [30]	3.99	5.67	0.8603	0.9240	0.9222	0.9296	0.9561
Ours	2.06	2.96	0.8792	0.9354	0.9185	0.9557	0.9624

Table 3. Results of ablation experiments on the ISIC2018 dataset

Model	Params	GMac	IOU	Dice	Acc
U-Net [7]	31.04	218.92	0.8703	0.9298	0.9297
PLTS	2.58	5.86	0.7856	0.8767	0.9509
Conv+PLTS	2.05	2.96	0.7993	0.8852	0.9537
Conv+PLTS+MSBN	2.06	2.96	0.8792	0.9354	0.9185

Table 4. Multi-channel experimental results on the ph2 dataset

Channel	Params	GMac	IOU	Dice	Acc
{32, 64, 128, 256, 512}	5.48	4.35	0.8497	0.9177	0.9513
{32, 64, 128, 160, 256}	2.06	2.96	0.8792	0.9354	0.9624
{16, 32, 64, 128, 160}	0.92	1.03	0.8614	0.9249	0.9574
{16, 32, 48, 64, 96}	0.33	0.57	0.8543	0.9204	0.9533
{8, 16, 24, 32, 64}	0.11	0.17	0.8252	0.9028	0.9452

Number of parameters and computation: U-Net has 31.04M parameters and 218.92GMac of computation, Attention U-Net has 34.88M parameters and 266.58GMac of computation, while our model has only 2.06M parameters and 2.96GMac of computation, which is much smaller than all other models. This indicates that our model is more lightweight, which can save storage space and runtime, and is suitable for deployment in mobile devices or low-configuration environments.

Segmentation metrics: the model proposed in this paper performs well on the IOU and Dice metrics, especially on the ISIC2018 dataset, as shown in Table 1. Comparing the IOU and Dice, our model improves 1.54% and 0.88% over UNeXt, respectively. As shown in Table 2, our model also achieves the highest IOU and Dice on the PH2 dataset, and also performs better on Precision and Recall. This indicates that our model can accurately recognize and segment skin lesion areas, and also exclude background noise and other interfering factors, which is very important for improving the accuracy and reliability of skin lesion diagnosis.

5 Discussion

To further explore the key factors and effectiveness of the dermatological image segmentation method proposed in this paper, we conducted a series of ablation experiments and multi-channel analysis. In the ablation experiments, we evaluated the impact of the key components of the method on the segmentation performance by removing them step by step. In addition, we conducted a multi-channel comparison experiment to find the optimal channel combination to improve the accuracy of the segmentation results by evaluating the effects of different channels on the segmentation performance.

5.1 Ablation Experiments

In this paper, U-Net is used as the baseline network, and as can be seen in Table 3, firstly the computation and complexity of the model is significantly reduced by using the PLTS module. Secondly, by using the convolution module and adjusting the framework structure of the whole model, the performance is improved and the computation is reduced at the same time. Finally the MSBN module is introduced to optimize the overall performance.

5.2 Multi-Channel Comparison Experiment

Based on the data in Table 4, it is found that as the number of channels increases, the parameters and computational complexity increase accordingly. When the number of channels is {32, 64, 128, 160, 256}, all the performance indicators reach the peak, while when the number of channels is {8, 16, 24, 32, 64}, all of these indicators decrease significantly. Therefore, the appropriate number of channels is selected according to different scenarios and requirements to achieve a balance between model complexity and performance.

6 Conclusion

In this study, we propose a lightweight Transformer network with a novel convolutional structure based on U-Net for medical image segmentation of skin lesions. The design of LUT-SLS integrates U-Net and Transformer frameworks and consists of three main modules: the convolution module, the PLTS module, and the MSBN module. These modules are combined within the five-layer U-Net structure to achieve effective modeling with reduced computation and complexity. Comparative experiments on several datasets demonstrate that our proposed structure optimizes comprehensive performance and outperforms current state-of-the-art methods.

Despite the progress made in dermatologic image segmentation, several areas require further exploration and improvement. Firstly, most current segmentation methods are validated and evaluated on specific datasets. Future research should focus on enhancing model generalization across different datasets to handle diverse real-world skin lesion scenarios. Secondly, dermatological image segmentation can benefit from incorporating other image modalities, such as infrared and multispectral images. Future research could integrate multimodal images to provide more comprehensive and accurate segmentation results.

In summary, while the field of dermatologic image segmentation faces many challenges, it also presents numerous opportunities. Through continuous research and innovation, we can further improve segmentation performance, offering more accurate and reliable support for clinical diagnosis and treatment.

Acknowledgement

Parts of the research were supported by the New Generation Information Technology Innovation Project 2022 for “Color perception Test Map Generation and Color perception Detection and Correction Assistant System” under grant no. 2022IT036.

References

[1] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1124−1137, September, 2004.

[2] J. Canny, A computational approach to edge detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6, pp. 679−698, November, 1986.

[3] J. Malik, P. Perona, Preattentive texture discrimination with early vision mechanisms, *Journal of the Optical Society of America, A, Optics, Image & Science*, Vol. 7, No. 5, pp. 923−932, May, 1990.

[4] N. Otsu, A threshold selection method from gray-level histograms, *Automatica*, Vol. 11, pp. 23−27, 1975.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40, No. 4, pp. 834−848, April, 2018.

[6] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 4700−4708.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, Munich, Germany, 2015, pp. 234−241.

[8] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y Hammerla, B. Kainz, B. Glocker, D. Rueckert, Attention u-net: Learning where to look for the pancreas, *1st Conference on Medical Imaging with Deep Learning (MIDL 2018)*, Amsterdam, The Netherlands, 2018, pp. 1−10.

[9] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, *Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis (DLMIA)*, Granada, Spain, 2018, pp. 3−11

[10] F. Chollet, Xception: Deep learning with depthwise separable convolutions, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, HI, USA, 2017, pp. 1251−1258.

[11] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: More features from cheap operations, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Seattle, WA, USA, 2020, pp. 1580−1589.

[12] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic Convolution: Attention Over Convolution

Kernels, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11027−11036.

[13] D. Li, J. Hu, C. Wang, X. Li, Q. She, L. Zhu, T. Zhang, Q. Chen, Involution: Inverting the inherence of convolution for visual recognition, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Nashville, TN, USA, 2021, pp. 12316−12325.

[14] B. Yang, G. Bender, Q. V. Le, J. Ngiam, Condconv: Conditionally parameterized convolutions for efficient inference, *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, BC, Canada, 2019, pp. 1305−1316.

[15] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-unet: Unet-like pure transformer for medical image segmentation, *European conference on computer vision*, Tel Aviv, Israel, 2022, pp. 205−218.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, Y. Zhou, *Transunet: Transformers make strong encoders for medical image segmentation*, arXiv preprint arXiv:2102.04306, February, 2021. https://arxiv.org/abs/2102.04306

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, arXiv preprint arXiv:2010.11929, October, 2020. https://arxiv.org/abs/2010.11929

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992−10002.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, California, USA, 2017, pp. 6000−6010.

[20] R. Azad, M. Heidari, M. Shariatnia, E. K. Aghdam, S. Karimijafarbigloo, E. Adeli, D. Merhof, Transdeeplab: Convolution-free transformer-based deeplab v3+ for medical image segmentation, *International Workshop on Predictive Intelligence In MEdicine*, Singapore, 2022, pp. 91−102.

[21] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, *International MICCAI Brainlesion Workshop*, Virtual Event, 2021, pp. 272−284.

[22] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, D. Zhang, Ds-transunet: Dual swin transformer u-net for medical image segmentation, *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, pp. 1−15, May, 2022.

[23] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, J. Li, Transbts: Multimodal brain tumor segmentation using transformer, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Strasbourg, France, 2021, pp. 109−119.

[24] S. Anwar, K. Hwang, W. Sung, Structured pruning of deep convolutional neural networks, *ACM Journal on Emerging Technologies in Computing Systems*, Vol. 13 No. 3, pp. 1−18, July, 2017.

[25] T. Zhang, S. Ye, K. Zhang, J. Tang, W. Wen, M. Fardad, Y. Wang, A systematic dnn weight pruning framework using alternating direction method of multipliers, *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, 2018, pp. 191−207.

[26] T. Xiang, C. Zhang, D. Liu, Y. Song, H. Huang, W. Cai, BiO-Net: learning recurrent bi-directional connections for encoder-decoder architecture, *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, Lima, Peru, 2020, pp. 74−84.

[27] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, *IEEE Transactions on Medical Imaging*, Vol. 39, No. 10, pp. 3008−3018, October, 2020.

[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, Salt Lake City, UT, USA, 2018, pp. 7132−7141.

[29] S. Woo, J. Park, J.-Y. Lee, I.-S. Kweon, Cbam: Convolutional block attention module, *Proceedings of the European conference on computer vision (ECCV)*, Munich, Germany, 2018, pp. 3−19.

[30] J. M. J. Valanarasu, V. M. Patel, Unext: Mlp-based rapid medical image segmentation network, *International conference on medical image computing and computer-assisted intervention*, Singapore, 2022, pp. 23−33.

[31] F. Yuan, Z. Zhang, Z. Fang, An effective CNN and Transformer complementary network for medical image segmentation, *Pattern Recognition*, Vol. 136, Article No. 109228, April, 2023.

[32] A. Roy, M. Saffar, A. Vaswani, D. Grangier, Efficient content-based sparse attention with routing transformers, *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 53−68, February, 2021.

[33] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, V. M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, *Proceedings of the 24th International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI 2021)*, Strasbourg, France, 2021, pp. 36−46.

[34] Z. Xia, X. Pan, S. Song, L. E. Li, G. Huang, Vision Transformer with Deformable Attention, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 4784−4793.

[35] A. Tragakis, C. Kaul, R. Murray-Smith, D. Husmeier, The Fully Convolutional Transformer for Medical Image Segmentation, *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023, pp. 3649−3658

[36] W. Yu, M. Luo, P, Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, S. Yan, MetaFormer is Actually What You Need

for Vision, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 10809−10819,

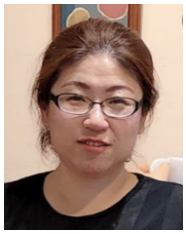
[37] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, *IEEE Transactions on Medical Imaging*, Vol. 38, No. 10, pp. 2281−2292, October, 2019.

[38] J. Ruan, S. Xiang, M. Xie, T. Liu, Y. Fu, MALUNet: A Multi-Attention and Light-weight UNet for Skin Lesion Segmentation, *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 1150−1156

Biographies

Ming Zhao received the master's degrees in computer aided Instruction from Huazhong Normal University, China in 2006, and he got Ph.D. in computer science and technology from Harbin Institute of Technology, China in 2015. He is currently a professor in Yangtze University, China. His research interests include computational intelligence, image and signal processing, pattern recognition etc. He is an IEEE Senior Member.

Bingxue Zhou enrolled at Yangtze University in 2018 and received her bachelor's degree in Computer Science and Technology in 2022. Her research interests include computational intelligence, image and signal processing, and pattern recognition.



Ling-Ju Hung is an associate professor in the Department of Creative Technologies and Product Design at National Taipei University. She has served as a guest editor for several international journals, including the *Journal of Computer and System Sciences*, *Algorithmica, Theoretical Computer Science*, *Journal of Combinatorial Optimization*, and *Discrete Applied Mathematics*. Currently, she is an associate editor for both the *Journal of Computer and System Sciences* and the *Journal of the Chinese Institute of Engineers*. In 2018, Dr. Hung applied her expertise in algorithms to address industrial challenges by joining AROBOT as a senior manager, where she led the Department of Algorithms. Under her leadership, the team developed a speech recognition engine that won an Excellent Industrial System Award at the Formosa Speech Recognition Challenge in 2018. Her research focuses on the design and analysis of algorithms for optimization problems with applications in networks.