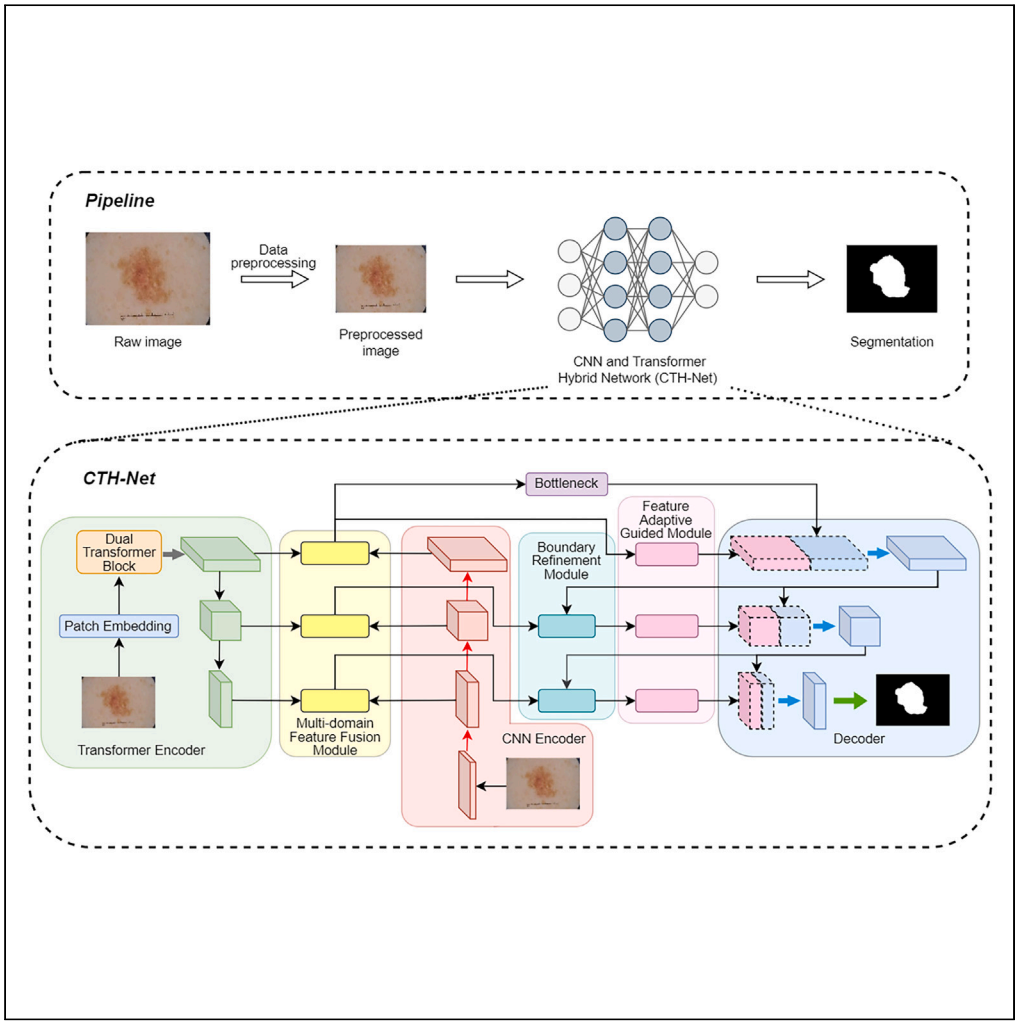
iScience

CellPress
OPEN ACCESS

Article

CTH-Net: A CNN and Transformer hybrid network for skin lesion segmentation



Yuhan Ding, Zhenglin Yi, Jiatong Xiao, Minghui Hu, Yu Guo, Zhifang Liao, Yongjie Wang

zfliao@csu.edu.cn (Z.L.)
yongjiewang@csu.edu.cn (Y.W.)

Highlights

We proposed an automatic skin lesion segmentation network, named CTH-Net

The multi-domain feature fusion module can effectively fuse dual encoder features

The boundary refinement module utilizes contextual information to refine boundaries

Experimental results on four skin lesion datasets show the effectiveness of CTH-Net

Ding et al., iScience 27, 109442
April 19, 2024
https://doi.org/10.1016/j.isci.2024.109442

iScience

CellPress
OPEN ACCESS

Article

CTH-Net: A CNN and Transformer hybrid network for skin lesion segmentation

Yuhan Ding,[1] Zhenglin Yi,[2] Jiatong Xiao,[2] Minghui Hu,[2] Yu Guo,[3,4] Zhifang Liao,[1,*] and Yongjie Wang[3,4,5,*]

SUMMARY

Automatically and accurately segmenting skin lesions can be challenging, due to factors such as low contrast and fuzzy boundaries. This paper proposes a hybrid encoder-decoder model (CTH-Net) based on convolutional neural network (CNN) and Transformer, capitalizing on the advantages of these approaches. We propose three modules for skin lesion segmentation and seamlessly connect them with carefully designed model architecture. Better segmentation performance is achieved by introducing SoftPool in the CNN branch and sandglass block in the bottleneck layer. Extensive experiments were conducted on four publicly accessible skin lesion datasets, ISIC 2016, ISIC 2017, ISIC 2018, and PH^2 to confirm the efficacy and benefits of the proposed strategy. Experimental results show that the proposed CTH-Net provides better skin lesion segmentation performance in both quantitative and qualitative testing when compared with state-of-the-art approaches. We believe the CTH-Net design is inspiring and can be extended to other applications/frameworks.

INTRODUCTION

One of the most common risks to human health around the world is skin disease,[1] for example, melanoma is extremely deadly, with a less than 15% five-year survival rate.[2] Studies have shown that when melanoma is diagnosed early, the survival rate is as high as 90%.[3] Dermoscopy, a non-invasive imaging tool, is frequently used to examine skin lesions and their surrounding regions for screening and diagnosing skin illnesses. Traditionally, manual inspection of malignant melanoma based on images generated by dermoscopy has been performed by specialist dermatologists, but it is considered a time-consuming and skill-intensive endeavor.

Computer-aided diagnosis (CAD) tools have been extensively used to help dermatologists with these issues by increasing diagnostic accuracy and generating reliable outcomes.[4] Building CAD systems depends heavily on the automatic skin lesion segmentation process.[5] This is because the segmented lesions can provide quantitative information such as location, shape, size, etc., which is very meaningful for increasing the effectiveness and precision of skin lesion diagnostics. However, automatic and accurate segmentation of skin lesions is still a complex and challenging task for the following reasons. In dermoscopic images, for instance, patient-specific characteristics including skin color, texture, lesion size, lesion location form, and the presence of various artifacts such as body hair, reflections, air bubbles, shadows, uneven illumination, and markings may change randomly.[6] Figure 1 displays typical difficult instances.

Early automatic lesion segmentation techniques were usually based on edge detection and thresholding methods[7] and active contour models.[8] It relies on carefully selected handcrafted features and efficient image pre-processing or post-processing algorithms, which lack robustness, resulting in inadequate performance in challenging scenes. Deep learning algorithms, on the other hand, can automatically and adaptively learn high-dimensional features,[9] evading the drawbacks of conventional techniques and increasingly taking over the field of skin lesion segmentation.

Convolutional neural network (CNN)-based structures have been proposed in recent years to enhance the accuracy of segmentation. Fully convolutional neural network (FCN)[10] is one of the early attempts at image segmentation. To prevent the loss of shallow information and obtain outstanding segmentation efficiency, Ronneberger et al.[11] presented a U-Net with "skip connections" according to FCN. The network's performance was then improved by some work that expanded U-Net or introduced new information, including ResU-Net,[12] U-Net++,[13] Attention U-Net,[14] V-Net,[15] etc. To better address the challenge of the skin lesion segmentation problem, Shahin et al.[16] embedded the pyramid pooling module into the deep skip connection to merge the global context information. Similar to this, Hu et al.[17] created a unique attention synergy network by merging spatial and channel attention processes to improve the discriminative performance of skin lesion segmentation. Despite being successful in a variety of computer vision tasks, CNN models cannot provide global context or

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China
[2]Departments of Urology, Xiangya Hospital, Central South University, Changsha 410008, China
[3]Department of Burns and Plastic Surgery, Xiangya Hospital, Central South University, Changsha 410008, China
[4]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410008, China
[5]Lead contact
*Correspondence: zfliao@csu.edu.cn (Z.L.), yongjiewang@csu.edu.cn (Y.W.)
https://doi.org/10.1016/j.isci.2024.109442

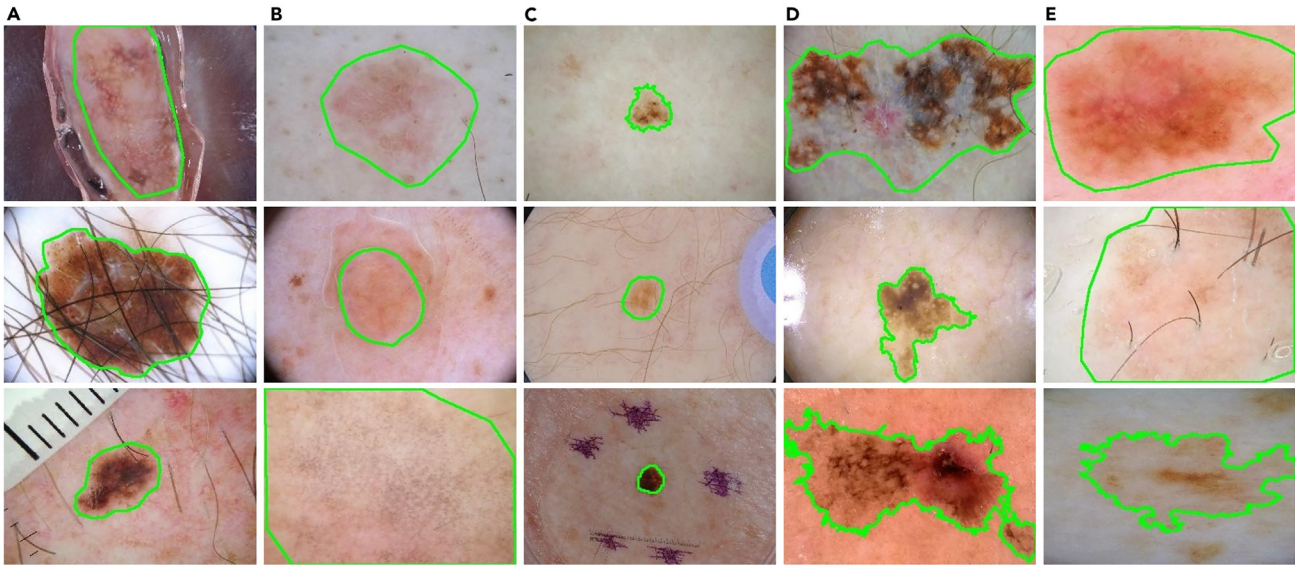


Figure 1. Commonly difficult situations in the public dermoscopic dataset ISIC 2018, including (A) lesions with artifact interference, (B) poor background contrast for the lesions, (C) small lesions, (D) irregularly shaped lesions Lesions, (E) Lesions with indistinct borders

The green outline is the ground truth.

long-distance relationships in images due to their constrained receptive fields and intrinsic inductive biases.[18] As a result, its ability to segment skin lesion images is restricted.

To address the limitations of CNN-based models in terms of global representation, Dosovitskiy et al.[19] proposed Vision Transformer (ViT) to capture global dependencies. The method first decomposes the image into token sequences and then injects positional embeddings into the token sequences when they are fed to the Transformer block. Compared to previous convolution-based algorithms, it achieves superior performance. The transformer excels at modeling global context, but it has trouble catching fine-grained details, particularly in medical images.[20] Because of the absence of spatial inductive bias when representing local information, pure Transformer-based segmentation networks like SETR[21] perform poorly. To solve the problem of weak local representation of the Transformer model, a method of constructing a hybrid CNN-Transformer network is proposed, and it encodes both global and local characteristics using the locality of CNN and the long-range dependency of the Transformer.[22] TransUNet, proposed by Chen et al.,[23] is the first model to combine Transformer and U-Net for the use of segmenting medical images. A large number of parameters and poor computational performance of TransUNet, however, are a drawback. Some subsequent methods, such as CoTr,[24] SegTran,[25] and TransBTS,[26] also use CNN-based networks as the backbone to supplement long-range dependencies with certain parts (such as encoders, bottlenecks, decoders, or skip connections) and achieve good results. However, due to the particularity of dermoscopic images, lesions often have different sizes, and the boundaries of some lesions are very blurred and difficult to define due to the lesions and surroundings having little contrast with one another.[27] In addition, artifacts such as ink blots, air bubbles, rulers, and hairs, which are abundantly present, may introduce additional noise. For the ability of skin lesion localization and fine boundary delineation, the study mentioned previously is by no means sufficient. It continues to be problematic to precisely segment skin lesions in dermoscopic images in this challenging setting.

To solve the problems listed previously, we propose an encoder-decoder model (CTH-Net) based on CNN and Transformer, which effectively utilizes the global long-range relation of Transformer and the local feature representation of CNN to achieve accurate skin lesion segmentation. By combining three well-designed core modules with an encoder-decoder structure, CTH-Net can better handle skin lesion segmentation tasks. Specifically, in CTH-Net, we first design a CNN-based encoder branch utilizing Res2Net50 for extracting fine-grained contextual features. Then a Transformer encoder branch with channel and spatial dual attention is designed using a dual transformer block to capture long-range dependency information. To better extract local spatial features, we introduce the SoftPool method in the CNN encoder, which can retain more useful information during the downsampling process, thereby improving the segmentation performance of fuzzy boundaries. For better cross-fusion enhancement of multi-domain features from two encoder branches, we design a multi-domain feature fusion module (MFFM). Next, we embed the boundary refinement module (BRM) and feature adaptive guided module (FAGM) in the skip connection. The former can achieve better performance in fine-grained boundary delineation by utilizing boundary information and neighborhood context information. The latter improves the learned lesion boundaries and better adaptively matches the feature distribution between the encoder and the decoder through a simple parallel convolution structure without increasing the number of parameters too much. Finally, the multi-scale encoder features after boundary refinement and feature adaptation are input into a progressive upsampling decoding layer to gradually obtain the final segmentation mask. Additionally, we offer a sandglass block that creates a quick connection among linear high-dimensional representations in the bottleneck layer to lower the number of parameters and better optimize network training. The following is a summary of this paper's main contributions.

(1) We propose an automatic skin lesion segmentation network called CTH-Net. In the encoder of CTH-Net, we use parallel dual-encoder branches instead of the traditional single-branch encoder structure. A Transformer branch and a CNN branch make up the dual encoder. The CNN encoder based on Res2Net and SoftPool[28] is mainly used to extract rich local spatial features. To segment skin lesions, the Transformer branch with a dual attention mechanism is utilized to gather global context information.
(2) We propose a multi-domain feature fusion module (MFFM). It combines self-attention and multi-domain fusion mechanism, which can realize feature complementation and fusion between CNN and Transformer. The segmentation accuracy is further improved by enhancing important information in both feature maps and suppressing insignificant features.
(3) We propose a boundary refinement module (BRM) and feature adaptive guided module (FAGM), which is embedded in skip connections. The former can achieve better performance in fine-grained boundary delineation by utilizing boundary information and neighborhood context information. The latter can learn and improve mismatched lesion boundaries while reducing the difference in features between the encoder and decoder.
(4) On four publicly accessible skin lesion datasets, extensive experimental findings show the efficacy and superiority of our proposed CTH-Net compared to competing approaches.

Related work

CNN-based segmentation networks

Long et al.[10] proposed a fully convolutional network for image semantic segmentation, which is the pioneering work of deep learning in the field of semantic segmentation. The FCN framework was further improved by Ronneberger et al.[11]'s unique convolutional segmentation network, known as U-Net, which included skip connections in every level of the encoder-decoder module. It achieves excellent performance in the segmentation of medical images. Although the U-shaped structure based on the encoder-decoder is simple, it exhibits powerful performance and is widely used in different image segmentation fields. In U-Net++,[29] the idea of deep supervision is introduced by adding dense connections to the U-Net network. At the same time, more skip connection paths and upsampling convolutional blocks are added to bridge the semantic gap between the encoder and decoder. Oktay et al.[14] proposed an Attention U-Net by generating gating signals to emphasize the attention to different spatial location features. It adds an attention submodule to each decoder layer to help the model learn more accurately how to distinguish foreground from background. With the use of atrous convolution, DeeplabV3+[30] provides an encoder-decoder structure to broaden the receptive field and increases the precision of semantic segmentation. MultiResUNet[31] mainly addresses two common problems in medical image segmentation: scale diversity and the semantic gap in the fusion between different levels of features. The MultiRes module and Res Path were proposed to solve it and achieved excellent performance in multimodal image segmentation. Feng et al.[32] proposed a new contextual pyramid fusion network (CPFNet) based on a U-shaped structure to fuse multi-scale context information by combining two pyramid modules. Karaali et al.[33] propose a new deep-learning pipeline that combines the efficiency of residual dense network blocks and residual squeeze and excitation blocks to achieve superior performance on retinal vessel segmentation. The consistent perception generative adversarial network (CPGAN[34]) is a semi-supervised consistent perception generative adversarial network that achieves accurate segmentation of stroke lesion areas by effectively capturing multi-scale feature information and introducing a consistent perception strategy. The symmetric driven generative adversarial network (SD-GAN)[35] models various symmetric changes in the normal brain in an unsupervised manner, completing the segmentation of brain tumors in magnetic resonance (MR) images and reducing reliance on manually labeled data.[36]

Transformer-based segmentation networks

Despite the positive outcomes that CNN models have produced, these techniques frequently perform poorly because their small receptive fields make it difficult to model long-range dependencies. Transformer-based models and CNN and Transformer hybrid models have recently gained more traction in the field of medical segmentation of images than CNN-only techniques. Chen et al.[23] proposed the first model that integrates the self-attention mechanism into medical image segmentation tasks: TransUNet, which brings together the benefits of Transformer and U-Net. For accurate localization, the decoder mixes the Transformer-encoded features with high-resolution CNN feature maps after upsampling them. The first entirely Transformer-based U-architecture is called Swin U-Net.[37] With the use of a patch extension layer and skip connections, a decoder upsamples the recovered contextual features and fuses them with multi-scale data from an encoder to restore the feature map's spatial resolution for future segmentation prediction. TransFuse[38] effectively captures global relationships and low-level spatial features in a shallower manner by combining Transformer and CNN in tandem. The multi-level characteristics of the two branches are effectively fused using a unique feature fusion technique. Azad et al.[39] reformulated the self-attention mechanism to extract spatial and channel relationships to cover all feature dimensions and redesigned skip connection paths to ensure feature reusability and enhance localization capabilities. A recent method called HiFormer[22] connects CNN and Transformer for medical image segmentation in an effective way. A Swin Transformer module and a CNN-based encoder are used to create two multi-scale feature representations that carefully combine local and global data. By introducing dynamically deformable convolutions in the CNN branch and combining it with the Transformer branch with a shift window adaptive complementary attention module, CiT-Net[40] combines the advantages of CNN and Transformer and performs well in medical image segmentation.

Skin lesion segmentation networks

Methods for segmenting skin lesions traditionally focus primarily on extracting and recognizing low-level image characteristics. CNN-based algorithms do not need detailed image definitions, in contrast to conventional feature-based techniques. Tang et al.[41] proposed a separable U-Net based on random weight averaging for skin lesion segmentation. It can significantly increase the pixel-level discriminative representation capability of fully convolutional networks by capturing contextual feature channel correlation and higher semantic feature information. Dai et al.[42] created a brand-new network for residual encoding and decoding on many scales to segment skin lesions, which can efficiently segment various lesions accurately and reliably. Using a coarse-to-fine approach, Liu et al.[27] developed a neighborhood contextual refinement network to accomplish accurate skin lesion segmentation. To localize skin lesions and define lesion boundaries, it comprises a shared encoder and two distinct but related decoders. Efficient group enhanced UNet (EGE-UNet)[43] combines the group multi-axis Hadamard product attention module (GHPA) and group aggregation bridge module (GAB) in a lightweight manner based on U-Net, achieving excellent performance in skin lesion segmentation.

Wu et al.[44] presented a feature adaptive Transformer network based on the encoder-decoder architecture, known as FAT-Net, to better capture local detail information and long-range relationships. It incorporates a further Transformer branch to effectively gather data on global context and distant dependencies. J. Wang et al.[45] integrated a boundary attention gate into Transformer, which not only allows the network as a whole to efficiently model global long-range dependencies not only through Transformer but also captures more local detail prior knowledge. The pyramid Transformer inter-pixel correlation module and the local neighborhood metric learning module were created by Cao et al.[46] as part of their innovative technique for learning and modeling inter-pixel correlation from global and local factors. The majority of earlier works either use Transformers with restricted local feature representation or CNNs without global features for feature extraction, which lacks an effective complementarity between long-distance dependencies and local features. In the hybrid model, for the multi-domain features extracted in different fields, only a simple feature fusion mechanism is used, which cannot guarantee the consistency of features between different scales. In the information transmission of the codec, contextual information cannot be used to describe and guide the fuzzy boundaries of skin lesions in a fine-grained manner. Therefore, we propose an encoder-decoder framework CTH-Net based on CNN and Transformer, which effectively utilizes Transformer's global long-range relation and CNN's local feature representation for an accurate skin lesion segmentation task. For better cross-fusion enhancement of multi-domain features from two encoder branches, we design a multi-domain feature fusion module. Next, we embed the boundary refinement module and feature adaptive guided module in the skip connection. To accurately segment skin lesion boundaries, they can learn from and improve mismatched lesion boundaries while narrowing the feature gap that exists between the encoder and decoder.

RESULTS

Datasets

Using four publicly available skin lesion segmentation datasets, we undertake comprehensive experiments: ISIC 2016,[47] ISIC 2017,[48] ISIC 2018,[49] and PH^2 [50] to demonstrate the effectiveness of our method. The International Skin Imaging Collaboration (ISIC) archive offers ISIC 2016, ISIC 2017, and ISIC 2018. The International Symposium on Biomedical Imaging (ISBI) sponsored three challenge datasets for "skin lesion analysis toward melanoma detection" in 2016, 2017, and 2018, respectively. The dermatology department of Hospital Pedro Hispano provides the PH^2 dataset, which is the other dataset (Matosinhos, Portugal). The four datasets' combined image counts and data partitions are as follows:

ISIC 2016: In the ISIC 2016 dataset, there are 1,279 RGB skin lesion images, 900 of which are used for training and 379 for testing.

ISIC 2017: There are 2,750 RGB skin lesion images in the ISIC 2017 dataset; 2,000 of them are used for training, 150 for validation, and the remaining 600 for testing.

ISIC 2018: A total of 3,694 RGB skin lesion images make up the ISIC 2018 dataset, of which 2,594 are utilized for training, 100 are used for validation, and the remaining 1,000 are used for testing. We re-partition ISIC 2018 into a training set (70%), validation set (10%), and test set at random (20%).

PH^2: The 200 8-bit RGB color dermoscopic images in the PH^2 dataset have a resolution of 768 × 560 pixels. We choose 140 images at random as the training set, 20 images for the validation set, and 40 images for the test set.

Evaluation metrics

To assess the effectiveness of various algorithms, we employed seven standard semantic segmentation measures, including precision, recall, dice socre, Jaccard index, accuracy, frequency weighted intersection over union (FWIoU), and 95% Hausdorff distance (95%HD). The definitions are shown in Equations 1, 2, 3, 4, 5, 6, 7, and 8:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{(Equation 1)}$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{(Equation 2)}$$

Table 1. Comparison of skin lesion segmentation performance of different networks on ISIC 2016

Methods	Dice Score	Jaccard Index	Accuracy	Params(M)
Rank #1	0.910	0.843	0.953	–
Rank #2	0.897	0.829	0.949	–
Rank #3	0.895	0.822	0.952	–
Rank #4	0.885	0.811	0.944	–
Rank #5	0.888	0.810	0.946	–
DeepLabV3+[30]	0.926	0.843	0.952	59.5
Swin Unet[37]	0.935	0.857	0.954	27.2
nnUnet[51]	0.938	0.868	0.955	29.9
HiFormer[22]	0.943	0.867	0.960	**25.5**
DAE-Former[39]	0.948	0.876	0.962	48.1
Ours	**0.954**	**0.887**	**0.968**	27.4

The best outcomes are highlighted in bold.

$$\text{Dice Socre} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP} \quad \text{(Equation 3)}$$

$$\text{Jaccard Index} = \frac{TP}{TP + FN + FP} \quad \text{(Equation 4)}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{(Equation 5)}$$

$$\text{FWIoU} = \frac{(TP + FN) \cdot TP}{(TP + TN + FP + FN) \cdot (TP + FP + FN)} \quad \text{(Equation 6)}$$

$$\text{hd}_{95}(\text{G}, \text{P}) = \max\left(\min\left(d(g, p)\right), 1 - 0.95\right) \quad \text{(Equation 7)}$$

$$95\%\ \text{HD}(\text{G}, \text{P}) = \max(\text{hd}_{95}(\text{G}, \text{P}), \text{hd}_{95}(\text{P}, \text{G})) \quad \text{(Equation 8)}$$

where TP, TN, FP, FN represent true positive, true negative, false positive, and false negative, respectively. G, P represent the boundary point sets of ground truth and predicted mask, respectively, $g \in$ G, $p \in$ P, and $d(g, p)$ represent the Euclidean distance from point g to point p, 0.95 represents a 95% confidence level, i.e., the percentage of the distance considered. The two most significant segmentation evaluation factors for rating the competitors in the ISIC Challenge are the dice score and the Jaccard index. We will thus give the dice score and Jaccard index more weight when statistically measuring network performance.

Results on the ISIC 2016 and ISIC 2017 dataset

Quantitative study

On the predefined ISIC 2016 and ISIC 2017 datasets, Tables 1 and 2 objectively compare CTH-Net's performance to that of the top five competing approaches and five popular semantic segmentation models. The competition's top five finishers are based on the results of the official leaderboard. Compared with other state-of-the-art methods, CTH-Net always keeps ahead in various indicators. In ISIC 2016, compared with the first place in the challenge, CTH-Net significantly improved the dice socre and Jaccard index from 0.910 to 0.843 to 0.954 and 0.887, respectively. At the same time, compared with the most competitive DAE-Former, our model improves dice score, Jaccard index, and accuracy by 0.6%, 1.1%, and 0.6%, respectively. Compared with HiFormer with a size of 25.5M, our method increases the dice score and Jaccard index by 1.1% and 2.0%, respectively, while only increasing the number of parameters by 1.9M. This shows that CTH-Net has achieved a good balance between computing resources and performance. In ISIC 2017, there are more types and more complex skin lesions, with blurred borders and indistinguishable from the background. Compared with the first-ranked solution, CTH-Net improved the dice socre and Jaccard index from 0.849 and 0.765 to 0.934 and 0.819, respectively. Compared with the most competitive nnUnet and DAE-Former, our method improves dice score, Jaccard index, and accuracy by 1.2%, 1.0%, and 0.9%, respectively. It is noteworthy that compared with DAE-Former, the number of parameters of our method is reduced by 43.0%. Compared with the DAE-Former of the pure Transformer architecture, the excellent performance of CTH-Net benefits from the design of parallel dual encoders, which can exactly segment the boundaries of skin lesions by combining local context features while capturing global context information.

Table 2. Comparison of skin lesion segmentation performance of different networks on ISIC 2017

Methods	Dice Score	Jaccard Index	Accuracy	Params(M)
Rank #1	0.849	0.765	0.934	–
Rank #2	0.847	0.762	0.932	–
Rank #3	0.844	0.760	0.934	–
Rank #4	0.842	0.758	0.934	–
Rank #5	0.839	0.754	0.931	–
DeepLabV3+[30]	0.911	0.776	0.950	59.5
TransUNet[23]	0.919	0.790	0.955	105.3
nnUnet[51]	0.921	0.801	0.957	29.9
FAT-Net[44]	0.919	0.804	0.953	30.0
DAE-Former[39]	0.922	0.809	0.955	48.1
Ours	**0.934**	**0.819**	**0.966**	**27.4**

The best outcomes are highlighted in bold.

Qualitative study

The results of the visual segmentation using various models on the ISIC 2016 and ISIC 2017 are qualitatively compared in Figures 2 and 3. In ISIC 2016, we conducted a visual comparison of several approaches for several common hard circumstances, such as blurring boundaries, poor background contrast, and the presence of artifacts. We selected DeepLabV3+, Swin Unet, nnUnet, HiFormer, and DAE-Former as comparisons. It can be seen that compared with the other five competitors, our method has achieved superior segmentation results in skin lesion segmentation. Even in the case where the lesion is light in color and indistinguishable from the background (the image in the first row of

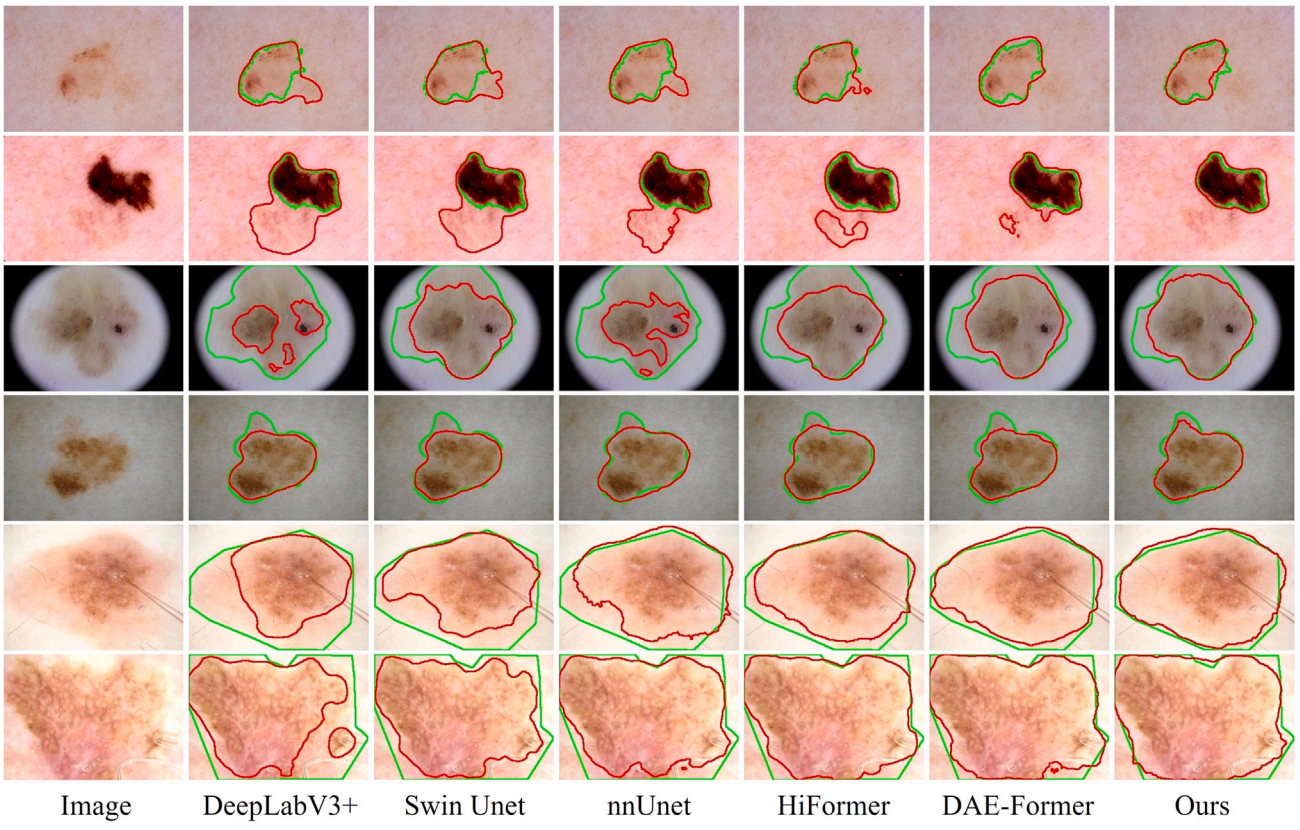


Figure 2. Visual comparison with the state-of-the-art on ISIC 2016

The red outline represents the segmentation outcome of the corresponding algorithm, and the green outline represents the ground truth.

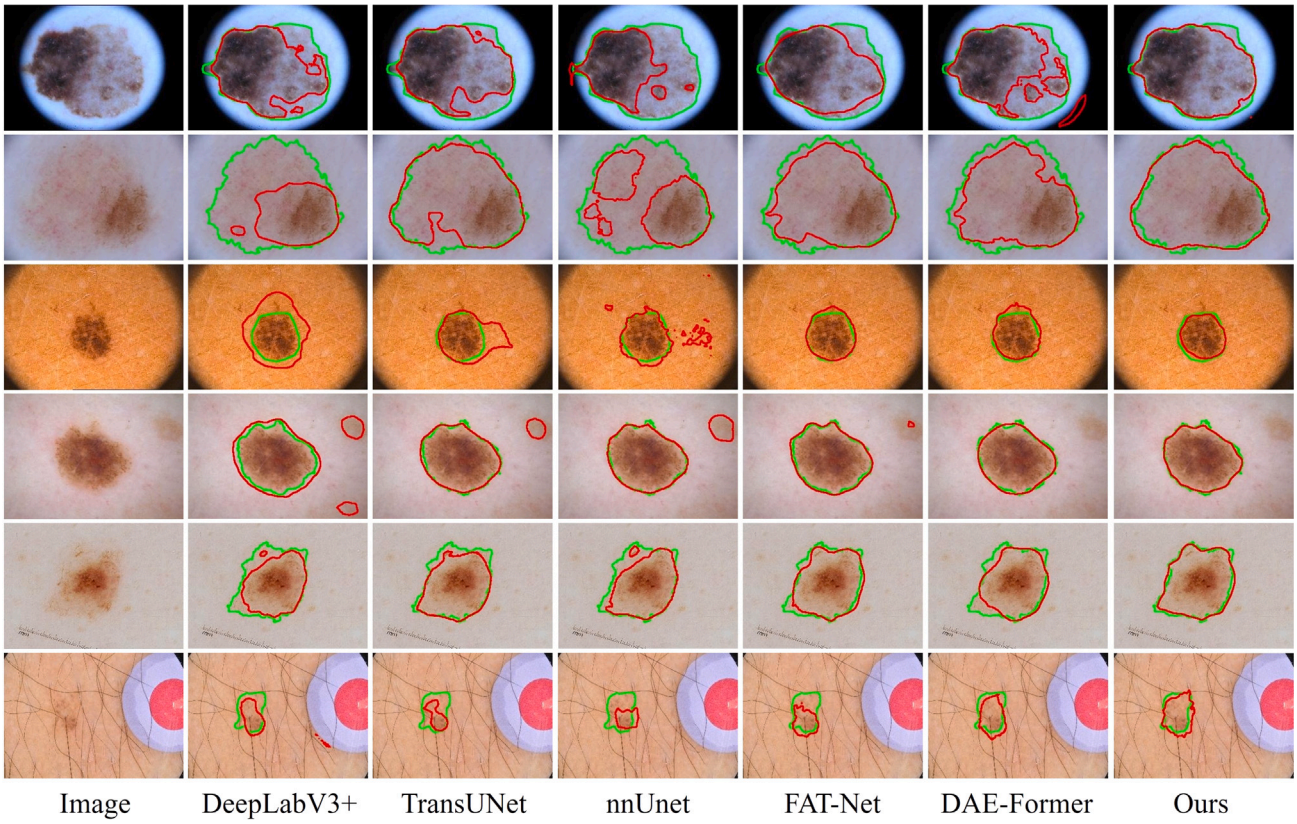


Figure 3. Visual comparison with the state-of-the-art on ISIC 2017

The red outline represents the segmentation outcome of the corresponding algorithm, and the green outline represents the ground truth.

Figure 2), our method can still accurately locate the boundary, which is very close to the real situation. While at ISIC 2017, we utilized DeepLabV3+, TransUNet, nnUnet, FAT-Net, and DAE-Former with our method to generate visual comparison results for typical challenging cases. Compared with FAT-Net, which is also specially designed for skin lesion segmentation, CTH-Net has achieved better performance in lesion identification and fine-grained boundary delineation in the face of dermoscopic images with artifacts and small lesions (the image in the sixth row of Figure 3). The efficiency of the suggested strategy in the task of skin lesion segmentation is completely demonstrated by these outcomes. In contrast to FAT-Net, our proposed multi-domain feature fusion module is used to effectively fuse and complement the information extracted from CNN and Transformer branches.

Results on the ISIC 2018 dataset

Quantitative study

Table 3 quantitatively shows the comparison of skin lesion segmentation performance between CTH-Net and 10 mainstream segmentation algorithms on ISIC 2018, including U-Net, U-Net++, Attention U-Net, DeepLabV3+, TransUNet, Swin Unet, nnUnet, FAT-Net, HiFormer, and DAE-Former. To ensure a fair comparison, all competitors in our comparative experiments run on the same computing environment and undergo the same data processing, and the scores of all evaluation indicators are obtained via 5-fold cross-validation. Based on the classic U-Net, U-Net++ introduces more upsampling nodes and skip connections to achieve better results. To extract multi-scale features, DeepLabV3+ combines dilated convolution and inception structures based on the encoder-decoder structure and suggests an improved atrous spatial pyramid pooling module. As a result, its performance is better than U-Net and its variations. The Transformer overcomes the relatively limited shortcomings of CNN in modeling global information. Compared with pure Transformers such as Swin Unet, CTH-Net is robust to noise. Compared with the most competitive methods such as nnUnet, FAT-Net, and HiFormer, our approach has produced the best results across all indicators. Especially in terms of dice score, Jaccard index, accuracy, and FWIoU, it reached 0.959, 0.893, 0.975, and 0.952, respectively, and compared with HiFormer at 95% Hausdorff distance, it increased by 0.606 mm. Compared with DeepLabV3+ based on the CNN method, our method reduces the number of parameters by 32.1M while increasing the dice score and Jaccard index by 2.0% and 5.0%, respectively. This once again shows that CTH-Net achieves a good balance between the number of model parameters and segmentation performance. The results of the comparative experiments clearly show how successful the dual encoder design and multi-domain

Table 3. Skin lesion segmentation performance of different networks on ISIC 2018

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD	Params(M)
U-Net[11]	0.910 ± 0.013	0.903 ± 0.016	0.926 ± 0.005	0.820 ± 0.004	0.956 ± 0.005	0.924 ± 0.006	8.238 ± 2.081	32.9
U-Net++[13]	0.916 ± 0.010	0.904 ± 0.017	0.933 ± 0.005	0.827 ± 0.013	0.960 ± 0.004	0.929 ± 0.007	6.984 ± 1.086	34.9
Attention U-Net[14]	0.909 ± 0.015	0.923 ± 0.018	0.934 ± 0.009	0.836 ± 0.022	0.958 ± 0.008	0.925 ± 0.012	5.287 ± 1.027	33.3
DeepLabV3+[30]	0.908 ± 0.014	0.928 ± 0.007	0.939 ± 0.003	0.843 ± 0.013	0.964 ± 0.002	0.934 ± 0.004	5.212 ± 1.517	59.5
TransUNet[23]	0.914 ± 0.012	0.929 ± 0.007	0.941 ± 0.004	0.849 ± 0.013	0.964 ± 0.003	0.934 ± 0.004	4.308 ± 1.308	105.3
Swin Unet[37]	0.922 ± 0.007	0.927 ± 0.007	0.947 ± 0.006	0.857 ± 0.008	0.968 ± 0.007	0.942 ± 0.011	3.953 ± 1.790	27.2
nnUnet[51]	0.929 ± 0.010	0.944 ± 0.010	0.953 ± 0.005	0.877 ± 0.012	0.969 ± 0.003	0.943 ± 0.004	3.485 ± 1.712	29.9
FAT-Net[44]	0.927 ± 0.011	0.943 ± 0.013	0.952 ± 0.003	0.875 ± 0.002	0.969 ± 0.004	0.944 ± 0.007	3.801 ± 1.258	30.0
HiFormer[22]	0.941 ± 0.007	0.938 ± 0.007	0.954 ± 0.003	0.883 ± 0.008	0.969 ± 0.002	0.943 ± 0.003	2.160 ± 0.289	**25.5**
DAE-Former[39]	0.931 ± 0.005	0.943 ± 0.011	0.952 ± 0.004	0.878 ± 0.009	0.969 ± 0.004	0.943 ± 0.006	2.750 ± 0.922	48.1
Ours	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**	27.4

The best outcomes are highlighted in bold. Data are represented as mean ± std.

feature fusion module in CTH-Net are at accurately segmenting skin lesions using the global distant relationship of the Transformer and the local feature representation of CNN.

We conducted descriptive statistics on the two important indicators of the dice score and Jaccard index at ISIC 2018. Figure 4 shows the boxplots of all the important indicators of the above models. It can be shown that CTH-Net has the highest median value and the best score distribution, demonstrating the superiority of our method over other comparable networks.

We use the frequently used paired t test for evaluation to confirm the validity of the performance increase of the suggested strategy over competing methods. Table 4 displays the analysis findings for the four performance evaluation indicators we used for statistical analysis (dice score, Jaccard index, FWIoU, and 95% HD). The fact that all of the paired t tests' p values are less than 0.05 clearly shows that the proposed model's performance increase is statistically significant. As a result, the viability and dependability of the suggested CTH-Net are further confirmed.

Qualitative study

On the ISIC 2018 dataset, the performance of various networks is qualitatively compared in Figure 5. Typical difficult samples include tiny lesions, artifact interference, blurred borders, blurred lesions, and blurred borders. The images in the first row of Figure 5 display the segmentation outcomes of various models in the presence of boundary-blurred images with low contrast. Whether the network can extract richer feature representations will determine how well it can segment blurred objects. U-Net, U-Net++, and DeepLabV3+ made wrong predictions for the transitional color difference regions around the lesion because they all failed to effectively identify the boundary between the lesion and the background. Our CTH-Net shows the best performance in low-contrast fuzzy boundary recognition, thanks to our boundary refinement module and feature adaptive guided module used in skip connections. Without adding too many parameters, it can improve the learned lesion border and more adaptively match the feature distribution between the encoder and decoder. The segmentation outcomes

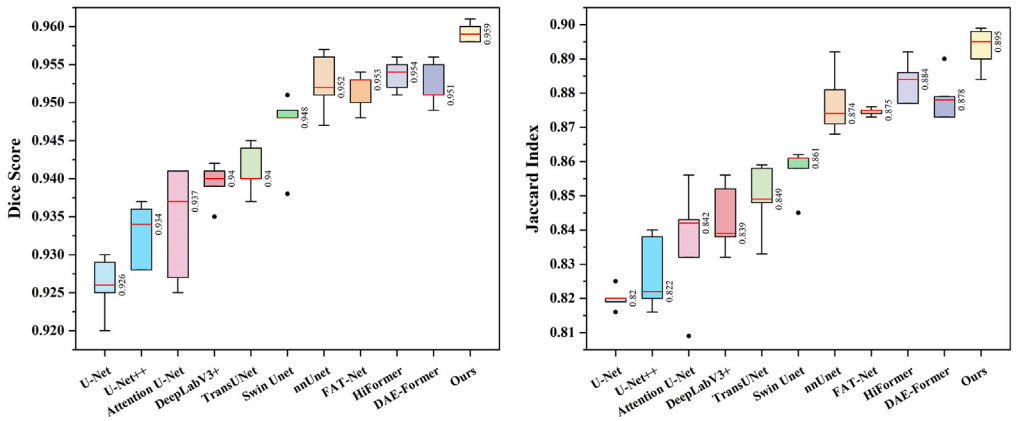


Figure 4. Performance comparison of different networks on ISIC 2018

Boxes in different color indicate the score range of different models, the line inside each box represents the median value, box limits include interquartile ranges Q1 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 times the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers.

Table 4. Statistical analysis (p value) of the proposed CTH-Net and other state-of-the-art methods

Methods	Dice Score	Jaccard Index	FWIoU	95%HD
U-Net vs. Proposed	1.360E-04	2.480E-06	1.141E-04	9.065E-04
U-Net++ vs. Proposed	1.794E-04	1.025E-04	1.108E-03	1.666E-04
Attention U-Net vs. Proposed	2.027E-03	1.690E-03	3.645E-03	1.163E-03
DeepLabV3+ vs. Proposed	1.376E-04	8.033E-04	1.025E-03	3.732E-03
TransUNet vs. Proposed	2.921E-04	2.018E-04	2.333E-04	3.750E-03
Swin Unet vs. Proposed	4.356E-03	1.320E-03	2.786E-02	2.238E-02
nnUnet vs. Proposed	2.766E-02	3.381E-02	9.548E-03	4.803E-02
FAT-Net vs. Proposed	6.254E-03	2.861E-03	5.614E-03	1.373E-02
HiFormer vs. Proposed	1.097E-03	1.786E-03	4.221E-03	1.451E-02
DAE-Former vs. Proposed	1.543E-02	1.747E-02	7.951E-03	2.849E-02

of various networks for large lesions with clear color differences inside are shown in the image in the fifth row of Figure 5. It can be seen that most of the results of the network are seriously under-segmented, and the features cannot be understood from the perspective of combining the global and the local, and the ability to capture the overall shape is poor. Through the combination of dual encoder and MFFM, the multi-domain features of CNN and Transformer can be effectively fused to obtain more comprehensive and compact fusion features, to get the closest ground truth and the best segmentation results.

Overall, our approach outperforms rivals' segmentation techniques on ISIC 2018, notably for difficult instances with weak background contrast and hazy boundary lines.

Results on the PH^2 dataset

Quantitative study

The segmentation performance of various networks on the PH^2 dataset is quantitatively displayed in Table 5. CTH-Net, DAE-Former, and HiFormer have the top three comprehensive results in Table 5. Owing to the utilization of different useful modules such as MFFM, BRM, and FAGM, CTH-Net's dice score, Jaccard index, accuracy, FWIoU, and 95%HD reached 0.960, 0.908, 0.971, 0.945, and 0.785 mm, respectively, which is substantially superior over alternative networks. Compared with the most competitive DAE-Former, our method improves the dice score and Jaccard index by 0.5% and 1.1%, respectively, and the 95% HD is reduced by 0.607cmm. Experimental results show that CTH-Net also performs well on small datasets.

We performed descriptive statistics on two important indicators on PH^2: dice score and Jaccard index. Figure 6 shows the boxplots of all the important indicators of the aforementioned models. It is clear that CTH-Net has the highest median value and the best score distribution, and the deviation is minimal, demonstrating the superiority of our method over other networks.

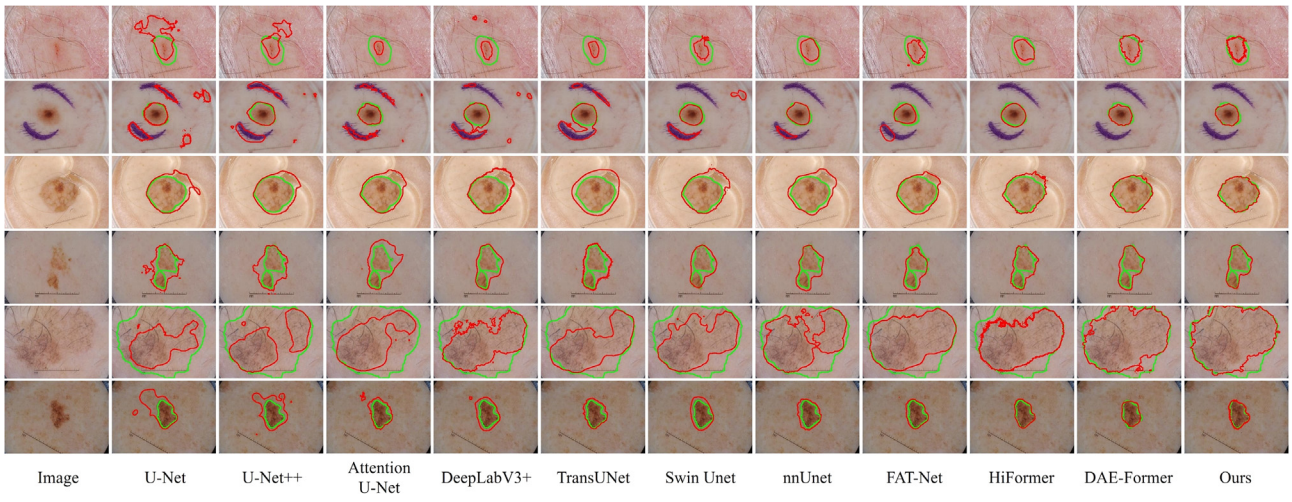


Figure 5. Visual comparison with the state-of-the-art on ISIC 2018

The red outline represents the segmentation outcome of the corresponding algorithm, and the green outline represents the ground truth.

Table 5. Skin lesion segmentation performance of different networks on PH²

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD	Params(M)
U-Net[11]	0.924 ± 0.055	0.913 ± 0.028	0.918 ± 0.040	0.840 ± 0.063	0.938 ± 0.031	0.900 ± 0.036	6.448 ± 0.760	32.9
U-Net++[13]	0.928 ± 0.072	0.921 ± 0.024	0.924 ± 0.035	0.855 ± 0.063	0.944 ± 0.028	0.907 ± 0.037	5.294 ± 1.438	34.9
Attention U-Net[14]	0.941 ± 0.048	0.917 ± 0.014	0.923 ± 0.038	0.861 ± 0.055	0.941 ± 0.032	0.906 ± 0.038	4.480 ± 1.127	33.3
DeepLabV3+[30]	0.946 ± 0.043	0.915 ± 0.022	0.929 ± 0.023	0.865 ± 0.040	0.948 ± 0.019	0.914 ± 0.021	4.070 ± 0.559	59.5
TransUNet[23]	0.936 ± 0.046	0.929 ± 0.017	0.936 ± 0.020	0.869 ± 0.047	0.956 ± 0.014	0.922 ± 0.018	3.247 ± 1.233	105.3
Swin Unet[37]	0.956 ± 0.021	0.918 ± 0.020	0.936 ± 0.036	0.876 ± 0.038	0.955 ± 0.025	0.924 ± 0.031	3.885 ± 1.666	27.2
nnUnet[51]	0.949 ± 0.041	0.925 ± 0.017	0.937 ± 0.026	0.877 ± 0.051	0.956 ± 0.019	0.925 ± 0.025	2.851 ± 0.560	29.9
FAT-Net[44]	0.949 ± 0.041	0.928 ± 0.015	0.938 ± 0.026	0.879 ± 0.047	0.957 ± 0.018	0.927 ± 0.024	2.862 ± 1.605	30.0
HiFormer[22]	0.957 ± 0.022	0.924 ± 0.021	0.943 ± 0.019	0.886 ± 0.022	0.958 ± 0.016	0.923 ± 0.024	1.845 ± 0.742	**25.5**
DAE-Former[39]	0.958 ± 0.012	0.936 ± 0.024	0.955 ± 0.007	0.897 ± 0.026	0.967 ± 0.007	0.937 ± 0.011	1.392 ± 0.939	48.1
Ours	**0.966±0.01**	**0.939±0.017**	**0.96±0.003**	**0.908±0.022**	**0.971±0.005**	**0.945±0.008**	**0.785±0.584**	27.4

The best outcomes are highlighted in bold. Data are represented as mean ± std.

Qualitative study

Using the results of the visual segmentation, Figure 7 qualitatively analyzes the performance of various networks on PH². The segmentation outcomes of various networks for skin lesions, when there is hair interference in the dermoscopic image, are shown in the image in the third row of Figure 7. Most of the contrast methods will mistake the surrounding hair for the lesion. The lesion's border is still precisely delineated by CTH-Net, which is also quite near to reality. The segmentation outcomes from several networks in the presence of hazy borders are depicted in line 6 of Figure 7. CTH-Net performs better at segmentation even when there is very little difference between the disease area and the surrounding healthy skin. The size of the PH² dataset is very small, containing only 200 dermoscopic images, we used 140 images as the training set and 40 images as the testing set. Despite having fewer samples and more challenging training, the proposed CTH-Net performed exceptionally well in terms of evaluation index scores and visual segmentation outcomes. This demonstrates once more how very efficient and effective CTH-Net is at segmenting skin lesions.

Cross-validation on ISIC 2018 and PH²

We performed cross-validation on ISIC 2018 and PH² to further confirm the generalization capability of CTH-Net on various data distributions. Table 6 displays how well various models generalize when cross-validated using ISIC 2018 and PH². "ISIC 2018 → PH²" indicates the performance tested on the full PH² dataset using the model obtained in ISIC 2018. And "PH² → ISIC 2018" shows how well the model developed in PH² performed on 40 randomly chosen ISIC 2018 test data. In the comparative experiment of PH², 40 images were extracted from the PH² dataset as the test set, which accounted for 20% of the dataset. Table 6 demonstrates that CTH-Net outperforms other comparison models in terms of generalization performance. Among them, the model obtained in ISIC 2018 has shown good generalization ability on the PH² dataset, while the model obtained in PH² has a poor generalization effect on the ISIC 2018 dataset. This is so that

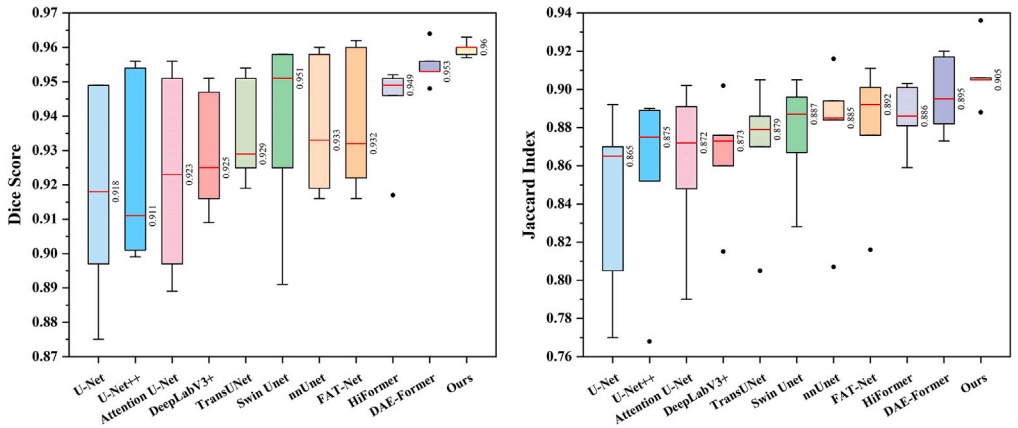


Figure 6. Performance comparison of different networks on PH²

Boxes in different color indicate the score range of different models, the line inside each box represents the median value, box limits include interquartile ranges Q1 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 times the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers.

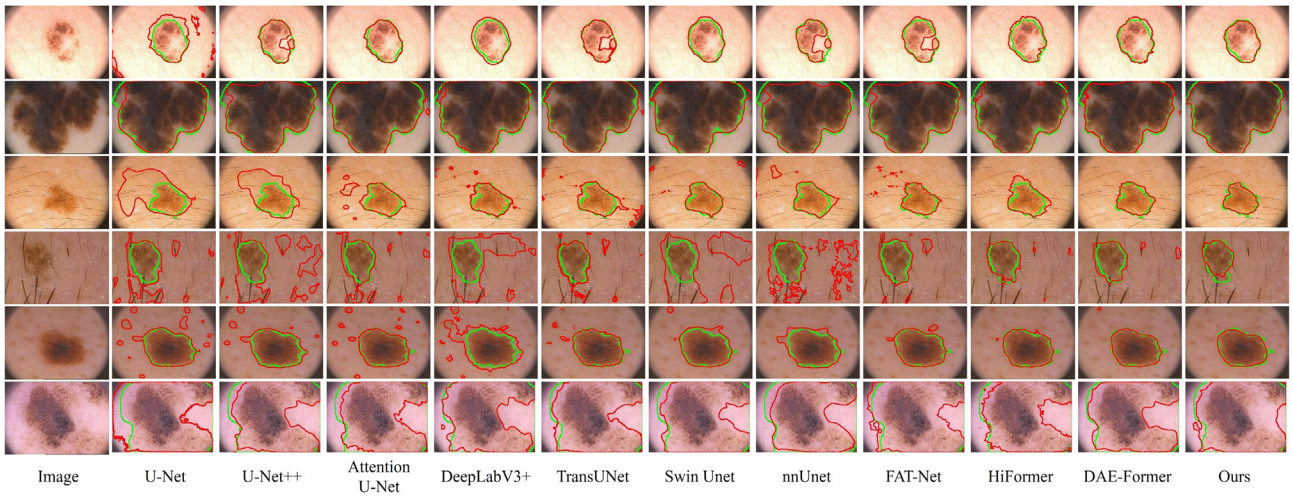


Figure 7. Visual comparison with the state-of-the-art on PH^2

The red outline represents the segmentation outcome of the corresponding algorithm, and the green outline represents the ground truth.

the model can better understand the properties of many forms of skin lesions. The ISIC 2018 dataset comprises a total of 3,694 images of skin lesions, covering a wide variety of skin lesion types. The PH^2 dataset only contains 200 dermoscopic images, and most of the lesions have obvious contrast with the background, and the segmentation difficulty is low, so it cannot be well generalized to test images with different distributions. The excellent performance of CTH-Net in the bidirectional generalization experiment, on the one hand, benefits from the fact that the CNN encoder based on Res2Net and SoftPool can extract rich local spatial features. The Transformer branch implemented by the dual attention mechanism can capture the global context information of skin lesion segmentation. On the other hand, it can learn and enhance the mismatched lesion boundary while minimizing the feature gap between the encoder and decoder due to the boundary refinement module and feature adaptive guided module contained in the skip connection. This is crucial to enhancing the model's capacity for generalization.

DISCUSSION

Ablation analysis

We conduct an extensive ablation analysis on CTH-Net to show the efficacy of various components in the proposed model, including the encoder network, dual encoder, key components, MFFM, bottleneck layer, and upsampling method. The ISIC 2018 dataset is used for all experiments, and 5-fold cross-validation is used to determine the average performance of each assessment indicator.

Table 6. Cross-validate the generalization ability of different methods on ISIC 2018 and PH^2

	ISIC 2018 → PH^2					PH^2 → ISIC 2018				
Methods	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
U-Net[11]	0.905	0.824	0.924	0.881	6.286	0.721	0.622	0.840	0.782	39.552
U-Net++[13]	0.917	0.830	0.934	0.885	5.827	0.794	0.631	0.853	0.796	36.548
Attention U-Net[14]	0.925	0.840	0.944	0.903	4.636	0.785	0.637	0.846	0.783	38.080
DeepLabV3+[30]	0.929	0.844	0.945	0.904	4.152	0.806	0.632	0.875	0.821	35.885
TransUNet[23]	0.931	0.852	0.948	0.910	2.649	0.803	0.646	0.861	0.802	40.141
Swin Unet[37]	0.940	0.859	0.955	0.919	3.400	0.791	0.659	0.864	0.806	39.345
nnUnet[51]	0.937	0.864	0.949	0.911	2.867	0.745	0.668	0.886	0.831	34.692
FAT-Net[44]	0.941	0.870	0.956	0.919	1.758	0.775	0.676	0.879	0.822	38.750
HiFormer[22]	0.936	0.871	0.949	0.908	2.638	0.818	0.685	0.870	0.811	33.162
DAE-Former[39]	0.947	0.877	0.960	0.927	2.494	0.845	0.681	0.906	0.861	**26.163**
Ours	**0.948**	**0.882**	**0.962**	**0.930**	**1.627**	**0.857**	**0.714**	**0.910**	**0.865**	32.031

The best outcomes are highlighted in bold.

Table 7. Performance comparison between different cnn encoder networks

Encoder Network	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
ResNet50[53]	0.930 ± 0.012	0.924 ± 0.007	0.945 ± 0.005	0.859 ± 0.01	0.966 ± 0.004	0.939 ± 0.006	3.723 ± 1.707
ResNeXt50[54]	0.929 ± 0.008	0.923 ± 0.017	0.944 ± 0.004	0.856 ± 0.012	0.965 ± 0.002	0.937 ± 0.003	3.959 ± 1.315
DenseNet121[55]	0.935 ± 0.011	0.921 ± 0.018	0.945 ± 0.005	0.860 ± 0.012	0.966 ± 0.004	0.938 ± 0.006	3.928 ± 1.157
EfficientNet-B0[56]	0.927 ± 0.016	0.924 ± 0.012	0.945 ± 0.003	0.856 ± 0.010	0.967 ± 0.002	0.940 ± 0.002	2.962 ± 0.725
Res2Net50[52]	0.929 ± 0.008	0.931 ± 0.021	0.946 ± 0.008	0.865 ± 0.013	0.967 ± 0.003	0.940 ± 0.005	2.78 ± 0.494
MobileNet[57]	0.926 ± 0.018	0.927 ± 0.008	0.945 ± 0.004	0.858 ± 0.012	0.967 ± 0.004	0.940 ± 0.006	3.545 ± 1.851
Ours	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

Ablation study for CNN encoder network

Because different pre-trained encoder networks will extract local features of varying quality, choosing the right CNN encoder network is essential for CTH-Net. Table 7 shows the ablation experiment results of different encoder networks on ISIC 2018. We selected six mainstream backbone networks for comparative experiments. Compared to other networks, Res2Net50 [52] is regarded as the most competitive method. It builds a feature pyramid structure inside each residual block and performs multi-scale convolution inside the feature layer to form different receptive fields, thereby obtaining different fine-grained features. Compared with the original Res2Net50, the dice score and Jaccard index have greatly improved using our method from 0.946 and 0.865 to 0.959 and 0.893, respectively. On FWIoU and 95% HD, it increased by 1.2% and 1.226%, respectively. Such performance improvement is due to the fast and efficient SoftPool, which in the downsampling activation map keeps more information and can obtain better pixel-by-pixel classification accuracy.

Ablation study for dual encoder

We conducted an ablation study to compare the dual encoder's performance to a single-branch encoder that only contains the Transformer encoder or the CNN encoder to further confirm the dual encoder's efficacy. Table 8 quantitatively shows the comparison of the performance results of encoders from different branches on ISIC 2018. In contrast to a single CNNs encoder, our dual-encoder method achieves 1.6%, 4.0%, and 0.9% improvements in dice score, Jaccard index, and accuracy, respectively. At the same time, the FWIoU and 95% HD were increased from 0.937 and 3.924 mm to 0.952 and 1.554 mm respectively. Moreover, our dual-encoder approach accomplishes 1.3%, 0.6%, 1.1%, 1.9%, and 0.296 mm improvement compared to the single Transformer encoder in dice score, Jaccard index, accuracy, FWIoU, and 95%HD. CNN is better at extracting spatially relevant information and maintaining spatial details than Transformer, and Transformer is better at capturing long-range dependencies than CNN. As a result, integrating CNN and Transformer branches as the encoder of the model can mitigate the drawbacks of the two models while enhancing their strengths, enhancing the model's ability to segment skin lesions.

We depict the attention map of the output of the final layer of the CNN encoder and Transformer encoder, as shown in Figure 8, to more easily comprehend which feature regions are highlighted by the CNN encoder branch and the Transformer encoder branch. The accuracy of CTH-Net to recognize skin lesions from the global receptive field can be greatly increased by using our Transformer encoder, which employs dual attention to capture long-range dependencies. Transformer's abilities to capture long-range relationships are determined by its calculation principle, which also limits its capacity to capture local aspects. The CNN encoder, which simultaneously models the local receptive field of the input image through progressive convolution and pooling processes, is better able to identify local details and features. Thus, by incorporating high-efficiency CNN and dual-attention Transformer branches into CTH-Net, it is possible to extract rich local features and crucial global contextual data for skin lesion segmentation.

Ablation study for key components

To assess each critical component's performance in the proposed network, we conduct a step-by-step ablation study using several comparative models:

Baseline: Choose TransFuse[38] as the baseline, and complete our network design based on this.

Model 1: Use CNN and Transformer dual encoder instead of the original encoder in TransFuse.

Model 2: Use the multi-domain feature fusion module instead of the feature fusion module in TransFuse.

Model 3: Add a multi-domain feature fusion module based on Model 1.

Model 4: Add sandglass block based on Model 3.

Model 5: Add boundary refinement module based on Model 4.

Model 6 (Ours): Add feature adaptive guided module based on Model 5.

Table 9 displays the comprehensive quantitative experimental results for the baseline and the six designs we proposed. Compared to the starting point, Model 1 improves the performance by 0.7%, 2.0%, 0.2%, 0.3%, and 1.66 mm in terms of dice score, Jaccard index, accuracy, FWIoU, and 95%HD by using dual encoders. Compared to the baseline, the dice score and Jaccard index of Model 2 have increased by 0.8%

Table 8. Performance comparison of the dual encoder

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
A single CNNs encoder	0.935 ± 0.017	0.913 ± 0.026	0.943 ± 0.006	0.853 ± 0.011	0.966 ± 0.005	0.937 ± 0.008	3.924 ± 0.822
A single transformer encoder	0.938 ± 0.003	0.945 ± 0.008	0.946 ± 0.001	0.887 ± 0.006	0.964 ± 0.001	0.933 ± 0.002	1.850 ± 0.369
Dual encoder with CNNs and transformer (Ours)	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

and 2.0%, respectively, which proves the usefulness of the multi-domain feature fusion module. Compared with Model 1 and Model 2, Model 3 improves dice score and Jaccard index by 1.1%, 1.0%, 2.6%, and 2.6%, respectively, indicating that using dual encoders and multi-domain feature fusion module at the same time can enhance the model's performance even more. Compared with Model 3, Model 4 improves Jaccard index and 95% HD by 0.7% and 0.589 mm, respectively, which proves that sandglass block can effectively minimize the chance of information loss and gradient confusion. Compared with Model 4, Model 5 improves the performance scores of dice score, Jaccard index, and FWIoU by 0.5%, 0.4%, and 0.9%, which shows that boundary refinement module has a significant effect on the fine-grained description of the boundary of the skin lesion area. Compared with the baseline, the dice score, Jaccard index, accuracy, FWIoU, and 95%HD of CTH-Net are significantly improved by 2.6%, 6.4%, 1.5%, 2.4%, and 3.889 mm, respectively. This demonstrates the great segmentation performance of the suggested network.

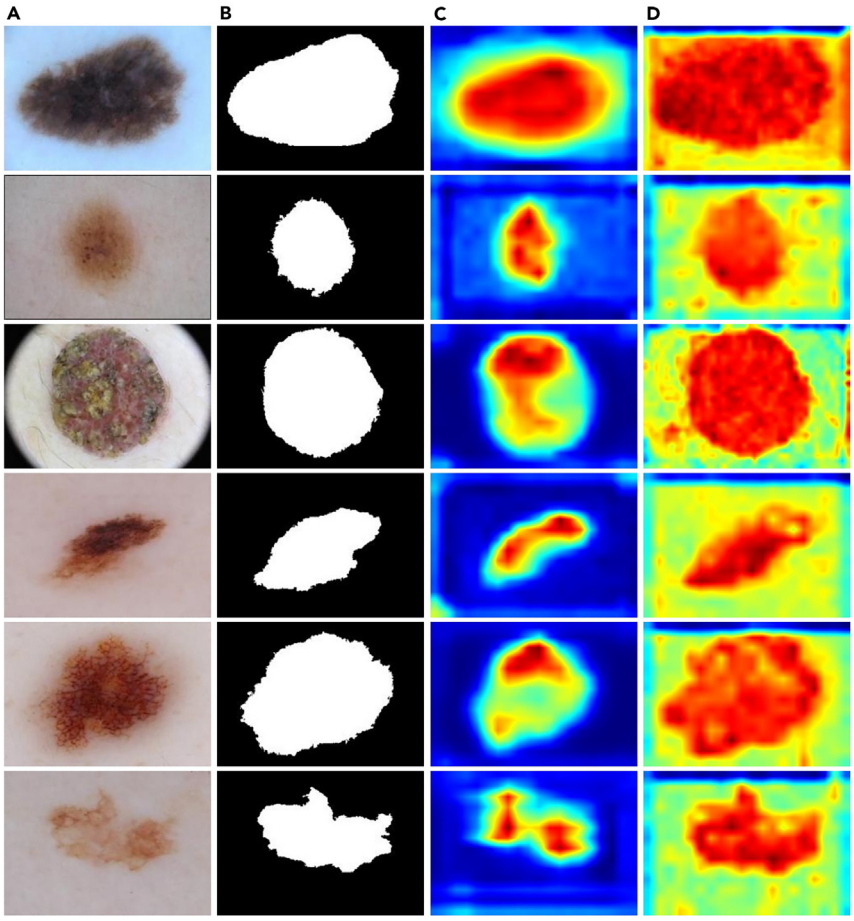


Figure 8. Visual comparison of different attention maps

(A) Input image.

(B) Ground truth.

(C) The attention map of the last layer of the CNN encoder.

(D) The attention map of the last layer of the Transformer encoder.

Table 9. Performance comparison between baseline and different proposed models

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
Baseline	0.905 ± 0.013	0.916 ± 0.026	0.933 ± 0.006	0.829 ± 0.013	0.960 ± 0.004	0.928 ± 0.007	5.443 ± 2.125
Model 1	0.909 ± 0.011	0.934 ± 0.013	0.940 ± 0.006	0.849 ± 0.016	0.962 ± 0.003	0.931 ± 0.005	3.783 ± 1.008
Model 2	0.918 ± 0.011	0.922 ± 0.008	0.941 ± 0.003	0.849 ± 0.011	0.964 ± 0.003	0.934 ± 0.005	3.537 ± 0.387
Model 3	0.926 ± 0.005	0.942 ± 0.008	0.951 ± 0.004	0.875 ± 0.007	0.967 ± 0.005	0.938 ± 0.009	3.019 ± 0.291
Model 4	0.932 ± 0.005	0.945 ± 0.007	0.952 ± 0.004	0.882 ± 0.007	0.968 ± 0.004	0.940 ± 0.006	2.430 ± 0.538
Model 5	0.939 ± 0.004	0.943 ± 0.009	0.957 ± 0.003	0.886 ± 0.007	0.973 ± 0.003	0.949 ± 0.005	2.004 ± 0.583
Model 6(Ours)	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

Figure 9 qualitatively shows the visual segmentation results of the baseline and the proposed models. The segmentation outcomes for small-area lesions are displayed in the image in the first row of Figure 9. After the addition of two encoders, the model's capacity to find and recognize small lesion sites improved as compared to the baseline. The segmentation outcomes of lesions with irregular shapes are displayed in the image in the fourth row of Figure 9. When Model 3 and Model 1 findings are compared, it is clear that the multi-domain feature fusion module considerably enhanced the network's performance for irregularly shaped lesions. This confirms that MFFM can realize the feature complementation and fusion between CNN and Transformer, enhance the important information in the two feature maps suppress the insignificant features, and further enhance the segmentation ability of the model. The segmentation outcomes of low-contrast lesions are displayed in the images in the fifth and sixth rows of Figure 9. Comparing Model 4, Model 5, and Ours, it can be seen that after the introduction of the boundary refinement module and feature adaptive guided module, the model has achieved significant improvement in the fine depiction of the segmentation result boundary. This confirms that our boundary refinement module and feature adaptive guided module embedded in the skip connection can narrow the difference in features between encoders and decoders. At the same time, it learns and improves the mismatched lesion boundaries, and obtains more accurate skin lesion boundary segmentation results.

To better observe the feature representations learned by CTH-Net at each stage of the encoder-decoder, we give the visualization results、the corresponding ground truth, and the predicted mask of the attention map of CTH-Net in various phases of the encoder and decoder in Figure 10. It can be seen that with the deepening of the encoder level, the visualization results of the attention map gradually present more accurate lesion localization and boundary delineation effects. The focus is gradually transformed from shallow features such

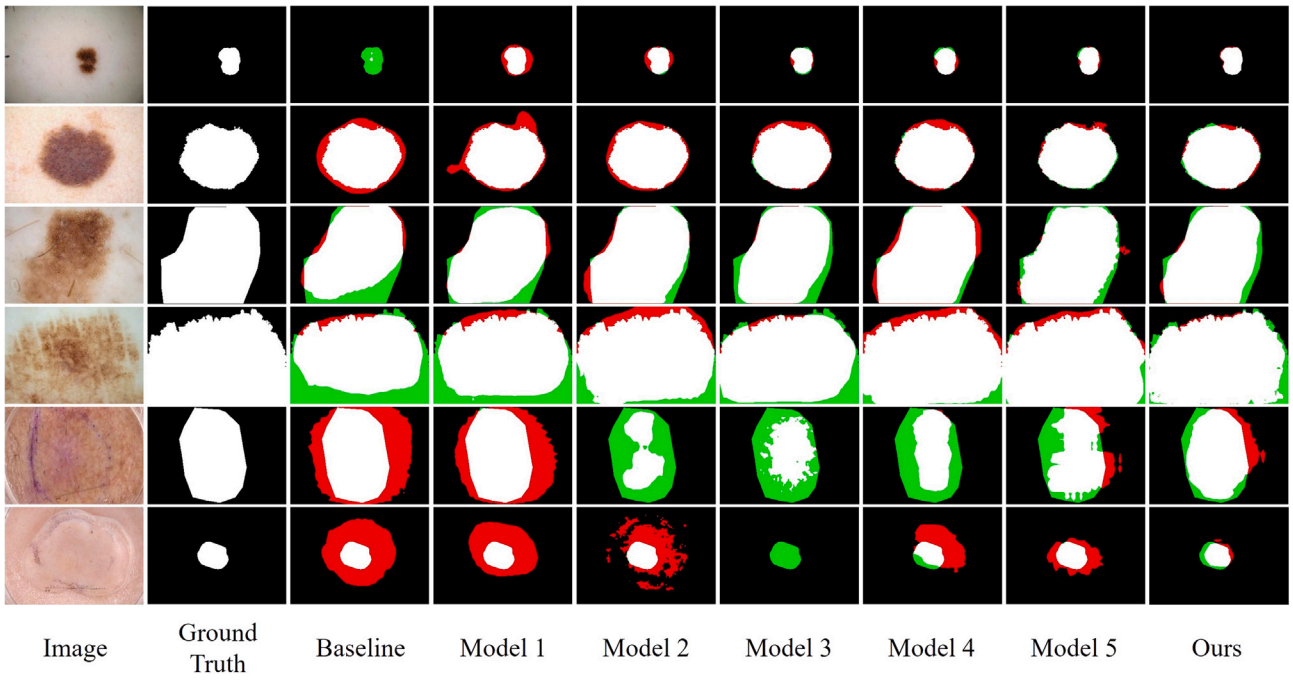


Figure 9. Visual comparison between baseline and different proposed models
White, green, and red, respectively, stand for proper segmentation, under-segmentation, and over-segmentation.

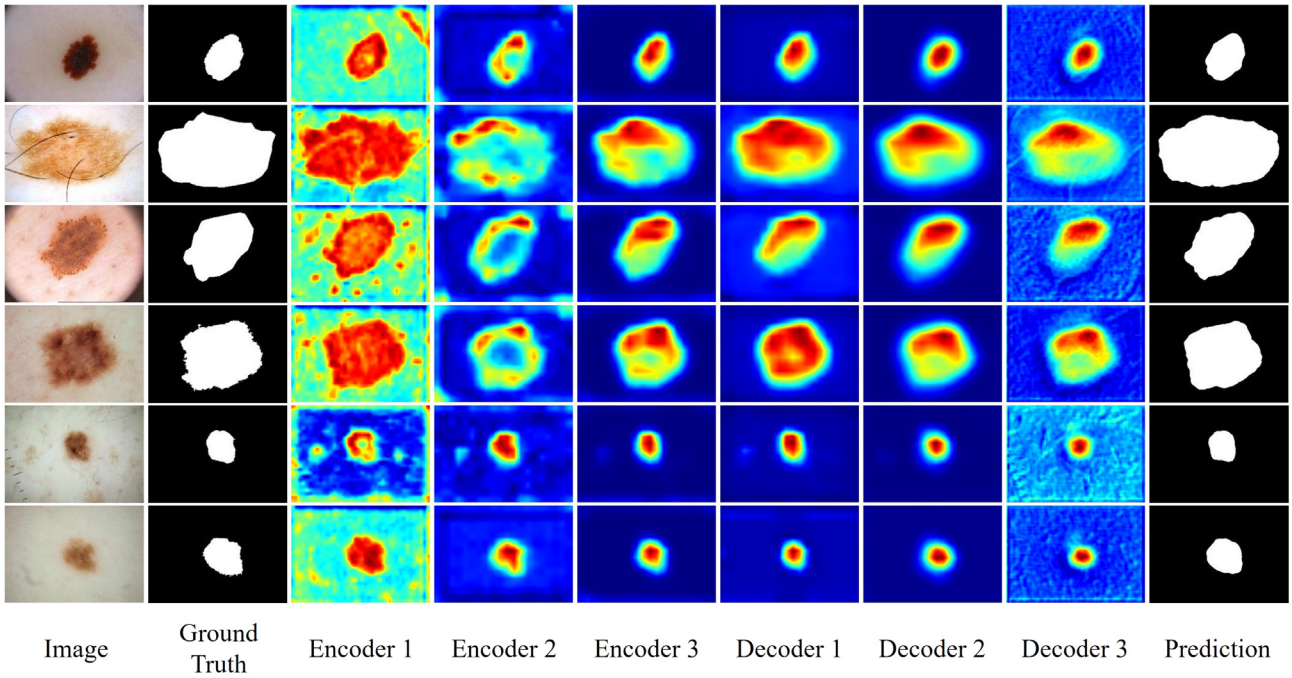


Figure 10. Visual comparison of different attention maps for each stage in CTH-Net

as edges and textures to high-level semantic features for overall position and context. And with the deepening of the decoder layer, the model can better utilize the contextual information in the encoder to guide the prediction at the pixel level. To get more precise pixel-by-pixel segmentation results features from various levels are integrated simultaneously to create a more comprehensive semantic context. Visualizing the attention maps of each stage in the model can not only help understand the attention distribution of the encoder and decoder at different stages but also make the segmentation findings more understandable.

Ablation study for MFFM

To check out the performance of each block in the multi-domain feature fusion module, we designed an ablation experiment by gradually increasing different blocks. Table 10 quantitatively shows the step-by-step ablation results of channel attention block (CAB), spatial attention block (SAB), cross-domain enhancement block (CDEB), and feature fusion block (FFB) in MFFM. Table 10 shows the segmentation results after using both the channel attention block and the spatial attention block. Compared with the result of only using the feature fusion block, the dice score, Jaccard index, and 95% HD are significantly improved by 1.1%, 2.9%, and 2.694 mm. It has been amply demonstrated that the simultaneous usage of CAB and SAB may successfully achieve the mixing of channels and self-attention while also promoting global information from the Transformer branch. Additionally, it can accentuate regional specifics while suppressing unimportant areas. Our method achieves the best performance in ablation studies, improving the Jaccard index by 3.6%, 1.5%, and 0.7%, respectively, compared to the other three variants. It demonstrates how each MFFM component exhibits its distinct benefits.

The visual segmentation outcomes of various approaches in the MFFM step-by-step ablation investigation are qualitatively displayed in Figure 11. The segmentation outcomes for minor lesions are displayed in the images in the top row of Figure 11. It can be seen that compared with the misjudgment of other skin regions in (a), there are more under-segmented or over-segmented areas in (b) and (c). In (d), by using CDEB, the important information in the two feature maps of the dual-branch encoder is fused and enhanced and the insignificant features are suppressed. The segmentation outcomes of lesions with irregular edges are displayed in the images in the second, third, and sixth rows of Figure 11. In (a), only the segmentation results of FFB are used, and the outline of complex irregular boundaries is far from meeting the requirements of accurate skin lesion segmentation. In contrast, more boundary information is mined and the lesion boundary is optimized more successfully in (c) to produce more precise segmentation visualization. All in all, both quantitative and qualitative experiments have fully proved that MFFM plays an important role in CTH-Net.

Ablation study for bottleneck

To verify the effectiveness of sandglass block, we conducted ablation studies on different types of bottleneck layers, including no bottleneck layer, residual block,[53] inverted residual block,[58] and sandglass block. Table 11 shows the segmentation performance comparison of models using different bottleneck layers on ISIC 2018. It can be seen that compared with the case of not using the bottleneck layer, the CTH-Net using residual block and inverted residual block performs better in various performance indicators. The model using sandglass bbock obtained the

Table 10. Performance comparison of different block combinations in MFFM

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
FFB	0.926 ± 0.011	0.926 ± 0.009	0.945 ± 0.005	0.857 ± 0.008	0.967 ± 0.006	0.939 ± 0.009	4.57 ± 1.579
+ CAB	0.936 ± 0.008	0.939 ± 0.005	0.948 ± 0.002	0.878 ± 0.007	0.968 ± 0.002	0.94 ± 0.004	2.046 ± 0.439
+ SAB	0.939 ± 0.004	0.943 ± 0.009	0.956 ± 0.005	0.886 ± 0.009	0.972 ± 0.004	0.948 ± 0.006	1.876 ± 0.449
+ CDEB	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

best performance score in the experiment. Compared with the most competitive inverted residual block, the dice score, Jaccard index, accuracy, FWIoU and 95% HD are improved by 0.3%, 0.6%, 0.3%, 0.4%, and 0.267 mm, respectively. It shows that sandglass block can assist the network's expressiveness and segmentation performance.

Ablation study for upsampling

In each decoding stage of the CTH-Net decoder, the features of the decoder are first concat with the feature map of the corresponding layer skip connection of the encoder. Convolute the spliced map next to change the number of channels, and then use the upsampling technique

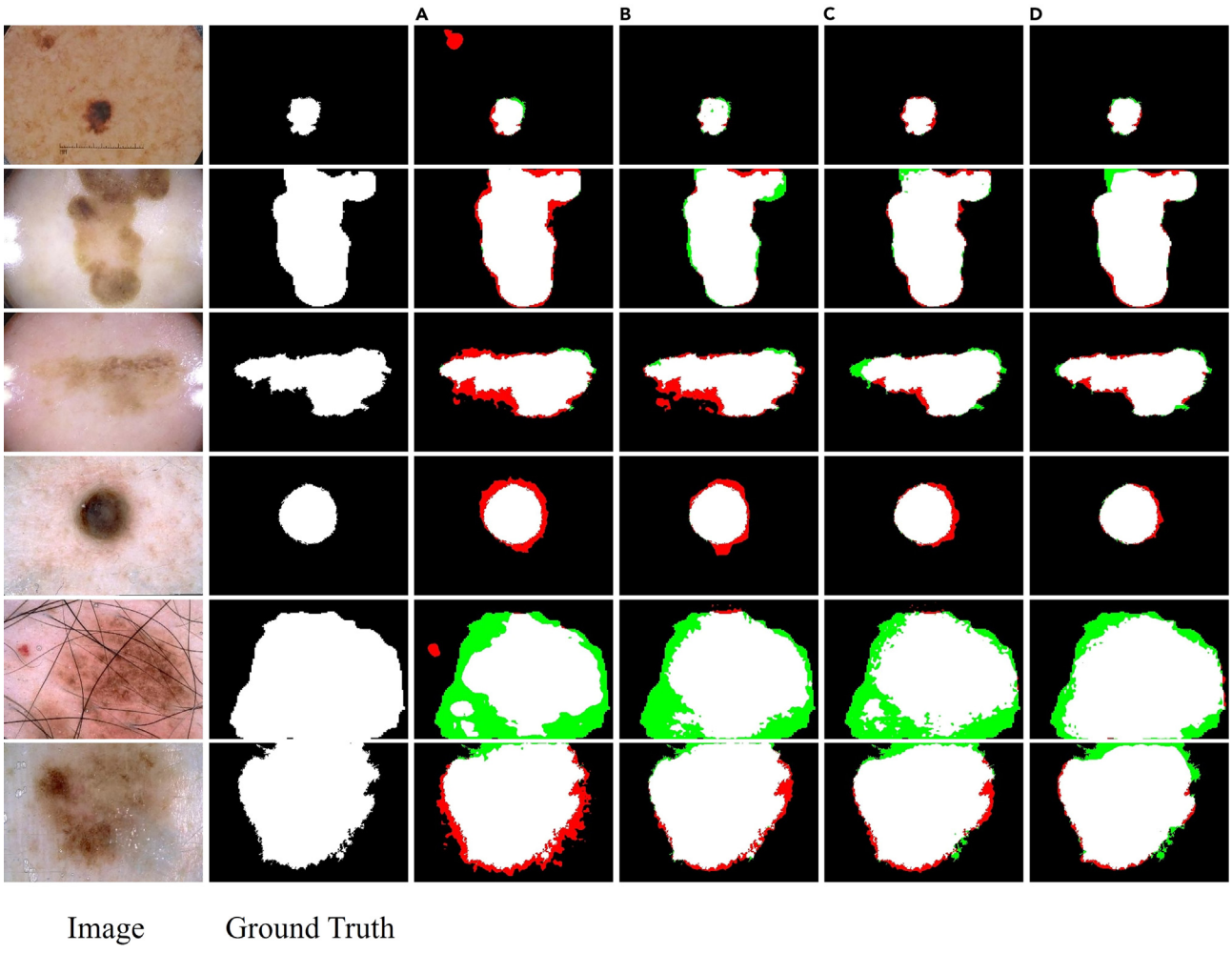


Figure 11. Visual comparison of different block combinations in MFF

(A) FFB.

(B) + CAB.

(C) + SAB.

(D) + CDEB. White, green, and red, respectively, stand for proper segmentation, under-segmentation, and over-segmentation.

Table 11. Performance comparison between different bottleneck layers

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
w/o Bottleneck layer	0.936 ± 0.007	0.936 ± 0.012	0.946 ± 0.004	0.877 ± 0.008	0.966 ± 0.004	0.937 ± 0.007	2.065 ± 0.623
Residual Block	0.937 ± 0.006	0.942 ± 0.008	0.948 ± 0.002	0.883 ± 0.009	0.968 ± 0.002	0.939 ± 0.004	2.597 ± 0.999
Inverted Residual Block	0.940 ± 0.002	0.941 ± 0.007	0.956 ± 0.003	0.887 ± 0.006	0.972 ± 0.003	0.948 ± 0.006	1.821 ± 0.439
Sandglass Block	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

to double the size of the feature map and cut in half the number of feature map channels before sending it to the following decoding stage. To explore the most effective upsampling method, Table 12 shows the ablation study on the performance of skin lesion segmentation using different upsampling methods in the decoder. Transposed convolution has demonstrated the best performance, as can be observed, which is significantly improved by 3.9% and 4.6% in terms of the Jaccard index compared with bilinear iInterpolation and UnPooling.

Comparison between different loss functions

We select the weighted loss function of binary cross entropy (BCE) and SoftDice to optimize the network throughout the end-to-end training of CTH-Net. First, we designed comparative experiments to find the optimal correlation importance weights λ. Figure 12 intuitively shows the changing trend of the scores of different evaluation indicators in the process of λ increasing from 0.1 to 0.9. It can be seen that when the value λ is set to 0.8, CTH-Net obtains better segmentation performance.

We employed five loss functions to optimize the network to examine the effects of various loss functions on the performance of CTH-Net, including BCE loss (loss 1), Dice loss (loss 2), SoftDice loss (loss 3), BCE+Dice loss (loss 4), and BCE+SoftDice loss (loss 5). The performance comparison results of five different loss functions on skin lesion segmentation are shown in Table 13 and Figure 13. It is clear that loss 5 has outperformed the other loss functions in terms of performance. When using loss 1 and loss 2 to optimize the performance of the network, the performance of the network is comparable, but loss 4 after the combination of the two has achieved a performance improvement of 0.7% and 0.5% respectively on the Jaccard index. This is because different loss functions have different concerns for different aspects of model training, and each loss function can capture different feature information. By using a weighted combination of multiple loss functions, the needs of multiple aspects can be considered comprehensively, providing more comprehensive and accurate training signals, and helping the model learn more details. At the same time, it strengthens the model's robustness and lessens reliance on a single loss function. Additionally, it can be seen that the performance for loss 3 is superior to the performance for loss 2, and the performance for loss 5 is superior to the performance for loss 4. This is because SoftDice loss introduces a smoothing factor based on Dice loss, which converts the binary Dice coefficient into a continuous probability value. This smoothness can alleviate the extreme binarization of the prediction results, making the model more stable during the gradient descent process. Overall, in ISIC 2018, by using the weighted loss function of BCE+SoftDice, the segmentation performance of the model can be improved.

Efficiency study

Learning efficiency

To compare the learning efficiency of different models in the training and verification process, we monitored the changing trend of the Jaccard Index and loss values with epochs. The outcomes are displayed in Figure 14. It is clear that CTH-Net is simpler to train and converge than alternative approaches. Compared with HiFormer, which is the most competitive performance in ISIC 2018, our method has faster learning speed and lower training loss during the training process and only needs 60 epochs on the training set to complete the convergence. When using 2,586 images as training samples, CTH-Net only needs 68 s to train an epoch on a single NVIDIA GeForce RTX 4090 GPU. This indicates that after training for roughly 70 min, a skin lesion segmentation model with good performance can be obtained. The aforementioned experimental findings conclusively show that the proposed CTH-Net is simple to train.

Pre-trained and data augmentation

By choosing appropriate learning techniques, such as pre-training and data augmentation operations, the performance of the model can be improved to a certain extent and reach peak performance. We performed a comparative experiment on ISIC 2018 to assess the impact of

Table 12. Performance comparison between different upsampling

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
Bilinear Interpolation	0.933 ± 0.013	0.914 ± 0.014	0.944 ± 0.004	0.854 ± 0.01	0.966 ± 0.004	0.938 ± 0.007	3.409 ± 0.742
UnPooling	0.919 ± 0.014	0.921 ± 0.012	0.941 ± 0.005	0.847 ± 0.012	0.966 ± 0.004	0.937 ± 0.007	3.856 ± 0.876
Transposed Convolution	**0.944±0.011**	**0.946±0.006**	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

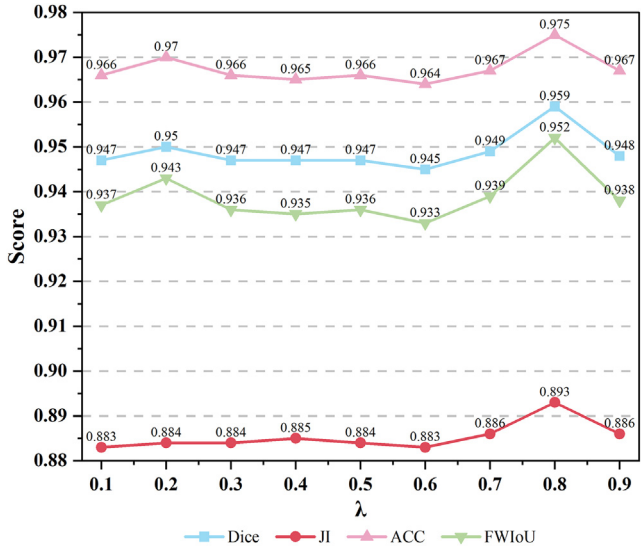


Figure 12. Performance comparison between different importance weights

pre-training and data augmentation operations on CTH-Net. The pre-trained model is obtained by training the encoder of CTH-Net on ImageNet 2012[59] and then fine-tuning the specific skin lesion segmentation task. Data augmentation strategies used horizontal or vertical flipping, random rotation (−20 to 20°), cropping, scaling, and adjusting brightness and contrast (−3%–3%). The effect of using the pre-trained model and the data augmentation operation on the model performance is intuitively illustrated in Figure 15. According to the results, the pre-training model and data augmentation both assist CTH-Net in performing better. Compared with the case of not using the pre-training model and data augmentation, the dice score, Jaccard index, accuracy, and FWIoU of CTH-Net increased by 2.6%, 6.5%, 1.6%, and 2.5%, respectively. This shows that CTH-Net has strong learning ability, and by combining some appropriate learning techniques, it can achieve excellent segmentation performance.

Inspired by the powerful representation capabilities of CNN and Transformer, this paper proposes a new hybrid encoder-decoder model CTH-Net based on CNN and Transformer. It can effectively utilize Transformer's global long-range relationship and CNN's local feature representation to achieve accurate and reliable skin lesion segmentation. Specifically, we build a CNN encoder branch based on Res2Net50 and SoftPool that can extract fine-grained features, while using a Transformer branch with channel and spatial dual attention in parallel to capture long-range dependencies. We create a multi-domain feature fusion mdule to more effectively cross-fuse multi-domain features from two encoder branches. Next, we embed a boundary refinement module and a feature adaptive guided module in skip connections. By using contextual information to fine-grained outline the lesion boundary, the learned lesion boundary is improved, and the feature distribution between the encoder and decoder is better adaptively matched. Extensive tests on four datasets of skin lesions that are available to the public show that the proposed CTH-Net provides cutting-edge segmentation performance in both quantitative and qualitative analysis. We will extend CTH-Net in the future to support medical image segmentation tasks in other fields based on the great performance of the current technology.

Limitations of the study

Although our method achieved satisfactory segmentation results, however, it still has some limitations. Similar to most existing state-of-the-art methods, our method still fails to accurately outline the boundaries of skin lesion areas when the contrast between the

Table 13. Performance comparison between different loss functions

Methods	Precision	Recall	Dice Score	Jaccard Index	Accuracy	FWIoU	95%HD
BCE (loss 1)	0.933 ± 0.006	0.938 ± 0.006	0.943 ± 0.005	0.876 ± 0.011	0.963 ± 0.004	0.932 ± 0.008	2.256 ± 0.520
Dice (loss 2)	0.924 ± 0.010	**0.949±0.011**	0.945 ± 0.003	0.878 ± 0.012	0.965 ± 0.002	0.935 ± 0.003	2.129 ± 0.786
SoftDice (loss 3)	0.942 ± 0.009	0.936 ± 0.007	0.946 ± 0.003	0.883 ± 0.011	0.965 ± 0.002	0.935 ± 0.003	1.948 ± 0.471
BCE + Dice (loss 4)	0.938 ± 0.009	0.937 ± 0.011	0.947 ± 0.004	0.879 ± 0.014	0.966 ± 0.004	0.937 ± 0.007	1.722 ± 0.504
BCE+SoftDice (loss 5)	**0.944±0.011**	0.946 ± 0.006	**0.959±0.002**	**0.893±0.007**	**0.975±0.002**	**0.952±0.003**	**1.554±0.262**

The best outcomes are highlighted in bold. Data are represented as mean ± std.

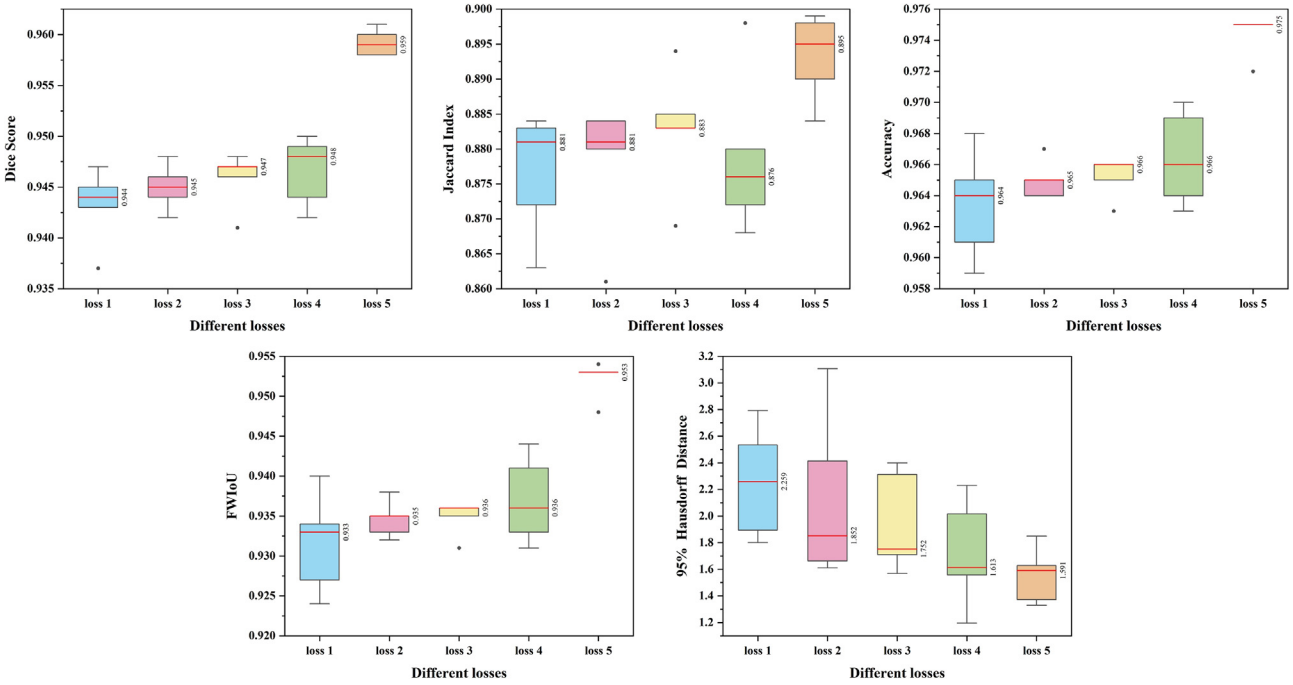


Figure 13. Performance comparison between different loss functions on ISIC 2018

Loss 1 is BCE, loss 2 is Dice, loss 3 is SoftDice, loss 4 is BCE + Dice, loss 5 is BCE + SoftDice. Boxes in different color indicate the score range of different models, the line inside each box represents the median value, box limits include interquartile ranges Q1 and Q3 (from 25% to 75% of samples), upper and lower whiskers are computed as 1.5 times the distance of upper and lower limits of the box, and all values outside the whiskers are considered outliers.

skin lesion and the background tissue in the dermatoscopic image is extremely low, or when the color inside the skin lesion changes too much. However, the segmentation effect of our method is closest to the real situation and outperforms other competitors. Second, CTH-Net is specifically designed for the task of skin lesion segmentation and has not yet explored its potential for other medical image segmentation tasks. To address the above limitations, in our future work, we will explore more model structure design and boundary

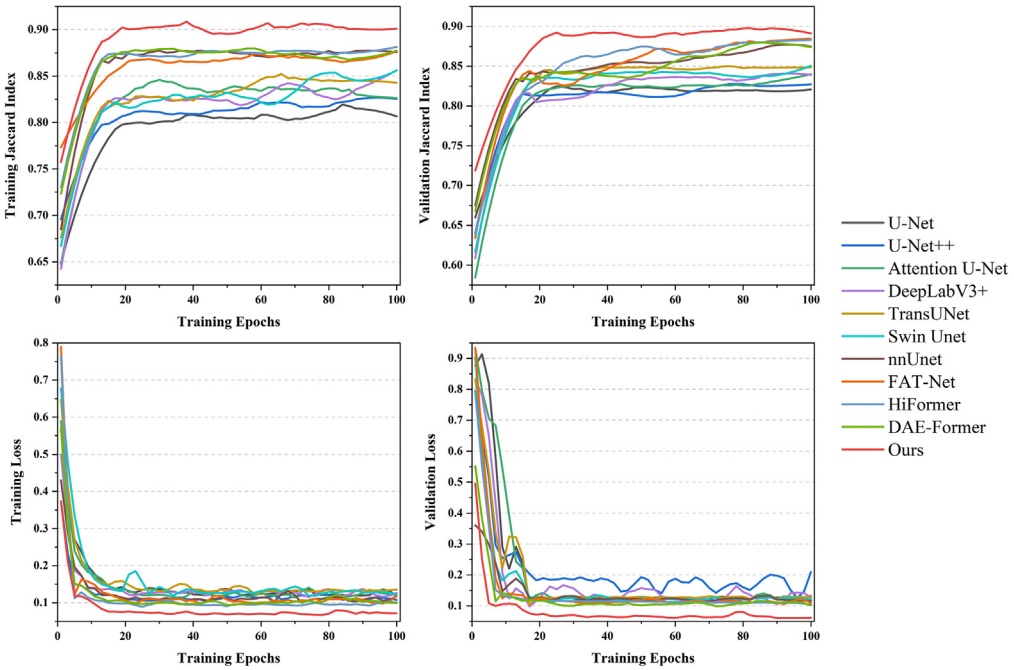


Figure 14. Comparison of the learning efficiency between different models on the training set and the validation set

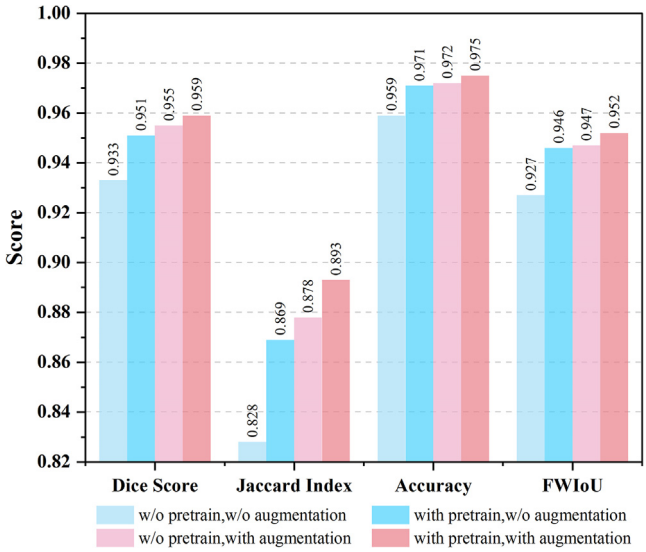


Figure 15. Performance comparison of whether to use the pre-trained model and data augmentation

refinement strategies to further improve the performance of skin lesion segmentation in dermoscopy images. Meanwhile, we will continue to explore the potential of the proposed CTH-Net and apply it to medical image segmentation tasks in other fields.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Network architecture
  - CNN and Transformer dual encoder
  - Multi-domain feature fusion module
  - Sandglass Block
  - Boundary Refinement Module
  - Feature Adaptive Guided Module
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109442.

ACKNOWLEDGMENTS

This work was supported by the grant from Hunan Provincial Natural Science Foundation of China (2021JJ41026) and the Fundamental Research Funds for the Central Universities of Central South University.

AUTHOR CONTRIBUTIONS

D.Y.H., W.Y.J., and L.Z.F. conceived and supervised the study. Y.Z.L., H.M.H., and G.Y. contributed to data collection and assembly. D.Y.H., Y.Z.L., and X.J.T. performed data analysis and interpretation. D.Y.H., H.M.H., and G.Y. performed software, visualization, and validation. All authors contributed to writing the manuscript. All authors reviewed and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 22, 2023
Revised: January 25, 2024
Accepted: March 4, 2024
Published: March 6, 2024

REFERENCES

1. Siegel, R.L., Miller, K.D., and Jemal, A. (2019). Cancer statistics, 2019. CA. Cancer J. Clin. *69*, 7–34. https://doi.org/10.3322/caac.21551.
2. Wang, X., Jiang, X., Ding, H., and Liu, J. (2019). Bi-Directional Dermoscopic Feature Learning and Multi-Scale Consistent Decision Fusion for Skin Lesion Segmentation. IEEE Trans. Image Process. *29*, 3039–3051. https://doi.org/10.1109/TIP.2019.2955297.
3. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., and Garnavi, R. (2017). Skin Disease Recognition Using Deep Saliency Features and Multimodal Learning of Dermoscopy and Clinical Images. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2017 Lecture Notes in Computer Science, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D.L. Collins, and S. Duchesne, eds. (Springer International Publishing), pp. 250–258. https://doi.org/10.1007/978-3-319-66179-7_29.
4. Sarker, M.M.K., Rashwan, H.A., Akram, F., Banu, S.F., Saleh, A., Singh, V.K., Chowdhury, F.U.H., Abdulwahab, S., Romani, S., Radeva, P., et al. (2018). SLSDeep: Skin Lesion Segmentation Based on Dilated Residual and Pyramid Pooling Networks. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 Lecture Notes in Computer Science, A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds. (Springer International Publishing), pp. 21–29. https://doi.org/10.1007/978-3-030-00934-2_3.
5. González-Díaz, I. (2019). DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis. IEEE J. Biomed. Health Inform. *23*, 547–559. https://doi.org/10.1109/JBHI.2018.2806962.
6. Mishra, N.K., and Celebi, M.E. (2016). An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.1601.07843.
7. Mahbod, A., Tschandl, P., Langs, G., Ecker, R., and Ellinger, I. (2020). The effects of skin lesion segmentation on the performance of dermatoscopic image classification. Comput. Methods Progr. Biomed. *197*, 105725. https://doi.org/10.1016/j.cmpb.2020.105725.
8. Ximenes Vasconcelos, F.F., Medeiros, A.G., Peixoto, S.A., and Rebouças Filho, P.P. (2019). Automatic skin lesions segmentation based on a new morphological approach via geodesic active contour. Cognit. Syst. Res. *55*, 44–59. https://doi.org/10.1016/j.cogsys.2018.12.008.
9. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444. https://doi.org/10.1038/nature14539.
10. Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.
11. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 Lecture Notes in Computer Science, N. Navab, J. Hornegger, W.M. Wells, and A.F. Frangi, eds. (Springer International Publishing), pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
12. Zhang, Z., Liu, Q., and Wang, Y. (2018). Road Extraction by Deep Residual U-Net. Geosci. Rem. Sens. Lett. IEEE *15*, 749–753. https://doi.org/10.1109/LGRS.2018.2802944.
13. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., and Liang, J. (2020). UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans. Med. Imag. *39*, 1856–1867. https://doi.org/10.1109/TMI.2019.2959609.
14. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al. (2018). Attention U-Net: Learning Where to Look for the Pancreas. Preprint at arXiv. https://doi.org/10.48550/arXiv.1804.03999.
15. Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. https://doi.org/10.1109/3DV.2016.79.
16. Shahin, A.H., Amer, K., and Elattar, M.A. (2019). Deep Convolutional Encoder-Decoders with Aggregated Multi-Resolution Skip Connections for Skin Lesion Segmentation. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 451–454. https://doi.org/10.1109/ISBI.2019.8759172.
17. Hu, K., Lu, J., Lee, D., Xiong, D., and Chen, Z. (2022). AS-Net: Attention Synergy Network for skin lesion segmentation. Expert Syst. Appl. *201*, 117112.
18. Yuan, F., Zhang, Z., and Fang, Z. (2023). An effective CNN and Transformer complementary network for medical image segmentation. Pattern Recogn. *136*, 109228. https://doi.org/10.1016/j.patcog.2022.109228.
19. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at arXiv. https://doi.org/10.48550/arXiv.2010.11929.
20. Ding, Y., Yi, Z., Li, M., long, J., Lei, S., Guo, Y., Fan, P., Zuo, C., and Wang, Y. (2023). HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT. Digit. Health *9*, 20552076231207197. https://doi.org/10.1177/20552076231207197.
21. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H.S., et al. (2021). Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. Preprint at arXiv. https://doi.org/10.1109/CVPR46437.2021.00681.
22. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J., and Merhof, D. (2023). HiFormer: Hierarchical Multi-scale Representations Using Transformers for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.1109/WACV56688.2023.00614.
23. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., and Zhou, Y. (2021). TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.04306.
24. Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.04306.
25. Li, S., Sui, X., Luo, X., Xu, X., Liu, Y., and Goh, R. (2021). Medical Image Segmentation Using Squeeze-and-Expansion Transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.09511.
26. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., and Zha, S. (2021). TransBTS: Multimodal Brain Tumor Segmentation Using Transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2103.04430.
27. Liu, Q., Wang, J., Zuo, M., Cao, W., Zheng, J., Zhao, H., and Xie, J. (2022). NCRNet: Neighborhood context refinement network for skin lesion segmentation. Comput. Biol. Med. *146*, 105545.
28. Stergiou, A., Poppe, R., and Kalliatakis, G. (2021). Refining activation downsampling with SoftPool. Preprint at arXiv. https://doi.org/10.48550/arXiv.2101.00440.
29. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., and 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 Lecture Notes in Computer Science, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J.M.R.S. Tavares, A. Bradley, J.P. Papa, and V. Belagiannis, et al., eds. (Springer International Publishing), pp. 3–11. https://doi.org/10.1007/978-3-030-00889-5_1.
30. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Computer Vision – ECCV 2018 Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds. (Springer International Publishing), pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
31. Ibtehaz, N., and Rahman, M.S. (2020). MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Network. *121*, 74–87. https://doi.org/10.1016/j.neunet.2019.08.025.
32. Feng, S., Zhao, H., Shi, F., Cheng, X., Wang, M., Ma, Y., Xiang, D., Zhu, W., and Chen, X. (2020). CPFNet: Context Pyramid Fusion

Network for Medical Image Segmentation. IEEE Trans. Med. Imag. *39*, 3008–3018. https://doi.org/10.1109/TMI.2020.2983721.
33. Karaali, A., Dahyot, R., and Sexton, D.J. (2022). DR-VNet: Retinal Vessel Segmentation via Dense Residual UNet. In Pattern Recognition and Artificial Intelligence: Third International Conference, ICPRAI 2022, Paris, France, June 1–3, 2022, Proceedings, Part I (Springer-Verlag), pp. 198–210. https://doi.org/10.1007/978-3-031-09037-0_17.
34. Wang, S., Chen, Z., You, S., Wang, B., Shen, Y., and Lei, B. (2022). Brain stroke lesion segmentation using consistent perception generative adversarial network. Neural Comput. Appl. *34*, 8657–8669. https://doi.org/10.1007/s00521-021-06816-8.
35. Wu, X., Bi, L., Fulham, M., Feng, D.D., Zhou, L., and Kim, J. (2021). Unsupervised brain tumor segmentation using a symmetric-driven adversarial network. Neurocomputing *455*, 242–254.
36. Gong, C., Jing, C., Chen, X., Pun, C.M., Huang, G., Saha, A., Nieuwoudt, M., Li, H.-X., Hu, Y., and Wang, S. (2023). Generative AI for brain image computing and brain network computing: a review. Front. Neurosci. *17*, 1203104.
37. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2105.05537.
38. Zhang, Y., Liu, H., and Hu, Q. (2021). TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2102.08005.
39. Azad, R., Arimond, R., Aghdam, E.K., Kazerouni, A., and Merhof, D. (2023). DAE-Former: Dual Attention-guided Efficient Transformer for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2212.13504.
40. Lei, T., Sun, R., Wang, X., Wang, Y., He, X., and Nandi, A. (2023). CiT-Net: Convolutional Neural Networks Hand in Hand with Vision Transformers for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2306.03373.
41. Tang, P., Liang, Q., Yan, X., Xiang, S., Sun, W., Zhang, D., and Coppola, G. (2019). Efficient skin lesion segmentation using separable-Unet with stochastic weight averaging. Comput. Methods Progr. Biomed. *178*, 289–301. https://doi.org/10.1016/j.cmpb.2019.07.005.
42. Dai, D., Dong, C., Xu, S., Yan, Q., Li, Z., Zhang, C., and Luo, N. (2022). Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation. Med. Image Anal. *75*, 102293. https://doi.org/10.1016/j.media.2021.102293.
43. Ruan, J., Xie, M., Gao, J., Liu, T., and Fu, Y. (2023). Ege-unet: an efficient group enhanced unet for skin lesion segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer), pp. 481–490.
44. Wu, H., Chen, S., Chen, G., Wang, W., Lei, B., and Wen, Z. (2022). FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Med. Image Anal. *76*, 102327.
45. Wang, J., Wei, L., Wang, L., Zhou, Q., Zhu, L., and Qin, J. (2021). Boundary-Aware Transformers for Skin Lesion Segmentation. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021 Lecture Notes in Computer Science, M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, eds. (Springer International Publishing), pp. 206–216. https://doi.org/10.1007/978-3-030-87193-2_20.
46. Cao, W., Yuan, G., Liu, Q., Peng, C., Xie, J., Yang, X., Ni, X., and Zheng, J. (2023). ICL-Net: Global and Local Inter-Pixel Correlations Learning Network for Skin Lesion Segmentation. IEEE J. Biomed. Health Inform. *27*, 145–156. https://doi.org/10.1109/JBHI.2022.3162342.
47. Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. (2016). Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). Preprint at arXiv. https://doi.org/10.48550/arXiv.1605.01397.
48. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al. (2018). 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). Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.05006.
49. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al. (2019). Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). Preprint at arXiv. https://doi.org/10.48550/arXiv.1902.03368.
50. Mendonca, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., and Rozeira, J. (2013). PH^2 - a dermoscopic image database for research and benchmarking. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. *2013*, 5437–5440. https://doi.org/10.1109/EMBC.2013.6610779.
51. Isensee, F., Jäger, P.F., Kohl, S.A.A., Petersen, J., and Maier-Hein, K.H. (2021). Automated Design of Deep Learning Methods for Biomedical Image Segmentation. Nat. Methods *18*, 203–211. https://doi.org/10.1038/s41592-020-01008-z.
52. Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2021). Res2Net: A New Multi-Scale Backbone Architecture. IEEE Trans. Pattern Anal. Mach. Intell. *43*, 652–662. https://doi.org/10.1109/TPAMI.2019.2938758.
53. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. Preprint at arXiv. https://doi.org/10.48550/arXiv.1512.03385.
54. Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.
55. Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K.Q. (2018). Densely Connected Convolutional Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.10108.
56. Tan, M., and Le, Q.V. (2020). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.11946.
57. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for MobileNetV3. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.02244.
58. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 4510–4520. https://doi.org/10.1109/CVPR.2018.00474.
59. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (IEEE), pp. 248–255.
60. Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.
61. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. *40*, 834–848. https://doi.org/10.1109/TPAMI.2017.2699184.
62. Huang, X., Deng, Z., Li, D., and Yuan, X. (2021). MISSFormer: An Effective Medical Image Segmentation Transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2109.0716.
63. Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., and Hu, S.-M. (2022). Attention Mechanisms in Computer Vision. Comput. Vis. Media (Beijing) *8*, 331–368. https://doi.org/10.1007/s41095-022-0271-y.
64. Zhuoran, S., Mingyuan, Z., Haiyu, Z., Shuai, Y., and Hongsheng, L. (2021). Efficient Attention: Attention with Linear Complexities. In 2021 IEEE Winter Conference on Applications of Computer Vision (WACV) (IEEE), pp. 3530–3538. https://doi.org/10.1109/WACV48630.2021.00357.
65. El-Nouby, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al. (2021). XCiT: Cross-Covariance Image Transformers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2106.09681.
66. Hendrycks, D., and Gimpel, K. (2016). Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units.
67. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195.
68. Qin, Z., Zhang, P., Wu, F., and Li, X. (2021). FcaNet: Frequency Channel Attention Networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2012.11879.
69. Fu, J., Liu, J., Jiang, J., Li, Y., Bao, Y., and Lu, H. (2021). Scene Segmentation With Dual Relation-Aware Attention Network. IEEE Transact. Neural Networks Learn. Syst. *32*, 2547–2560. https://doi.org/10.1109/TNNLS.2020.3006524.
70. Zhou, J., Wang, P., Wang, F., Liu, Q., Li, H., and Jin, R. (2021). ELSA: Enhanced Local Self-Attention for Vision Transformer. Preprint at arXiv. https://doi.org/10.48550/arXiv.2112.12786.
71. Daquan, Z., Hou, Q., Chen, Y., Feng, J., and Yan, S. (2020). Rethinking Bottleneck Structure for Efficient Mobile Network

Design. Preprint at arXiv. https://doi.org/10.48550/arXiv.2007.02269.

72. Sankararaman, K.A., De, S., Xu, Z., Huang, W.R., and Goldstein, T. (2020). The Impact of Neural Network Overparameterization on Gradient Confusion and Stochastic Gradient Descent. In Proceedings of the 37th International Conference on Machine Learning (PMLR), pp. 8469–8479.
73. He, K., Lian, C., Zhang, B., Zhang, X., Cao, X., Nie, D., Gao, Y., Zhang, J., and Shen, D. (2021). HF-UNet: Learning Hierarchically Inter-Task Relevance in Multi-Task U-Net for Accurate Prostate Segmentation in CT Images. IEEE Trans. Med. Imag. *40*, 2118–2128. https://doi.org/10.1109/TMI.2021.3072956.
74. Basak, H., Kundu, R., and Sarkar, R. (2022). MFSNet: A multi focus segmentation network for skin lesion segmentation. Pattern Recogn. *128*, 108673. https://doi.org/10.1016/j.patcog.2022.108673.
75. Dayananda, C., Yamanakkanavar, N., Nguyen, T., and Lee, B. (2023). AMCC-Net: An asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images. Eng. Appl. Artif. Intell. *122*, 106154.
76. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2004.08790.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ISIC 2016	Gutman et al.[47]	https://challenge.isic-archive.com/data/#2016
ISIC 2017	Codella et al.[48]	https://challenge.isic-archive.com/data/#2017
ISIC 2018	Codella et al.[49]	https://challenge.isic-archive.com/data/#2018
PH^2	Mendonca et al.[50]	https://www.fc.up.pt/addi/ph2%20database.html
Software and algorithms		
Python	Python Software Foundation	https://www.python.org/
PyTorch	PyTorch Foundation	https://pytorch.org/
Pycharm	JetBrains	https://www.jetbrains.com/pycharm/
CTH-Net	This paper	https://doi.org/10.5281/zenodo.10732004

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yongjie Wang (yongjiewang@csu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Github (https://github.com/0LeahD/CTH-Net) and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The PyTorch library is used to implement our suggested method end-to-end, and an NVIDIA GeForce RTX 4090 GPU is used for training. Training epochs are set at 100 and the batch size is 32. The stochastic optimization method of Adam[60] is adopted, the learning rate is initialized to 1e-4, the weight decay is 1e-7, and the "poly" learning rate strategy[61] is used for decay.

Referring to the setting in Dai et al.,[42] considering that the aspect ratio of most dermoscopic images is approximately 3:4, all datasets were resampled to 224×320 pixels and normalized. To broaden the variety of image samples, we also adopted a variety of data augmentation strategies, including horizontal or vertical flip, random rotation (-20 to 20 degrees), cropping, scaling, adjusting brightness, and contrast (-3% to 3%).

We carry out 5-fold cross-validation on ISIC 2018 and PH^2 and give the average performance of all assessment criteria to lessen the impact of randomness and create a fair comparison with other approaches.

Network architecture

Inspired by the powerful representation capabilities of CNN and Transformer, we propose a hybrid network (CTH-Net) based on CNN and Transformer to precisely and dependably segment dermoscopic images of skin lesions. The overall architecture is shown in Figure S1. Our approach primarily comprises five parts, including a dual encoder for enhanced feature encoding, a Multi-domain Feature Fusion Module (MFFM) for efficiently fusing encoded features of CNN and Transformer, a Boundary Refinement Module (BRM), Feature Adaptive Guided Module (FAGM), and a decoder that can perform feature decoding layer by layer. Meanwhile, we introduce two efficient methods: SoftPool and Sandglass Block.

Specifically, rich local characteristics as well as significant global contextual information for skin lesion segmentation can be extracted by merging high-efficiency CNN and dual-attention Transformer branches into CTH-Net. The Softpool method introduced in the CNN branch can preserve more information in the downsampled activation map, resulting in better classification accuracy, while being computationally

and memory efficient. Secondly, MFFM can effectively fuse the multi-domain features of CNN and Transformer to obtain more comprehensive and compact fusion features. We use a Sandglass Block in the bottleneck layer to increase model performance and decrease the number of parameters and calculations. This block is effective in reducing gradient confusion and information loss. Furthermore, we design a Boundary Refinement Module (BRM) to precisely guide and delineate the fuzzy contours of lesion boundaries by utilizing the fine-grained neighborhood contextual information and boundary information of the dual encoder fusion features. Combined with the use of a Feature Adaptive Guided Module (FAGM), the mismatched lesion boundaries can be learned and improved while reducing the feature gap between the encoder and decoder. Finally, the processed multi-scale features enter the decoder for layer-by-layer progressive upsampling feature decoding to obtain the dense prediction segmentation results.

CNN and Transformer dual encoder

Transformer branch

The design of the Transformer branch follows the classic encoder-decoder structure. For the input image $X \in \mathbb{R}^{H\times W\times C}$, firstly, overlapping patch tokens of size 4×4 are extracted from the input image using the overlapping patch embedding module.[62] Then, the tokenized input $z \in \mathbb{R}^{n\times d}$ passes through the Transformer encoder to generate multi-scale features, where n is the number of patches and d is the embedding dimension. The Transformer encoder is made up of three stacked encoder blocks, each of which is made up of two sequential dual transformer blocks and a patch merging layer and d is set to 64, 128, 320 and 512 respectively.

According to attention mechanism research,[63] combining spatial attention and channel attention can allow the model to catch more contextual features than it can with just a single attention. Therefore, we use a dual transformer block that combines efficient attention (spatial attention) and transpose attention (channel attention).[39]

Compared with the standard self-attention with quadratic computational complexity, the complexity of the dual transformer block is greatly reduced to the linear dimension. In Figure S2, the detailed structure is shown.

Efficient attention is proposed by Zhuoran et al.,[64] which proposes an efficient method for computing the self-attention process for the case where conventional self-attention will generate redundant context matrices. Efficient attention produces a new representation by first normalizing the key and query, then multiplying the key and value, and finally multiplying the resulting global context vector with the query. See Equation 9 for the calculation process:

$$E(Q,K,V) = \rho_q(Q)\left(\rho_k(K)^T V\right) \quad \text{(Equation 9)}$$

where Q, K and V denote query, key, and value vectors, respectively, and ρ_q and ρ_k are softmax regularization functions for queries and keys, respectively. When using ρ_q and ρ_k, this process produces an equivalent dot-product attention output.

Transpose attention, a channel attention mechanism that can effectively capture the full channel dimension, was originally proposed by El-Nouby et al.[65] as shown in Equations 10 and 11:

$$T(Q,K,V) = V\mathcal{C}_T(K,Q) \quad \text{(Equation 10)}$$

$$\mathcal{C}_T(K,Q) = \mathrm{Softmax}\left(K^T Q / \tau\right) \quad \text{(Equation 11)}$$

where $\mathcal{C}_T$ is the context vector of transpose attention, and τ is the temperature parameter. The temperature parameter was introduced to counteract the scaling of the l_2 norm applied to queries and keys before computing attention weights.

Therefore, the dual transformer block is made up of efficient attention followed by an add&norm, and transpose attention followed by an add&norm. The calculation process is shown in Equations 12, 13, 14, 15, and 16:

$$E_{block}(X,Q_1,K_1,V_1) = E(Q_1,K_1,V_1) + X \quad \text{(Equation 12)}$$

$$FFN_1(E_{block}) = FFN(LN(E_{block})) \quad \text{(Equation 13)}$$

$$T_{block}(E_{block},Q_2,K_2,V_2) = T(FFN_1(E_{block}) + E_{block}) + FFN_1(E_{block}) \quad \text{(Equation 14)}$$

$$FFN_2(T_{block}) = FFN(LN(T_{block})) \quad \text{(Equation 15)}$$

$$\mathrm{DualTransformer}(T_{block}) = FFN_2(T_{block}) + T_{block} \quad \text{(Equation 16)}$$

where $E(\cdot)$ and $T(\cdot)$ represent efficient attention and transpose attention respectively, E_{block} represents efficient attention block, and T_{block} represents transpose attention block. Q_1, K_1, V_1 are the keys, queries, and values calculated based on the input feature X, and Q_2, K_2, V_2 are the keys, queries, and values calculated based on the input of the transpose attention block. FFN stands for Mix-FFN feedforward network.[62] The computation method is presented in Equation 17:

$$FFN(X) = FC(GELU(DW-Conv(FC(X)))) \quad \text{(Equation 17)}$$

where FC stands for fully connected layer, GELU stands for GELU activation function,[66] and DW-Conv stands for depth convolution.[67]

In patch merging, we combine 2×2 patch tokens to minimize the spatial dimension while doubling the channel dimension, similar to how CNN frequently uses pooling to execute downsampling operations to gather contextual data. This enables the Transformer encoder to obtain hierarchical multi-scale representations.[18] The output of the last encoder block is subjected to layer normalization to produce the encoded sequence $z^L \in \mathbb{R}^{n\times d}$. Next, the encoder features are decoded using a progressive upsampling method.[21] Specifically, we first reshape the encoder output to $t^0 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$, which can be viewed as a 2D feature map with 4C channels. Then, to restore the spatial resolution, two successive standard upsampled convolutional layers are employed,[38] resulting in features $t^1 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$ and $t^2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, respectively. The decoder's multi-scale feature maps t^0, t^1 and t^2, along with the matching feature maps extracted by the CNN branch, will be fused.

CNN branch

To capture contextual features and preserve certain spatial details through convolutional neural networks, we use Res2Net50[52] as the backbone network of the CNN encoder. Traditionally, encoder features are progressively downsampled to $\frac{H}{32}\times\frac{W}{32}$. Combined with the advantage that the Transformer can capture the global context information, we delete the last encoding block of the original CNN, and the remaining four encoding blocks each perform a downsampling operation with a ratio of 2. We fuse the outputs of the fourth ($f^0 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$), third ($f^1 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$), and second ($f^2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$) encoding blocks with the corresponding Transformer-decoder feature maps t^0, t^1 and t^2. These three maps, namely f^0, f^1 and f^2, contain rich spatial details and contextual semantics for improving the feature representation of the Transformer decoder.

Meanwhile, to make the network retain more useful information and improve the performance of boundary segmentation during downsampling, we present the SoftPool innovative pooling algorithm in the CNN encoder. Commonly used pooling layers mainly include two types: maximum pooling and average pooling. However, a large number of experimental analyses indicated that these two processes will significantly degrade the performance of the entire network by losing the majority of the image information when pooling. Unlike maximum pooling and average pooling, SoftPool is a fast and efficient pooling method, and SoftPool can accumulate activations in an exponentially weighted manner. More information is preserved in the downsampled activation map, which leads to better classification accuracy while being computationally and memory efficient.

SoftPool uses a smooth maximum approximation of the activation value in the kernel area R, and each activation value of the index i to a_i will be multiplied by a weight w_i, which is equal to the natural exponent of the activation value divided by the natural exponent sum of all activation values. In the area adjacent to the kernel R, the output value $\tilde{a}$ of SoftPool can be obtained by summing all the weighted activation values. For the specific operation process, see Equations 18 and 19:

$$w_i = \frac{e^{a_i}}{\sum_{j\in R} e^{a_j}} \quad \text{(Equation 18)}$$

$$\tilde{a} = \sum_{i\in R} w_i * a_i \quad \text{(Equation 19)}$$

Multi-domain feature fusion module

The Transformer method designed for NLP tasks and the CNN method designed for vision tasks have completely different feature extraction methods and generation domains. We provide a unique Multi-domain Feature Fusion Module (MFFM) to efficiently combine the encoded features of CNN and Transformer. It combines self-attention and multi-domain fusion mechanism, which can realize the feature complementary function between CNN and Transformer and consists of Channel Attention Block (CAB), Spatial Attention Block (SAB), Cross-Domain Enhancement Block (CDEB), and Feature Fusion Block (FFB). The Figure S3 shows the detailed structure.

Channel Attention Block

Channel Attention Block (CAB) refers to Multi-Spectral Channel Attention.[68] It introduces more information by promoting the global average pooling (GAP) with more frequency components to promote the global information from the Transformer branch, effectively achieving a mixture of channels and self-attention. First, the input X (t^i) is divided into multiple blocks along the channel, recorded as $[X^0, X^1, \cdots, X^{n-1}]$, each of which is $X^i \in \mathbb{R}^{H\times W\times C'}$, $i \in \{0, 1, \cdots, n-1\}$, $C' = \frac{C}{n}$ and each block is assigned a two-dimensional DCT component, then the output of each block is shown in Equation 20.

$$\text{Freq}^i = \text{2DDCT}^{u,v}(X^i) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X^i_{:,h,w} B^{u,v}_{h,w} \quad \text{(Equation 20)}$$

where $[u, v]$ represents the component subscript of the two-dimensional DCT. Different frequency components are used for each block. After all, blocks are concat, the multispectral vector $\text{Freq} \in \mathbb{R}^C$ will be obtained, and then this vector will be delivered to the fully connected layer frequently used in channel attention for learning. Obtain the attention map $\hat{t}^i \in \mathbb{R}^{H\times W\times C}$ of the final output and the specific operation is shown in Equations 21 and 22:

$$\text{Freq} = \text{concat}\left(\left[\text{Freq}^0, \text{Freq}^1, \cdots, \text{Freq}^{n-1}\right]\right) \quad \text{(Equation 21)}$$

$$\text{ChannelAttention} = \text{Sigmoid}(\text{FC}(\text{Freq})) \quad \text{(Equation 22)}$$

Spatial Attention Block

Since low-level CNN features could be noisy, Compat Position Attention[69] is used as a spatial filter to improve local details and suppress irrelevant regions. It captures multiple aggregation centers with various contexts and enhances the relation-aware center-weighted sum of each spatial pixel through a simple pooling operation. First, the given feature $X \in \mathbb{R}^{C\times H\times W}$ (f^i) is input to the multi-scale pooling layer, and using a 1×1 convolutional layer, the pooling features with bin sizes of 1×1, 2×2, and 3×3 are generated. Then, each bin of the pooled features is considered a cluster center, and the features are reshaped with a bin size of $L\times L$ to $\mathbb{R}^{C\times L^2}$. Finally, the aggregation centers are obtained by concatenating the bins of all pooled features, where M is the sum of the bin numbers of all pooled features.

Next, cluster centers are adaptively integrated into each pixel according to semantic relevance. We feed features X and F into 1×1 convolutional and fully connected layers, obtaining $B \in \mathbb{R}^{\overline{C}\times H\times W}$ and $C \in \mathbb{R}^{\overline{C}\times M}$, respectively. The spatial attention map $S \in \mathbb{R}^{N\times M}$ is created using a softmax layer and matrix multiplication, $N = H\times W$ is the number of pixels. The cluster center F is then input to the fully connected layer to obtain the feature $D \in \mathbb{R}^{C\times M}$. The result is then reshaped to $\mathbb{R}^{H\times W\times C}$ by performing a matrix multiplication between it and the transpose of S. To obtain the final output attention map $\widehat{f}^i \in \mathbb{R}^{H\times W\times C}$, we multiply it by a scale parameter and execute an element-wise sum operation with the feature X. Equations 23 and 24 illustrate the standard calculation procedure.

$$s_{ji} = \frac{\exp(B_j \cdot C_i)}{\sum_{i=1}^{M} \exp(B_j \cdot C_i)} \quad \text{(Equation 23)}$$

$$\text{SpatialAttention}_j = \alpha \sum_{i=1}^{M} (s_{ji} D_i) + X_j \quad \text{(Equation 24)}$$

where s_{ji} measures the relationship between the i-th center and the j-th pixel, and α is the scale parameter, starts with a value of 0, and gradually learns to add more weights. Introducing the learnable α allows the network to first rely on cues in the local neighborhood (because this is easier) and then gradually learn to assign more weight to non-local evidence. The reason for this is that we want to learn simple tasks first and then gradually increase the complexity of the tasks.

Cross domain enhancement block

The Cross-Domain Enhancement Block (CDEB) uses the Bilinear Hadamard product[70] to model the cross-domain correlation between the features of the two transform domains of the Transformer and the CNN encoder, and after passing through the convolutional layer, cross-domain fusion features $\widehat{b}^i \in \mathbb{R}^{H\times W\times C}$ are obtained. It can enhance important information in both feature maps and suppress insignificant features. By using CDEB, we extract mutually salient features in CNN and Transformer branches to further improve accuracy.

Feature Fusion Block

The feature fusion block (FFB) generates the final multi-domain fusion feature map $m^i \in \mathbb{R}^{H\times W\times C}$ by using the residual and reshaping operations after deep-stitching the cross-domain fusion feature $\widehat{b}^i$ with the channel attention feature map $\widehat{t}^i$ and the spatial attention feature map $\widehat{f}^i$. The specific operation is shown in Equations 25 and 26:

$$m_0^i = \text{concat}\left(\widehat{b}^i, \widehat{t}^i, \widehat{f}^i\right) \quad \text{(Equation 25)}$$

$$m^i = \text{Conv}(m_0^i) + \text{PDBR}(m_0^i) \quad \text{(Equation 26)}$$

where PDBR is a block consisting of Depthwise convolution (DW-Conv) and Pointwise convolution (PW-Conv), batch normalization (BN), and rectified linear unit (ReLU),[58] used for fusing cascaded features while lowering the number of parameters. Specifically, in the Multi-domain Feature Fusion Module, we obtained the fusion feature representation of CNN and Transformer through the following operation process (Equation 27, 28, 29, and 30).

$$\widehat{t}^i = \text{ChannelAttention}\left(t^i\right) \quad \text{(Equation 27)}$$

$$\widehat{f}^i = \text{SpatialAttention}\left(f^i\right) \quad \text{(Equation 28)}$$

$$\widehat{b}^i = \text{Conv}\left(t^i W_1^i \odot f^i W_2^i\right) \quad \text{(Equation 29)}$$

$$m^i = \text{FeatureFusion}\left(\widehat{b}^i, \widehat{t}^i, \widehat{f}^i\right) \quad \text{(Equation 30)}$$

where $W_1^i \in \mathbb{R}^{D_i \times L_i}$, $W_2^i \in \mathbb{R}^{C_i \times L_i}$, $i = 0, 1, 2$, $\odot$ are Hadamard products, and Conv is the 3×3 convolutional layer.

Sandglass Block

By using the bottleneck layer, one may not only decrease the number of parameters and hence the quantity of calculation but also complete data training and feature extraction following dimensionality reduction more quickly and intuitively. By adopting two design principles, the inverted residual module[58] modifies the conventional residual bottleneck: learning to invert the residual and using a linear bottleneck, making it a commonly used bottleneck layer in the design of existing network architectures. Information loss and gradient ambiguity could yet result. So, we use the Sandglass Block,[71] a bottleneck design that conducts identity mapping and spatial transformation in higher dimensions, successfully minimizing information loss and gradient confusion. The Sandglass Block constructs shortcut connections between linear high-dimensional representations as opposed to the inverted residual block, which creates shortcuts between linear bottlenecks, and its structure protects more data transferred between blocks. Additionally, more gradients are propagated backward to better optimize network training as a result of high-dimensional residuals.[72] Additionally, Sandglass Block employs them in the extended high-dimensional feature space rather than placing the spatial convolution into the compression channel's bottleneck, which is a successful method to enhance the model's performance. To save on computing costs, pointwise convolution maintains the channel reduction and expansion process. Given an input of $F \in \mathbb{R}^{D_f \times D_f \times M}$, the output vector of the bottleneck block is $G \in \mathbb{R}^{D_f \times D_f \times M}$, and the specific operation process is shown in Equations 31 and 32:

$$\widehat{G} = \varphi_{1,p}\varphi_{1,d}(F) \quad \text{(Equation 31)}$$

$$G = \varphi_{2,d}\varphi_{2,p}(\widehat{G}) + F \quad \text{(Equation 32)}$$

where $\varphi_{i,p}$ and $\varphi_{i,d}$ are the i-th pointwise convolution and depthwise convolution, respectively. In comparison to inverting the residual block, richer feature representations can be retrieved because both depthwise convolutions are carried out in a high-dimensional space.

Boundary Refinement Module

Skin lesions usually have fuzzy lesion boundaries, and the localization results generated by conventional decoder step-by-step upsampling are far from meeting its accuracy requirements.[27]

According to studies, local context information at lesion boundaries has the greatest potential for boundary delineation, whereas boundary information has the potential to guide the work of feature extraction in segmentation by giving fine-grained boundary restrictions.[73] Therefore, we designed a Boundary Refinement Module (BRM), which accurately guides and depicts the fuzzy outline of the lesion boundary by using the fine-grained neighborhood context information and boundary information of the dual encoder fusion feature. The particular structure is displayed in Figure S4.

First, the i ($i = 1, 2$) th-level fusion feature $m^i \in \mathbb{R}^{H \times W \times C}$ from the Multi-domain Feature Fusion Module uses a series of convolutional layers and multiplication operations to generate the corresponding neighborhood prediction map $m_C^i \in \mathbb{R}^{H \times W \times C}$. It is depth concatenated with the boundary mask $m_B^i \in \mathbb{R}^{H \times W \times C}$ generated by the upsampled result $u^i \in \mathbb{R}^{H \times W \times C}$ of the $i-1$ ($i = 1, 2$) th-stage decoder. Convolutional layers are then employed to refine the boundary and correct prior predictions, driven by the contextual information, to produce the final output $o^i \in \mathbb{R}^{H \times W \times C}$. The overall process is shown in Equations 33 and 34.

$$m_C^i = \mathrm{Conv}(m^i) \otimes m^i \quad \text{(Equation 33)}$$

$$o^i = \mathrm{Conv}\left(\mathrm{concat}\left(m_C^i, m_B^i\right)\right) \quad \text{(Equation 34)}$$

The upsampling result u^i generates a binary segmentation map s^i through the process of Equations 35 and 36:

$$s^i(j) = \begin{cases} 1, & \text{if } \sigma\left[u^i(j)\right] > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad \text{(Equation 35)}$$

$$\sigma(x_i) = \frac{\exp x_m}{\sum_n \exp x_n} \quad \text{(Equation 36)}$$

where j is the index of the pixel position and σ is the softmax activation function.

The distance to the lesion boundary is then filled in at each pixel position of the lesion region using a distance transform applied to s^i.[74] By simply transposing s^i and conducting distance transformation, on the other hand, it is possible to determine the pixel distances of non-lesion regions. By normalizing and adding the two distance maps, the overall distance map is created, and then the border mask m_B^i can be acquired, as shown in Equation 37, 38, and 39:

$$\overline{s^i} = 1 - s^i \quad \text{(Equation 37)}$$

$$d^i = \frac{DT(s^i)}{\max_j DT[s^i(j)]} + \frac{DT(\overline{s^i})}{\max_j DT[\overline{s^i(j)}]} \quad \text{(Equation 38)}$$

$$m_B^i = 1 - d^i \quad \text{(Equation 39)}$$

where d^i is equal to 0 at the lesion boundary and 1 at the point furthest from the boundary, respectively, and $\overline{s^i}$ is the transpose of s^i.

Feature Adaptive Guided Module

Since we use a dual encoder based on CNN and Transformer, there is no need to use additional complex components to capture long-term dependencies. We design a Feature Adaptive Guided Module (FAGM). FAGM can learn and improve mismatched lesion boundaries while reducing the feature gap between the encoder and decoder.[75] Two parallel convolution branches make up FAGM, one of which has $k\times1$ and $1\times k$ convolution with a kernel size of 3. The other branch contains a 1×1 convolution, and the outputs of the two branches are summed element-wise to obtain the result. To better capture skin lesion boundaries, choose a convolutional layer with a kernel size of 3 to extract fine and local information. The number of nonlinear layers is increased via convolutions utilizing 1×1 kernels followed by ReLU activation layers without noticeably increasing the number of parameters or computation. We use a Feature Adaptive Guided Module in each skip connection. Given a feature input of $o^i \in \mathbb{R}^{H\times W\times C}$, as the encoder level deepens, 2, 4, and 6 basic blocks are used in FAGM0, FAGM1, and FAGM2 to match the feature distribution between the encoder and decoder, resulting in an output of $\hat{o}^i$. This is because the level of the extracted feature map is also changing from low to high as the encoder level deepens. Compared with attention gates[14] and multi-scale skip connections,[76] our FAGM is a memory-efficient and lightweight module, and its parameters are much smaller than the above two methods.

Subsequently, in the decoder block of each layer, the adaptive encoder feature $\hat{o}^i$ from FAGM and the upsampling feature u^i from the previous layer are deeply concatenated and then input into the convolutional layer and ReLU to obtain p^i, and the segmentation result S^i consists of a segmentation head generation with sigmoid activation function and 1×1 convolutional layer.

QUANTIFICATION AND STATISTICAL ANALYSIS

Dermoscopic image segmentation of skin lesions can be thought of as a binary classification task at the pixel level: background or skin lesion. Binary cross-entropy (BCE) loss and SoftDice loss are combined into a weighted total to train the complete network end-to-end. Following are the definitions for BCE loss, SoftDice loss, and weighted total loss (Equations 40, 41, 42, 43, and 44):

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} G_i \cdot \log(P_i) + (1 - G_i)\cdot\log(1 - P_i) \quad \text{(Equation 40)}$$

$$Dice = \frac{2\sum_{i=1}^{N} G_i \cdot P_i + \epsilon}{\sum_{i=1}^{N} G_i + \sum_{i=1}^{N} P_i + \epsilon} \quad \text{(Equation 41)}$$

$$Dice_b = \frac{2\sum_{i=1}^{N} (1 - G_i)\cdot(1 - P_i) + \epsilon}{\sum_{i=1}^{N} (1 - G_i) + \sum_{i=1}^{N} (1 - P_i) + \epsilon} \quad \text{(Equation 42)}$$

$$\mathcal{L}_{SoftDice} = 1 - (Dice + Dice_b)/2 \quad \text{(Equation 43)}$$

$$\mathcal{L} = \lambda\mathcal{L}_{BCE} + (1 - \lambda)\mathcal{L}_{SoftDice} \quad \text{(Equation 44)}$$

where $G_i \in \{0, 1\}$ and $P_i \in \{0, 1\}$ represent the ground truth of the i-th pixel and the probability of predicting that it belongs to the segmented area, $N = H\times W$ is the number of pixels, and $\epsilon \in R$ provides numerical stability to prevent the denominator from being 0. λ is the relative importance weight, which is set to 0.8 according to the experimental results.

We perform additional deep supervision on the decoder features t^2 at the third layer of the Transformer branch, the output m^0 of the Multi-domain Feature Fusion Module at the first layer, and the segmentation results $S^i(i = 0, 1, 2)$ at each layer of the decoder to improve the gradient flow. As a result, Equation 45 illustrates an extension of the total loss function.

$$\mathcal{L}_{sum} = \mathcal{L}(G, head(t^2)) + \mathcal{L}(G, head(m^0)) + \sum_{i=0,1,2} \mathcal{L}(G, head(S^i)) \quad \text{(Equation 45)}$$

where G represents ground truth, and the head represents the segmentation head.