Heliyon 10 (2024) e31395

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Enhancing skin lesion segmentation with a fusion of convolutional neural networks and transformer models

Zhijian Xu [a], Xingyue Guo [b], Juan Wang [b,*]

[a] *School of Electronic Information Engineering, China West Normal University, No. 1 Shida Road, Nanchong, Sichuan, 637009, China*
[b] *School of Computer Science, China West Normal University, No. 1 Shida Road, Nanchong, Sichuan, 637009, China*

ARTICLE INFO

*Keywords:*
Deep learning
Convolutional neural networks
Transformer
Skin lesion segmentation
Feature fusion
Dual branch encoder

ABSTRACT

Accurate segmentation is crucial in diagnosing and analyzing skin lesions. However, automatic segmentation of skin lesions is extremely challenging because of their variable sizes, uneven color distributions, irregular shapes, hair occlusions, and blurred boundaries. Owing to the limited range of convolutional networks receptive fields, shallow convolution cannot extract the global features of images and thus has limited segmentation performance. Because medical image datasets are small in scale, the use of excessively deep networks could cause overfitting and increase computational complexity. Although transformer networks can focus on extracting global information, they cannot extract sufficient local information and accurately segment detailed lesion features. In this study, we designed a dual-branch encoder that combines a convolution neural network (CNN) and a transformer. The CNN branch of the encoder comprises four layers, which learn the local features of images through layer-wise downsampling. The transformer branch also comprises four layers, enabling the learning of global image information through attention mechanisms. The feature fusion module in the network integrates local features and global information, emphasizes important channel features through the channel attention mechanism, and filters irrelevant feature expressions. The information exchange between the decoder and encoder is finally achieved through skip connections to supplement the information lost during the sampling process, thereby enhancing segmentation accuracy. The data used in this paper are from four public datasets, including images of melanoma, basal cell tumor, fibroma, and benign nevus. Because of the limited size of the image data, we enhanced them using methods such as random horizontal flipping, random vertical flipping, random brightness enhancement, random contrast enhancement, and rotation. The segmentation accuracy is evaluated through intersection over union and duration, integrity, commitment, and effort indicators, reaching 87.7 % and 93.21 %, 82.05 % and 89.19 %, 86.81 % and 92.72 %, and 92.79 % and 96.21 %, respectively, on the ISIC 2016, ISIC 2017, ISIC 2018, and PH2 datasets, respectively (code: https://github.com/hyjane/CCT-Net).

## 1. Introduction

Melanoma, which has a high mortality rate, is the most malignant skin lesion. Annually, 2–3 million people worldwide are diagnosed with skin cancer [1]. Computer-aided diagnosis can reduce the cost and stress of skin cancer screening [2]. Accurate

[*] Corresponding author.
*E-mail addresses:* wjuan0712@126.com, x18283338265@163.com (J. Wang).

https://doi.org/10.1016/j.heliyon.2024.e31395
Received 27 April 2023; Received in revised form 11 May 2024; Accepted 15 May 2024
Available online 17 May 2024

segmentation of skin lesions is extremely important [3,4] for their precise classification. However, typically, dermatologists manually outline the lesion area; this process is tedious, error-prone, and time-consuming. Automatic lesion separating is challenging [5] because of the indistinct borders of the lesions, low contrast of the lesions, and the presence of body hairs and shadows in the lesions. Traditional methods for manually designing features, such as gray [6–8], texture [9–11], morphology [12], region [13] and color features [14], include threshold [15,16], boundary [17,18], and region-based methods [19]. Deep learning methods have great potential in segmentation tasks. Past studies have proved that a symmetric encoding–decoding structure can extract rich features [20–22]. Convolution neural network (CNN)-based methods can provide an increased number of dermoscopic lesion image features to facilitate lesion segmentation without specific features and thresholds. Hong et al. [23] embedded a boundary-preserving structure in the U-Net, which could extract key boundary points of a skin lesion to preserve its boundary features. However, that method could not be used to extract global context information. Yun Jiang et al. [24] proposed a segmentation network based on the U-Net model that fused location information and context information of the skin lesion. Xu Q et al. [25] designed a deep and compact network, which effectively solved the problem of gradient disappearance with the increase in the number of layers and captured global context information. However, that method could cause severe overfitting by increasing the model depth. Generally, global features require a large convolution kernel or a deep network, which increases the computational complexity and loses certain shallow features important for segmentation.

Transformer is different from the method involving the expansion of the receptive field to capture contextual information, involving a multi-head attention mechanism to effectively capture long-distance dependencies of the sequences [26]. Vision Transformer (ViT) [27,28] exhibits better performance in computer vision tasks than CNN but requires a large number of data to support it. Inspired by the ViT, many scholars have combined a CNN and transformer for segmentation. Wu H et al. [29] designed a two-branch network to capture both local and global information of a skin lesion. However, the transformer they used was too deep for small datasets and for information interaction between the CNN and transformer branches. Lin H et al. [30] proposed a new cross-attention mechanism to capture the internal details of image patches and focus attention on the areas among the patches, which focuses on the information communication of the whole picture but increases computational complexity.

To address the shortcomings of the methods mentioned above, a network, which combines CNNs and transformers, named CCT-Net, was designed in this study. In designing the network, the classic encoder-decoder structure was adopted for the network, using a cross-dual-branch structure. For enhanced integration of local and global features, we designed a feature fusion module that would interact with the decoder layer to achieve accurate segmentation.

## 2. Methods

An overview of our proposed CCT-Net is illustrated in Fig. 1. CCT-Net comprises mainly three main modules: an across-fused feature encoder to capture features; a fusion module (FM) to weigh the channel features; and a decoder with skip connections that fuse encoder features to achieve segmentation. Because detailed and global features are equally important for splitting, we introduced a fusion module to merge features, and used the skip connection structure as a supplement during upsampling, to obtain segmentation results that closely represented the lesion area.

### 2.1. *Cross-fused transformer and CNNs*

Both global and local information are important to improve the accuracy. Thus, the fusion of the two types of information can

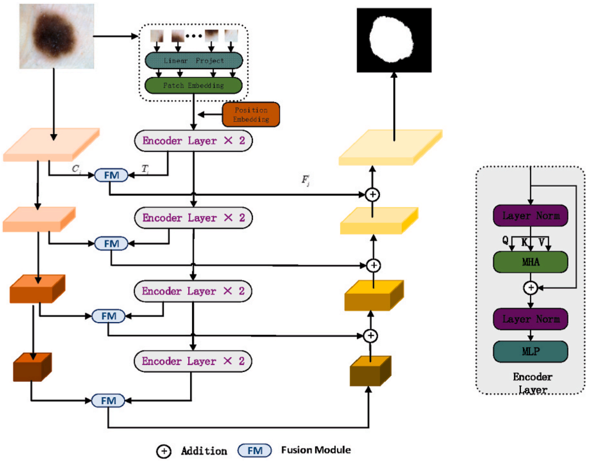

**Fig. 1.** CCT-Net network framework, containing a cross-dual-branch encoder for extracting global and local features. The FM can merge important features extracted from different channels by calculating channel attention distribution, while the skip connection forwards the encoded features to the decoder as supplementary inputs.

improve skin lesion prediction results. Because of blurred boundaries, low contrast, and irregular shape variations, skin lesion segmentation is complex. Thus, we proposed a cross-dual-branch encoder to capture the local and global features, as shown in Fig. 1. In our study, we adopted a parallel cross-fusion structure, which was in contrast to other networks that integrate CNN and Transformer by stacking them or simply using the Transformer to replace certain parts of CNN. Our network was more concise and achieved better results than the other networks. We employedResNet34 [31] as the backbone network of CNN branch for extracting local features, and designated the feature maps captured at each layer as $C_i, i = \{1, 2, 3, 4\}$ where $C_i$ represents the feature map output of the ith layer.

### 2.2. *Transformer encoder*

We introduced a Transformer branch, which could learn the global features of an image in the encoder in parallel to its CNN branch. To compensate for the lost spatial information of the image, we added the positional embedding to $X_i$, and the resulting embedding $X_j \in \mathbb{R}^{N\times C_0}$ is the input to the Transformer Layer. Typically, the core components of the Transformer layer are multi-head attention (MHA) and multilayer perceptron (MLP), as shown in Fig. 1. The MHA module updates the state of each embedding image patch over the input image sequence by capturing the global information in every layer. For every sequence $X_j$ or the result from the previous encoder layer $X_{l-1}$, the MHA module has to calculate three sequence vectors: query (q), key (k), and value (v), as shown in equation (1). The correlation between q and k is first calculated using a scoring function and then column normalized through the softmax function to ensure that all values are in the range of 0–1 and their sum is equal to 1. The result is finally multiplied by the v vector, as follows equation (2)

$$\mathrm{q} = W_q X_{l-1}, k = W_k X_{l-1}, v = W_v X_{l-1} \tag{1}$$

$$\mathrm{att}(X_{l-1}) = \mathrm{softmax}\left(\frac{Q \bullet K}{\sqrt{D_k}}\right)\mathrm{v} \tag{2}$$

where $W_q \in \mathbb{R}^{D_k\times C_0}, W_k \in \mathbb{R}^{D_k\times C_0}, and\ W_v \in \mathbb{R}^{D_k\times C_0}$ are the parameter matrices of linear mapping respectively, and $D_k$ is the dimension of the query vector. The MHA module can capture different interaction information in multiple projection spaces based on the attention mechanism with the output $\mathrm{MHA}(X_{l-1})$ expressed as equation (3).

$$\mathrm{MA}(X_{l-1}) = [att_1; att_2; att_3; \ldots; att_m] \tag{3}$$

where m implies that the self-attention mechanism has to be applied in m projection spaces. The output result obtained using an MLP, which consists of two linear layers and an activation function can be expressed as equation (4).

$$X_l = MLP(MA(X_{l-1})) + MA(X_{l-1}) \tag{4}$$

Finally, the encoder sequence result can be obtained through residual connection using the following equation (5).

$$X_{out} = Norm(MHA(X_{l-1})) + X_l \tag{5}$$

The Transformer branch has the same number of layers as the convolution branch and consists of four encoder layers (each layer contains only two encoders). The feature map of each layer can be expressed as $T_i, i = \{1, 2, 3, 4\}$ where $T_i$ represents the feature output by the ith layer.

### 2.3. *Feature fusion module*

Through the cross-dual-branch encoder, we collected the local details and global context information. To improve semantic segmentation results, we applied a concatenate operation at every layer, and doubled the number of channels. Not all channel features were equally important, and with the increase in the number of features, computation complexity also increased. As shown in Fig. 2, we introduced an FM to make full use of the channel features which can learn the dependencies between channels, further improving the segmentation accuracy and reducing the computational complexity. FM has channel-wise operation and can excite channels

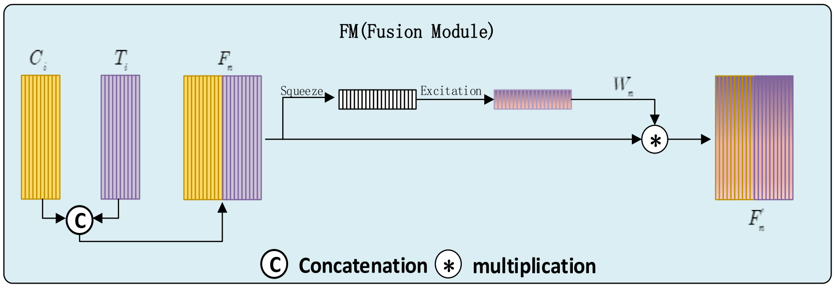

**Fig. 2.** Feature fusion module structure diagram.

containing important information and suppress channels containing unimportant information. For each channel, the FM first uses a global average pool operation. Each channel gets the result $F_n = Concatention(C_i, T_i)$ where $F_n$ represents the feature after concatenation and n is the number of channels. The weight $W_n$ is then calculated for each channel through the Full Connection layer (FC)-ReLU, activation layer-FC, and a sigmoid function. Finally, the weight $W_n$ is multiplied by the corresponding channel feature $F_n$ to obtain the new feature $F_n^{'}$. In the downsampling stage, the features extracted from the encoder stage are fused at each layer through the FM to make up for the information lost during sampling and combined with the upsampling layer using a skip connection. The four upsampling layers are implemented using deconvolution. The feature map of each layer can be expressed as $D_i, i = \{1,2,3,4\}$ $whereD_i$ represents the feature map of the ith layer. The output of the last layer of downsampling is used as the input of the first upsampling layer, expressed as $\mathrm{DLayer}_1 = F_4^{'}$. The feature map of the corresponding downsampling layer is integrated with the second layer and expressed as $\mathrm{DLayer}_i = F_j^{'} + D_k, i = \{2,3,4\}, j = \{3,2,1\}, k = \{1,2,3\}$ where $\mathrm{DLayer}_i$ represents the input feature vector sampled on the ith layer and $F_j^{'}$ represents the fusion feature of the jth layer. The result of the last layer is $D_4$, which passes through a prediction head with a sigmoid function and a convolution kernel of size 1 to provide the final segmentation result.

## 3. Loss function

For training our model for skin disease segmentation, we used two cost functions. To narrow the gap between the actual and predicted values, we used the dice loss function, which can be expressed as equation (6).

$$\mathrm{Loss1} = 1 - \frac{2 * (pred \cap true)}{pred \cup true} \tag{6}$$

where pred is the prediction, result set and the true is the actual label set. Loss2 is a binary cross-entropy loss function, which can be written as equation (7).

$$\mathrm{Loss2} = -\sum_i [(1 - p_i)\log(1 - q_i) + p_i \log(q_i)] \tag{7}$$

where $p_i$ is the actual label value and $q_i$ is the predicted value. The total loss function of the model, Loss, will depend on the two loss functions Loss1 and Loss2 as expressed equation (8).

$$\mathrm{Loss} = \omega_1 Loss1 + \omega_2 Loss2 \tag{8}$$

Through experiments, $\omega_1$ and $\omega_2$ were found to be 0.8 and 0.2, respectively.

## 4. Experiments and results

### 4.1. Datasets

To evaluate the effectiveness of our model, we conducted experiments on the three public datasets ISIC 2016 [32], ISIC 2017 [33], and ISIC 2018 [34,35], which were collected from different treatment centers andarchived by the International Skin Imaging Collaboration. The fourth dataset was PH2 [36], which was provided and described in detail by Mendonca et al. The details of the four datasets are as follows:

**ISIC 2016:** This dataset contains 1279 skin lesion images, with the test set 379 images and the training set containing 900 images.

**ISIC 2017:** This dataset contains 2750 images of melanoma, seborrheic keratosis, and other diseases. Its training, validation, and test sets contain 2,000, 150, and 600 images, respectively. **ISIC 2018:** This dataset contains a total of 2594 images, randomly divided into 1915 training, 259 validation, and 520 test.

**PH2:** This dataset contains 200 images of lesions caused by melanocyte cancer, including 80 images of common moles, 80 images of atypical moles, and 40 images of melanomas.

### 4.2. Implementation details

Our model used the Pytorch framework and was deployed on the 12G NVIDIA 3060. A large number of experiments proved that (448, 448) was the best image resolution. We uniformly resized the dataset to $(224, 224)$ when it passes into the Transformer layer. We employed Deit [37] and ResNet34 [31] as the prtrained model. The Deit was obtained by feeding the ImageNet dataset into Deit for training. We employed a cosine annealing decay function to dynamically regulate the learning rate. The corresponding mathematical expression is equation (9).

$$\mathrm{lr}_{\mathrm{new}} = \mathrm{lr}_{\mathrm{min}} + 0.5 * (\mathrm{lr}_0 - \mathrm{lr}_{\mathrm{min}}) \times \left(1 + \cos\left(\frac{E_C}{E_i}\pi\right)\right) \tag{9}$$

where $\mathrm{lr}_{\mathrm{new}}$ is the updated learning rate, $\mathrm{lr}_{\mathrm{min}}$ is the minimum learning rate, $\mathrm{lr}_0$ is the initial learning rate equal to 0.0001, $E_C$ is the number of epochs currently executed, and $E_i$ is the maximum value of the epoch equal to 200.

We determined the accuracy of the proposed model through two metrics: Intersection over Union (IoU) and Dice coefficient (Dice). The corresponding mathematical expression is equations (10) and (11).

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{10}$$

$$\mathrm{Dice} = \frac{2 * TP}{2 * TP + FP + FN} \tag{11}$$

where *the TP* is the number of pixels, classified as the skin lesion area, *FP* is the number of background pixels wrongly labeled as lesion pixels, and *FN* is the number of pixels of the lesion region wrongly labeled as background skin.

### *4.3. Ablation study*

As Table 1 shows, through ablation studies we demonstrated the requirement for two elements in the CCT-Net: the cross-dual-branch encoder and the feature fusion module, which contained a SE block to capture channel dependencies. We conducted experiments to compare prediction maps using a single encoder that included CNN or Transformer. The results are shown in Table 2, which indicates that the cross-dual-branch encoder is effective in combining the local details and global dependencies of the skin images for implementing segmentation detection, the feature fusion module can adaptively recalibrate channel-wise features, and the skip-connection architecture can replenish missing information. A comparison of the results of the ablation experiments are shown in Fig. 3, in which the green part is the actual disease labeling area, while the red area is the segmentation result of the corresponding network. From Table 2 and Fig. 3 show that the segmentation results are not accurate and that they lacked details owing to insufficient semantic information available when using the single CNN encoder branch.

The overall segmentation contour is close to the ground truth, but the boundary still has deviations owing to lack of local features coming from the single Transformer encoder branch. The results of M3 indicate that the results obtained using dual-branch encoding and by combining features at the last layer are more accurate than the two single-branch results in terms of visual effects. Table 2 shows that the two indicators are significantly better than M1-3, indicating the acceptability of our method. Because local features and global context information are equally important, the feature fusion module performs at every layer. Some information will be lost during encoding, and the encoder is connected to the decoder through skip connections to supplement each other. The analysis of the results on ISIC 2017 dataset obtained for the two indicators IoU and Dice reveal that their values obtained using M4 were higher by 7.25 % and 5.79 %, 8.68 % and 6.99 %, and 6.34 % and 4.98 %, respectively, than those obtained using M1, M2, and M3, respectively. In summary, the segmentation method proposed in our study was effective. In the proposed method, feature fusion was performed at each layer because local features and global context information carried equal importance. The information lost during sampling was recovered through skip connections.

### *4.4. Results*

To validate our pryoposed method, we conducted extensive experiments on four public datasets and compared the proposed method with five state-of-the-art segmentation methods, namely U-Net [20], U-Net++ [22], BAT [38], MultiResUNet [39], FAT [29], and Mpvit-tiny [40]. We used the metrics listed in Table 3 for evaluating the proposed method. The evaluation results indicated that the cross-fused dual-branch encoder was essential for combining local and global context features that the FM excites useful channel features and that it effectively reduces the computation complexity. In Tables 3 and 4, the values obtained for the IoU and Dice coefficient using our proposed method were 87.7 % and 93.21 %, 82.05 % and 89.19 %, 86.81 % and 92.72 % and 96.21 %, respectively, on the ISIC 2016, ISIC 2017, ISIC 2018, PH2 datasets, respectively. Moreover, visual comparisons and our subjective judgment indicate that our method generally outperforms other representative methods on all four datasets, as shown in Figs. 4–7, where the green was are the actual disease areas, and the red areas are the segmentation results of the corresponding networks. Because of the large errors associated with certain segmentation results, the red areas completely cover the green areas, and the color difference is unclear; thus, the uncolored originally labeled image was used in the comparisons. For example, the FAT and Ground Truth have different shapes in the first row in Fig. 3. U-Net, MultiResUnet, and U-Net++ are the networks that have CNN as the baselines and are more accurate in areas with clear textures than in other areas, but segmentation errors are still present with low-contrast images and blurred image boundaries. BAT incorporates a boundary-preserving structure into the Transformer, which can segment the disease areas with clear boundaries; however, it is still less accurate for areas with poorly defined edges. FAT uses dual-branch CNN and Transformer as the encoder and is better than the other methods. For an unbiased comparison, all comparison models were tested in the same

**Table 1**
Modules used in the ablation study.

| Method name | Transformer encoder | CNNs encoder | Feature fusion module |
|---|---|---|---|
| M1 | | ✓ | |
| M2 | ✓ | | |
| M3 | ✓ | ✓ | |
| M4 (CCT) | ✓ | ✓ | ✓ |

**Table 2**
Ablation experiment results.

| Method | ISIC 2017 | |
|---|---|---|
| | IoU (%) | Dice (%) |
| M1 | 74.80 | 83.40 |
| M2 | 73.37 | 82.20 |
| M3 | 75.71 | 84.21 |
| M4 (CCT) | 82.05 | 89.19 |

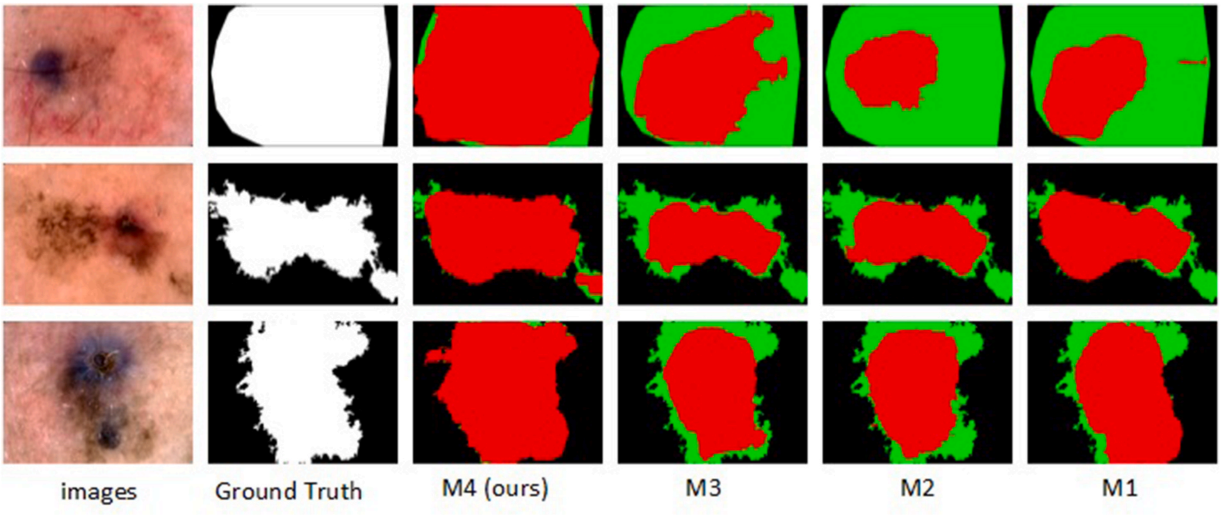

**Fig. 3.** Segmentation results of the ablation study.

**Table 3**
Statistics of the experimental results.

| Method | ISIC2016 | | ISIC2017 | | ISIC2018 | |
|---|---|---|---|---|---|---|
| | IoU (%) | Dice(%) | IoU (%) | Dice (%) | IoU (%) | Dice (%) |
| U-Net [20] | 80.00 | 88.00 | 67.40 | 77.00 | 82.60 | 89.00 |
| MultiResUnet [39] | 81.39 | 88.56 | 79.35 | 82.89 | 80.35 | 87.71 |
| U-Net++ [22] | 82.76 | 89.59 | 70.36 | 79.70 | 84.19 | 90.79 |
| BAT [38] | 76.40 | 85.30 | 77.16 | 85.00 | 85.98 | 92.23 |
| FAT [29] | 85.00 | 91.50 | 75.40 | 84.00 | 82.00 | 89.00 |
| Mpvit-tiny [40] | 84.07 | 90.44 | 74.36 | 83.15 | 76.12 | 84.04 |
| Ours (CCT) | 87.70 | 93.21 | 82.05 | 89.19 | 86.81 | 92.72 |

**Table 4**
Statistics of the experimental results.

| Method | PH2 | |
|---|---|---|
| | IoU (%) | Dice (%) |
| U-Net [20] | 78.20 | 85.80 |
| MultiResUnet [39] | 81.35 | 88.57 |
| U-Net++ [22] | 79.08 | 86.73 |
| BAT [38] | 90.00 | 94.80 |
| FAT [29] | 91.00 | 95.00 |
| Mpvit-tiny [40] | 88.14 | 93.47 |
| Ours (CCT) | 92.79 | 96.21 |

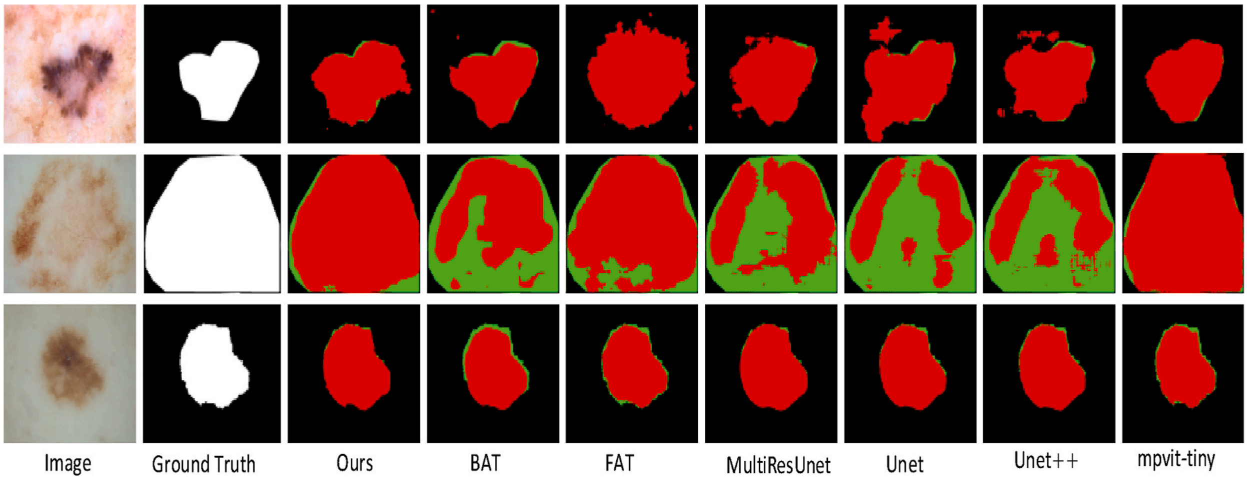

**Fig. 4.** Intuitive evaluation of the segmentation results of the ISIC 2016 dataset.

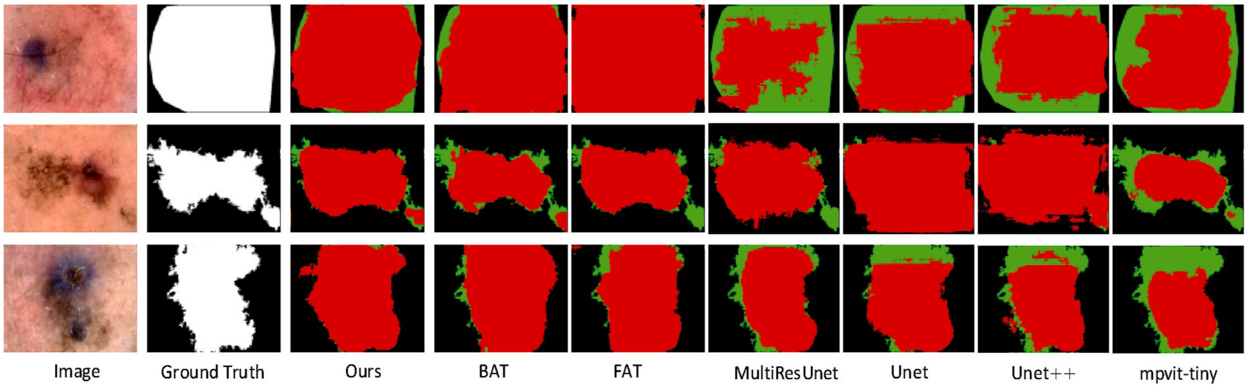

**Fig. 5.** Intuitive evaluation of the segmentation results of the ISIC 2017 dataset.

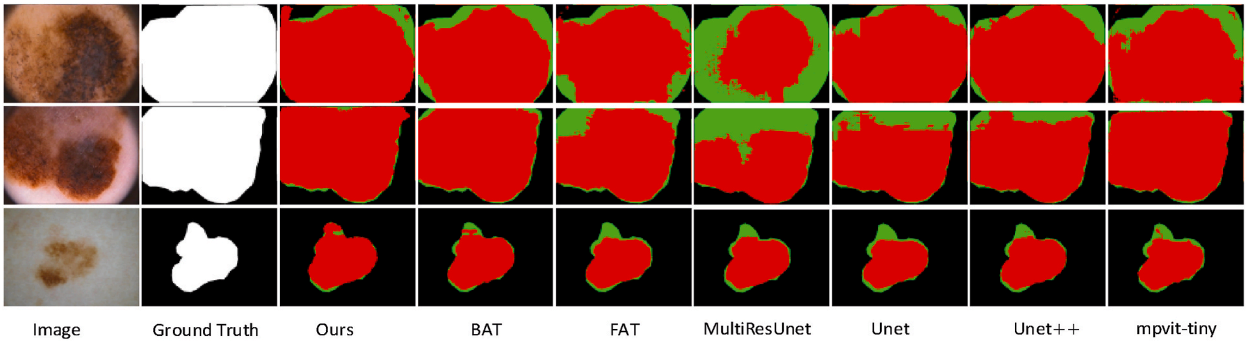

**Fig. 6.** Intuitive evaluation of the segmentation results of the ISIC 2018 dataset.

experimental environment. Due to device limitations, the batch size used in FAT was smaller than that used originally, although the evaluation index was lower than its original value only by 0.09 %.

## 5. Conclusion

In this study, we proposed a novel network, named CCT-Net, based on cross-fusion CNN and Transformer, which could effectively capture both global and local features. Extensive experiments on different datasets were conducted to confirm the effectiveness of the

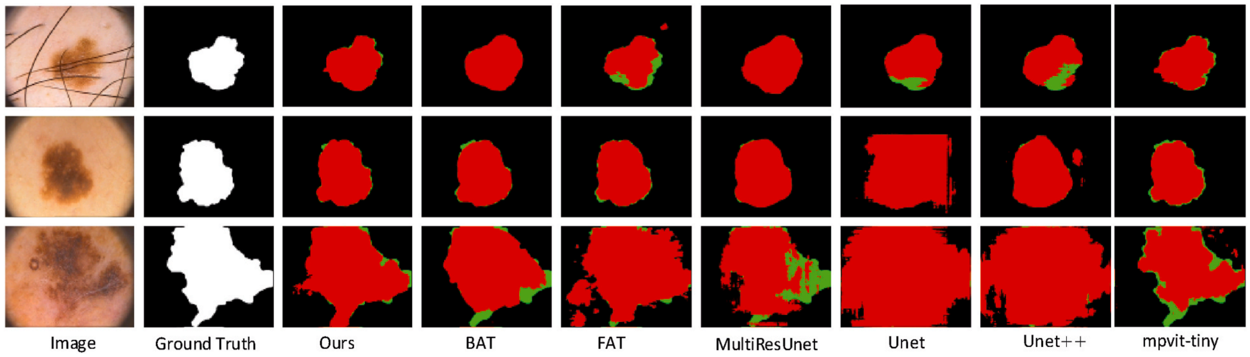

**Fig. 7.** Intuitive evaluation of the segmentation results of the PH2 dataset.

proposed method. Compared with pure convolutional coding, the method proposed by us could capture global context features of a skin lesion image more efficiently by using Transformer to significantly increase the receptive fields. We also employed a feature fusion block at each layer to excite channels containing important information and suppress other channels. Finally, the fusion features were supplemented by the upsampling layer, which effectively compensated for the information lost in the encoder layer. Through objective index evaluation and subjective visual comparison, our method could improve the accuracy of skin lesion segmentation, confirming the validation of our model. Due to the particularity of the dataset used, the types of skin diseases that the proposed network can learn are limited, limiting its practical applications. The proposed method solves only the segmentation problem and does not address the classification task. Thus, the proposed method still has limitations, which would be addressed through our future studies.

## Data availability statement

The data used in this article are four public datasets, and their download URL is:
**ISIC2016:** https://challenge.isic-archive.com/data/#2016.
**ISIC2017:** https://challenge.isic-archive.com/data/#2017.
**ISIC2018:** https://challenge.isic-archive.com/data/#2018.
**PH2:** https://www.fc.up.pt/addi/ph2%20database.html.
Everyone can download it for academic research, and the references are：

**ISIC2016:** D. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. K. Mishra, and A. Halpern.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). arXiv, 2016. doi: 10.48550/arXiv.1605.01397.

**ISIC7:** Codella, Noel C. F. and Gutman, David and Celebi, M. Emre and Helba, Brian and Marchetti, Michael A. and Dusza, Stephen W. and Kalloo, Aadi and Liopyris, Konstantinos and Mishra, Nabin and Kittler, Harald and Halpern, Allan. 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). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018:168–172. doi: 10.1109/ISBI.2018.8363547.

Codella, eronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris,18 Michael Marchetti et al. Skin lesion analysis toward melanoma detection 2018:A challenge hosted by the international skin imaging collaboration (isic). arXiv, 2019,03368. doi: hettps://doi.org/10.48550/arXiv.1902.03368.

**ISIC2018:** Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5, 180161 (2018). doi: https://doi.org/10.1038/sdata.2018.161.

**PH2:** Teresa Mendon¸ ca, Pedro M Ferreira, Jorge S Marques, Andr′; e RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013: 5437–5440. https://doi.org/10.1109/EMBC.2013.6610779.

## CRediT authorship contribution statement

**Zhijian Xu:** Writing – review & editing, Writing – original draft, Resources, Methodology, Formal analysis. **Xingyue Guo:** Validation, Supervision, Methodology, Investigation. **Juan Wang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported partly by General program of National Natural Science Foundation of China (Nos. 62176217), and the Supported by the Innovation TeamFunds of ChinaWest Normal University ( Grant Nos.KCXTD2022-3).

## References

[1] S. Khattar, R. Kaur, Computer assisted diagnosis of skin cancer: a survey and future recommendations, Comput. Electr. Eng. 104 (2022) 108431 doi:10.1016/j.compeleceng.2022.108431.

[2] M.K. Hasan, M.A. Ahamad, C.H. Yap, et al., A survey, review, and future trends of skin lesion segmentation and classification, Comput. Biol. Med. (2023) 106624, https://doi.org/10.1016/j.compbiomed.2023.106624.

[3] A. Mahbod, P. Tschandl, G. Langs, et al., The effects of skin lesion segmentation on the performance of dermatoscopic image classification, Comput. Methods Progr. Biomed. 197 (2020) 105725, https://doi.org/10.1016/j.cmpb.2020.105725.

[4] Dandu Ravi, M Vinayaka Murthy, Y.B. Ravi Kumar, Transfer learning for segmentation with hybrid classification to detect melanoma skin cancer, 4, Heliyon 9 (2023) e15416, https://doi.org/10.1016/j.heliyon.2023.e15416 [preprint]).

[5] N.K. Mishra, M.E. Celebi, An overview of melanoma detection in dermoscopy images using image processing and machine learning, arXiv:1601.07843 (2016), https://doi.org/10.48550/arXiv.1601.07843.

[6] S.U. Khan, N. Islam, Z. Jan, et al., A machine learning-based approach for the segmentation and classification of malignant cells in breast cytology images using gray level co-occurrence matrix (GLCM) and support vector machine (SVM), Neural Comput. Appl. (2022) 1–8, https://doi.org/10.1007/s00521-021-05697-1.

[7] S. Aouat, I. Ait-hammi, I. Hamouchene, A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix, Multimed. Tool. Appl. 80 (2021) 24027–24052, https://doi.org/10.1007/s11042-021-10634-4.

[8] J. Gao, B. Wang, Z. Wang, et al., A wavelet transform-based image segmentation method, Optik 208 (2020) 164123, https://doi.org/10.1016/j.ijleo.2019.164123 [preprint]).

[9] M. Waly, A. El-Hossiny, Detection of retinal blood vessels by using gabor filter with entropic threshold[J], arXiv (2020) 11508, https://doi.org/10.48550/arXiv.2008.11508.

[10] M. Baygin, T. Tuncer, S. Dogan, New pyramidal hybrid textural and deep features based automatic skin cancer classification model: ensemble DarkNet and textural feature extractor, arXiv (2022) 15090, https://doi.org/10.48550/arXiv.2203.15090.

[11] J. Ramya, H.C. Vijaylakshmi, H.M. Saifuddin, Segmentation of skin lesion images using discrete wavelet transform, Biomed. Signal Process Control 69 (2021) 102839, https://doi.org/10.1016/j.bspc.2021.102839.

[12] S. Gishkori, B. Mulgrew, Pseudo-Zernike moments based sparse representations for SAR image classification, IEEE Trans. Aero. Electron. Syst. 55 (2) (2019) 1037–1044, https://doi.org/10.1109/TAES.2018.2856321.

[13] M. Gao, H. Chen, S. Zheng, et al., Feature fusion and non-negative matrix factorization based active contours for texture segmentation, Signal Process. 159 (2019) 104–118, https://doi.org/10.1016/j.sigpro.2019.01.021.

[14] X. Zhou, T. Tong, Z. Zhong, et al., Saliency-CCE: exploiting colour contextual extractor and saliency-based biomedical image segmentation, Comput. Biol. Med. (2023) 106551, https://doi.org/10.1016/j.compbiomed.2023.106551.

[15] L. Ren, D. Zhao, X. Zhao, et al., Multi-level thresholding segmentation for pathological images: optimal performance design of a new modified differential evolution, Comput. Biol. Med. 148 (2022) 105910, https://doi.org/10.1016/j.compbiomed.2022.105910.

[16] Z. Ma, J.M.R.S. Tavares, A novel approach to segment skin lesions in dermoscopic images based on a deformable model, IEEE journal of biomedical and health informatics 20 (2) (2015) 615–623, https://doi.org/10.1109/JBHI.2015.2390032.

[17] P.M.M. Pereira, R. Fonseca-Pinto, R.P. Paiva, et al., Dermoscopic skin lesion image segmentation based on Local Binary Pattern Clustering: comparative study, Biomed. Signal Process Control 59 (2020) 101924, https://doi.org/10.1016/j.bspc.2020.101924.

[18] B. Gupta, A.K. Singh, A new computational approach for edge-preserving image decomposition, Multimed. Tool. Appl. 77 (15) (2018) 19527–19546, https://doi.org/10.1007/s11042-017-5401-7.

[19] D.A. Reddy, S. Roy, S. Kumar, et al., A scheme for effective skin disease detection using optimized region growing segmentation and autoencoder based classification, Procedia Comput. Sci. 218 (2023) 274–282, https://doi.org/10.1016/j.procs.2023.01.009.

[20] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: convolutional networks for biomedical image segmentation[C], Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (2015) 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

[21] D. Jha, P.H. Smedsrud, D. Johansen, et al., A comprehensive study on colorectal polyp segmentation with ResUNet++, conditional random field and test-time augmentation, IEEE journal of biomedical and health informatics 25 (6) (2021) 2029–2040, https://doi.org/10.1109/JBHI.2021.3049304.

[22] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, et al., Unet++: a nested u-net architecture for medical image segmentation[C], Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (2018) 3–11, https://doi.org/10.1007/978-3-030-00889-5_1.

[23] H.J. Lee, J.U. Kim, S. Lee, et al., Structure Boundary Preserving Segmentation for Medical Image with Ambiguous boundary[C], CVPR, 2020, pp. 4817–4826, https://doi.org/10.1109/CVPR42600.2020.00487.

[24] Y. Jiang, J. Dong, Y. Zhang, et al., PCF-Net: position and context information fusion attention convolutional neural network for skin lesion segmentation, Heliyon 9 (3) (2023) e13942, https://doi.org/10.1016/j.heliyon.2023.e13942.

[25] Q. Xu, Z. Ma, H.E. Na, et al., DCSAU-Net: a deeper and more compact split-attention U-Net for medical image segmentation, Comput. Biol. Med. 154 (2023) 106626, https://doi.org/10.1016/j.compbiomed.2023.106626.

[26] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need [C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010, https://doi.org/10.48550/arXiv.1706.03762.

[27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An image is worth 16x16 words: transformers for image recognition at scale, arXiv (2020) 11929, https://doi.org/10.48550/arXiv.2010.11929.

[28] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks[C], CVPR (2018) 7132–7141, doi:10.1109/TPAMI.2019.2913372.

[29] H. Wu, S. Chen, G. Chen, et al., FAT-Net: feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327, https://doi.org/10.1016/j.media.2021.102327.

[30] H. Lin, X. Cheng, X. Wu, et al., Cat: cross attention in vision transformer[C]. 2022 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2022, pp. 1–6, https://doi.org/10.1109/ICME52920.2022.9859720.

[31] K. He, X. Zhang, S. Ren, et al., Deep Residual Learning for Image recognition[C], CVPR, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[32] D. Gutman, N.C.F. Codella, M.E. Celebi, B. Helba, M.A. Marchetti, N.K. Mishra, A. Halpern, 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). arXiv (2016), https://doi.org/10.48550/arXiv.1605.01397.

[33] Noel C.F. Codella, David Gutman, Celebi, M. Emre, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern, 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), 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (2018) 168–172, https://doi.org/10.1109/ISBI.2018.8363547.

[34] Codella, eronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, 18 Michael Marchetti, et al., Skin lesion analysis toward melanoma detection 2018:A challenge hosted by the international skin imaging collaboration (isic), arXiv (2019) 03368, https://doi.org/10.48550/arXiv.1902.03368.

[35] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (2018) 180161. https://doi.org/10.1038/sdata.2018.161.

[36] Teresa Mendon¸ ca, Pedro M. Ferreira, Jorge S. Marques, Andr′ e RS. Marcal, Jorge Rozeira, Ph 2-a dermoscopic image database for research and benchmarking, in: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2013, pp. 5437–5440, https://doi.org/10.1109/EMBC.2013.6610779.

[37] H. Touvron, M. Cord, M. Douze, et al., Training data-efficient image transformers & distillation through attention[C].International conference on machine learning, PMLR (2021) 10347–10357, https://doi.org/10.1109/CVPR52688.2022.00714, 10.48550/arXiv.2012.12877 CVPR, 2022: 7287-7296.

[38] J. Wang, L. Wei, L. Wang, et al., Boundary-aware transformers for skin lesion segmentation[C], MICCAI (2021, 2021) 206–216, https://doi.org/10.1007/978-3-030-87193-2_20.

[39] N. Ibtehaz, M.S. Rahman, MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Network. 121 (2020) 74–87, https://doi.org/10.1016/j.neunet.2019.08.025.

[40] Y. Lee, J. Kim, J. Willette, et al., Mpvit: Multi-Path Vision Transformer for Dense prediction[C], CVPR, 2022, pp. 7287–7296, https://doi.org/10.1109/CVPR52688.2022.00714.