Pattern Analysis and Applications (2026) 29:22
https://doi.org/10.1007/s10044-025-01603-w

ORIGINAL ARTICLE

CRA-U: lightweight U-Net with component ranking attention for skin lesion segmentation

Zhan-Peng Ji[1] · Yan-Xu Chen[1] · Yu-Jie Xiong[1] · Xi-Jiong Xie[2] · Chun-Ming Xia[1]

Received: 8 August 2025 / Accepted: 20 December 2025 / Published online: 5 January 2026
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2026

Abstract

Melanoma is a highly malignant skin disease, for which early and accurate detection of lesions is crucial, as this significantly enhances the effectiveness of treatment. Compared with conventional manual examination methods, computer-aided diagnostic techniques based on automatic image segmentation have demonstrated significant application potential due to their high reproducibility and low-cost characteristics. Transformer is widely adopted in mainstream image segmentation methods due to its superior global image modeling capabilities. However, the self-attention in Transformers suffers from $O(n^2)$ time complexity, creating a bottleneck for real-time skin lesion segmentation. In this paper, we propose a new framework, termed **CRA-U**, which aims to accelerate the segmentation of skin lesions. First, images are inputted into two preprocessing stages to reduce the impact of interference factors on lesion region segmentation. Subsequently, the preprocessed images are fed into a modified U-Net for key region segmentation. In this process, this paper proposes to achieve global feature fusion through the **Component Ranking Attention (CR-Attention)**. This attention mechanism deeply integrates the low computational complexity of linear attention with a non-linear reweighting mechanism. Through the **Ranking Function**, CR-Attention effectively mitigates the deficiency of conventional linear attention in focusing capability. We evaluate our method on four public skin lesion datasets, demonstrating performance advantages over state-of-the-art methods. In addition, we evaluate the generalizability of our method on two typical medical image segmentation tasks. On the ISIC-2018 dataset, the proposed CRA-U model achieves 84.82% IoU and 91.56% Dice with only about 1/41 of the parameters of TransUNet. The code and datasets are available at https://github.com/jizhanpeng/CRA-U.

Keywords Skin melanoma · Medical image segmentation · Lightweight model · Component ranking attention

1 Introduction

Melanoma is a highly malignant skin disease, and early treatment is crucial for improving patient survival rates. Due to the morphological similarity between early-stage melanoma and benign skin lesions (e.g., spots, moles, normal pigmentation)[19], accurate clinical diagnosis often relies on highly skilled physicians. Currently, skin lesions are manually delineated by physicians for clinical analysis, a process that is time-consuming and error-prone. Fortunately, the

Zhan-Peng Ji and Yan-Xu Chen have contributed equally to this work.

The code and datasets are available at https://github.com/jizhanpeng/CRA-U.

✉ Yu-Jie Xiong
xiong@sues.edu.cn

Zhan-Peng Ji
jizhanpeng@sues.edu.cn

Yan-Xu Chen
yanxuchen287@hotmail.com

Xi-Jiong Xie
xiexijiong@nbu.edu.cn

Chun-Ming Xia
cmxia@sues.edu.cn

[1] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, 333 Longteng Road, Shanghai 201620, China

[2] School of Information Science and Engineering, Ningbo University, 818 Fenghua Road, Ningbo 315211, Zhejiang, China

integration of medicine with artificial intelligence (AI) has emerged as a promising solution. Deep learning-based medical image segmentation methods[14, 25] are acknowledged as effective tools for early melanoma detection. Skin images are analyzed through deep learning to automatically identify potential melanoma regions.

Convolutional neural networks (CNNs) are widely used in various computer vision tasks. However, their limitations in modeling global information become increasingly apparent. In 2017, Transformer was first applied to machine translation tasks. Due to its outstanding performance in global sequence modeling and parallel computing, Transformer-based methods now dominate artificial intelligence development. In the field of medical image segmentation, U-Net [31] and its variants are widely adopted due to their effectiveness in handling complex anatomical structures. Recent studies focus on integrating Vision Transformer (ViT) [13] with U-Net to better capture global information in medical images [4, 7].

Self-attention has $O(n^2)$ time complexity when processing a sequence of n input tokens. In medical image analysis tasks requiring real-time processing, the high computational complexity of self-attention leads to processing delays. Researchers focus on the design of lightweight attention mechanisms to balance high efficiency and global modeling capabilities, such as sparse attention [47], smaller attention windows [26], or linear attention [23]. The existing methods simplify the Transformer, but their attention mechanisms either miss critical information or fail to model long-term dependencies effectively, resulting in reduced accuracy.

In this paper, we propose CRA-U to address the issue of slow processing speed in existing skin lesion segmentation methods. Our method achieves fewer parameters and lower computational costs while maintaining performance superior to high-parameter methods. In the method, input images are first processed through a two-stage preprocessing pipeline, and then fed into a U-shaped model. In the model, we propose a large-kernel depthwise separable convolution (LKDW-Conv) to extract features. To fuse global features in the latent stage, we propose Component Ranking Attention, which addresses the limitation of linear attention in long-range dependencies modeling. Specifically, we replace the Softmax function with a carefully designed kernel function and adjust the order of matrix operations within the attention mechanism. This kernel function can adaptively assign higher weights to semantically relevant tokens through a novel non-linear reweighting strategy.

In summary, our main contributions are presented as follows:

- We propose a lightweight skin lesion segmentation method named CRA-U. The method enhances the quality of input images through a two-stage image preprocessing pipeline. In the initial stage, we use LKDW-Conv for feature extraction, while in the latent stage, we use the CR-Attention module to fuse features.
- We propose CR-Attention to address the quadratic computational complexity of self-attention, which effectively integrates the low-complexity characteristics of linear attention with a non-linear reweighting mechanism, adaptively assigning higher weights to semantically relevant tokens via a ranking function.
- We evaluate CRA-U on four public skin lesion datasets, and experimental results demonstrate that the proposed CRA-U surpasses existing approaches, while having a relatively low computational cost. Moreover, its generalization ability is evaluated on two typical medical image segmentation datasets.

2 Related work

In this section, deep learning methods for medical image segmentation based on different architectures are reviewed, including both CNN-based and Transformer-based methods, and their characteristics and advantages are discussed in detail. Moreover, current lightweight techniques for attention mechanisms are elaborated on.

2.1 CNN-based methods

CNNs are widely applied in medical image segmentation due to their ability to capture multiscale spatial structures and local features. The Fully Convolutional Network [27] marks a pivotal advance in image segmentation, replacing fully connected layers with convolutional layers to enable pixel-level segmentation while preserving spatial information. U-Net [31] further enhances performance with an encoder-decoder architecture, where the encoder extracts hierarchical features and the decoder generates pixel-wise segmentation maps. UNet++ [46] extends U-Net by introducing densely nested skip connections, enabling more comprehensive feature fusion across different network layers. This design reduces semantic gaps and improves segmentation accuracy. Attention gates are incorporated in AttU-Net [30] to dynamically recalibrate feature representations, emphasizing diagnostically relevant regions and suppressing irrelevant information, thereby enhancing both accuracy and robustness. UNeXt [34] is a lightweight MLP-based method proposed by Jeya et al., which employs the Token-MLP module in deeper layers to sequentially process features across height and width dimensions.

2.2 Mamba-based methods

Recent studies have shown that Vision Mamba has significant advantages in lightweight medical image segmentation. UltraLight VM-UNet [37] proposes a parallel Vision Mamba (PVM) layer that achieves a competitive Dice score with only 0.049M parameters; SK-VM++ [39] introduces Mamba to reconstruct skip connections in the UNet++ framework, significantly improving boundary-sensitive segmentation performance with lower FLOPs; H-vmunet [38], on the other hand, designs a high-order SS2D module that surpasses the accuracy of the original VM-UNet with fewer parameters while reducing redundant global information. These works collectively demonstrate Mamba's potential to balance efficiency and performance in dermoscopic image segmentation.

2.3 Transformer-based methods

In recent studies, Transformer-based methods are recognized as a dominant paradigm due to their exceptional ability to capture global contextual information. The integration of Transformer with U-Net was pioneered by Chen et al., who first introduced TransUNet [7]. In this model, self-attention is leveraged to encode image tokens and extract global features, enabling the architecture to capture long-range dependencies effectively while preserving U-Net's strengths in image segmentation. Swin-Unet [4] is developed by Cao et al., a fully Transformer-based method where hierarchical Swin Transformers with shifted windows are used for contextual feature extraction, and a symmetric decoder is employed for spatial resolution restoration. Huang et al. introduced MISSFormer [21], a model with redesigned Transformer modules that achieves adaptive feature alignment through the reconfiguration of the feed-forward network. APFormer [24] is proposed by Lin et al., in which adaptive pruning techniques, including query-wise and dependency-wise pruning, are employed to optimize efficiency by eliminating redundant tokens and dependencies based on adaptive thresholds.

2.4 Lightweight attention

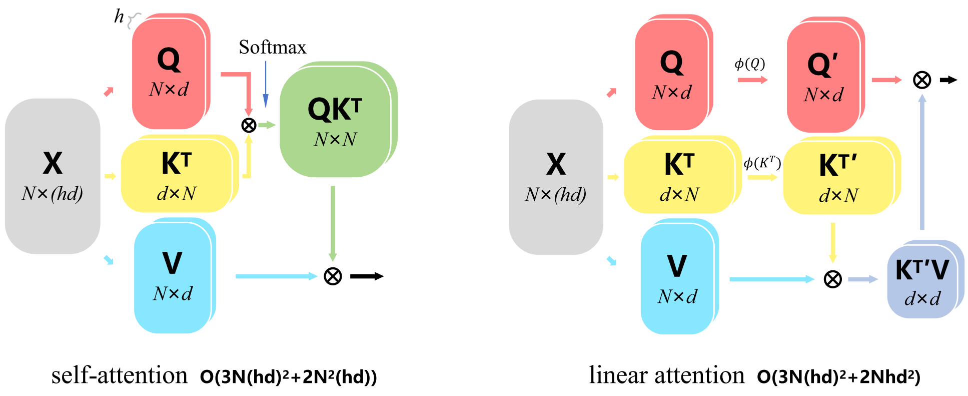
Although self-attention in Transformers performs exceptionally well on most tasks, its applications are limited by its high number of parameters. Currently, researchers are primarily developing more efficient attention modules from three dimensions: Local Attention, Token Sparsification, and Linear Attention.

Local attention Swin Transformer [26] employs window attention to enhance computational efficiency, where self-attention is performed within locally movable windows while global perception is improved through window shifting. To optimize hardware utilization, a decentralized local attention mechanism [35] is proposed by Vaswani et al., where images are partitioned into blocks with overlapping regions, and strided attention is employed for multiscale feature extraction. The approach is further advanced by Hassani et al. with pixel-wise attention [18], where self-attention is restricted to each pixel's nearest neighbors, achieving linear time and space complexity relative to image size.

Token Sparsification Its core is to identify and remove redundant parameters of the model to improve efficiency without significantly reducing performance. Zhu et al. [47] introduced L1 regularization during training to encourage dimension sparsity, automatically identifying and preserving crucial feature dimensions. Simultaneously, they pruned unimportant dimensions to reduce model parameters and computational load. Finally, they fine-tuned the model to recover the accuracy loss resulting from pruning. Chang et al. [5] significantly reduced computational complexity by generating a compact subset of semantic tokens to replace original image tokens. These semantic tokens dynamically aggregate information through self-attention, preserving local features while incorporating global information.

Linear attention To address the high computational complexity of self-attention, researchers propose replacing the Softmax function with multiple independent kernel functions [23]. Linear attention eliminates the need for precomputing the similarity matrix, reducing complexity, as shown in Fig. 1. For instance, Fast attention Via Positive

Fig. 1 Comparison between linear attention and self-attention



Orthogonal Random features is introduced by Choromanski et al. [9], where approximates the Softmax function without relying on sparsity or low-rank assumptions. Nyströmformer is proposed by Xiong et al. [40], where landmark points are selected to construct a small-scale matrix, and the Softmax matrix is approximated via singular value decomposition. A spectral angle similarity-based kernel attention is developed by You et al. [43], where the spectral angle between queries and keys is computed to define a distance function, with angles converted into similarity scores. This approach maintains global and local context modeling while improving inference efficiency. Linear attention's performance degradation in focus capability and feature diversity is analyzed by Han et al. [16], where self-attention's expressiveness is enhanced via a mapping function and rank-recovery module without increasing complexity. Moreover, Agent attention [17] is proposed, where proxy tokens are used to aggregate and broadcast information, reducing complexity while preserving global context modeling.

3 The proposed method

As shown in Fig. 3, we first use a two-stage image preprocessing, namely hair removal and contour enhancement, to enhance the quality of images input into the model. Then we design a U-shaped model where the LKDW-Conv is introduced in the initial stage and CR-Attention module is proposed in the latent stage.

3.1 Image preprocessing

The following challenges are identified in the skin lesion dataset: blurred boundaries between lesions and normal areas, interfering elements (such as hair and hair follicles), and small lesion size, as shown in Fig. 2. To address these challenges, a two-stage image preprocessing method is developed to enhance image quality.

3.1.1 Hair removal

During the first step of the preprocessing stage, the Dullrazor algorithm is employed to remove hair artifacts. This algorithm achieves the detection and suppression of hair structures through multi-scale morphological operations, thereby generating a mask of hair regions to be filled in subsequent steps. To restore the integrity and visual consistency of the image, the Telea algorithm is adopted for image inpainting, with the detailed procedure described in Algorithm 1.

```
Input: original image $I \in \mathbb{R}^{3\times h\times w}$
Output: clean hairless images $I_C \in \mathbb{R}^{3\times h\times w}$
# Step 1: convert to gray-scale image;
$I_G \leftarrow$ cvtColor($I$, color_rgb2gray );
# Step 2: perform Black-Hat Transform;
$I_B \leftarrow$ morphologyEx($I_G$, morph_blackhat, $kernel$);    // $kernel$ is a morphological structural element
# Step 3: obtain hair pixels by thresholding;
$mask_h \leftarrow$ threshold($I_B$, $t$, 255, thresh_binary);    // Pixels that exceed $t$ are regarded as hair.
# Step 4: use Telea algorithm to repair images;
$I_C \leftarrow$ inpaint($I$,$mask_h$, $r$, inpaint_Telea).
```

Algorithm 1 Dullrazor combined with Telea

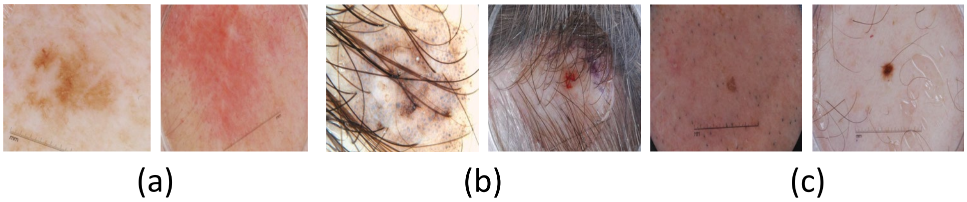


Fig. 2 Tricky examples of the skin lesion dataset. **a** The boundary is blurred. **b** There are interferences. **c** The lesion area is too small

Intermediate generated images are shown in Fig. 4. The Telea algorithm is employed for image inpainting, utilizing gradient information within a radius *r* around each pixel to reconstruct the blurred regions caused during hair removal.

3.1.2 Contour enhancement

During the second step of the preprocessing stage, manual feature extraction is implemented to highlight subtle gradient differences at lesion boundaries, enhancing the distinguishability of structural edges for improved boundary detection. Specifically, Sobel operators are utilized for edge feature extraction, where edge localization is achieved by analyzing gradient magnitude and direction to emphasize lesion contours. Additionally, this preprocessing step accelerates the model's convergence during training, reduces the number of epochs required for stabilization, and improves the efficiency of the training process.

3.2 Component ranking attention module

3.2.1 Analysis of linear attention

Self-attention, as the most critical component in Transformer, can be written as:

$$Q = xW^q, K = xW^k, V = xW^v \tag{1}$$

$$A(Q, K, V) = Softmax\left(QK^T/\sqrt{d_k}\right)V \tag{2}$$

where the input features $x \in \mathbb{R}^{N \times d}$ represent *N* tokens with dimension *d* and $W^q, W^k, W^v \in \mathbb{R}^{C \times C}$ respectively represent the projection matrices for query (Q), key (K), and value (V). The Softmax function is applied row-wise to QK^T to normalize attention weights. To compute the contextual relevance of each token to all others in the sequence, the similarity function $\mathrm{Sim}(\cdot, \cdot)$ is used to rewrite the formula as follows:

$$A_i = \sum_{j=1}^{N} \frac{\mathrm{Sim}(Q_i, K_j)}{\sum_{j=1}^{N} \mathrm{Sim}(Q_i, K_j)} V_j \tag{3}$$

where A_i denotes the output value of the *i*-th token, Q_i and K_j represent the *i*-th and *j*-th query and key vectors respectively, V_j denotes the *j*-th value vector, and N is the sequence length. In self-attention, the similarity function $\mathrm{Sim}(\cdot, \cdot)$ is concretized as:

$$\mathrm{Sim}(Q_i, K_j) = \exp\left(Q_i K_j^T/\sqrt{d_k}\right) \tag{4}$$

due to the requirement of the Softmax function for row-wise normalization of the attention scores in QK^T, the dot product between the query matrix Q and the key matrix K must first be computed to generate the QK^T score matrix. This process results in a computational complexity of $O(n^2)$ for self-attention. As the sequence length increases, the inference speed decreases significantly.

Linear attention is an effective method to reduce the computational complexity of self-attention from quadratic to linear. As shown in Fig. 1, its core idea is to introduce a kernel function $\phi(\cdot)$ for feature mapping of the query matrix Q and key matrix K, replacing the Softmax-based similarity normalization process in traditional self-attention. The formula is rewritten as:

$$A_i = \sum_{j=1}^{N} \frac{\phi(Q_i)\phi(K_j)^T}{\sum_{j=1}^{N} \phi(Q_i)\phi(K_j)^T} V_j \tag{5}$$

by leveraging the properties of matrix operations, $(QK^T)V = Q(K^T V)$, the formula can be further rewritten as:

$$A_i = \frac{\phi(Q_i)\left(\sum_{j=1}^{N} \phi(K_j)^T V_j\right)}{\phi(Q_i)\left(\sum_{j=1}^{N} \phi(K_j)^T\right)} \tag{6}$$

the overall computational complexity is reduced from $O(N^2 d)$ to $O(Nd^2)$. In general, the number of tokens N increases quadratically with the spatial resolution of the input image, and is therefore typically much larger than the token embedding dimensionality *d*.

However, in linear attention, simple feature mapping designs fail to effectively approximate the Softmax function, leading to significant degradation in sequence modeling performance, particularly for long-range dependencies. Currently, balancing computational efficiency and modeling performance in linear attention remains a significant challenge, as existing kernel-based approximations often trade accuracy for complexity reduction. Therefore, we propose the ranking function as the kernel function.

3.2.2 Ranking function

In self-attention, the Softmax function introduces a nonlinear reweighting mechanism, which normalizes attention scores into probability distributions. This mechanism assigns higher weights to relevant tokens while downweighting irrelevant ones. However, the attention matrix generated by linear attention exhibits a relatively smooth distribution, causing information-rich areas to receive insufficient focus.

```
Input: input tensor $Q \in \mathbb{R}^{h \times N \times (C/h)}$
Output: output tensor $Q_{out} \in \mathbb{R}^{h \times N \times (C/h)}$
# Step 1: Use ReLU activation function;
$Q \leftarrow \mathrm{RuLU}(Q) + \epsilon$ ;        // $\epsilon$ is the smoothing factor(set to $1e-6$)
# Step 2: Ranking tensor by last dimension;
$Q_R \leftarrow \mathrm{torch.argsort}(input = Q, dim = -1)$;
# Step 3: Hadamard product of input and $Q_R$;
$Q_1 \leftarrow Q \odot Q_R$;
# Step 4: Keep the same L2 norm as input;
$Q_{out} \leftarrow \frac{Q.norm(dim=-1)}{Q_1.norm(dim=-1)} Q_1$.
```

Algorithm 2 Ranking Function (Pytorch)

To address this, we design a ranking function that minimizes the distance between relevant query-key (Q-K) pairs while significantly expanding the distance between irrelevant ones. This relative distance adjustment effectively suppresses interactions between irrelevant Q-K pairs, thereby enhancing the attention mechanism's focus on semantically relevant tokens. The detailed steps are explained in Algorithm 2.

The ReLU function is applied before sorting to ensure the non-negativity of each component in the input tensor. Then, ascending sorting is performed along the last dimension to generate the index tensor Q_R, and the element-wise product of Q and Q_R is computed. Finally, the L2 norm of the output tensor is adjusted to match that of the input tensor. The formula for the ranking function is as follows:

$$\mathrm{RF}(Q) = \frac{\|Q\|_2}{\|(Q \odot R(Q))\|_2} \cdot (Q \odot R(Q)) \tag{7}$$

$$R(Q) = \mathrm{argsort}(\mathrm{ReLU}(Q) + \epsilon, \mathrm{dim} = -1) \tag{8}$$

In Fig. 5, five tensors with a L2 norm of 2 and dimensions of 3 (one query and four keys) are selected. Using the ranking function, the distance between relevant query-key pairs is minimized, while that between irrelevant query-key pairs is significantly maximized, as shown in Fig. 5c. Observing Q_1 and K_1, for semantically relevant query-key tensor pairs, the high similarity of ranking coefficients Q_R (assigned by the ranking function) ensures that the distance between them is only minimally increased. In contrast, for irrelevant pairs, dissimilar Q_R values lead the function to significantly increase the distance, thereby enhancing the discriminative gap between relevant and irrelevant interactions. The ranking function plays a role similar to the Softmax function, where the weights of relevant tensors are increased, while the weights of irrelevant tensors are suppressed, as shown in Fig. 5b.

3.2.3 Component ranking attention

In the design of CR-Attention, two improvements are made in addition to the ranking function.

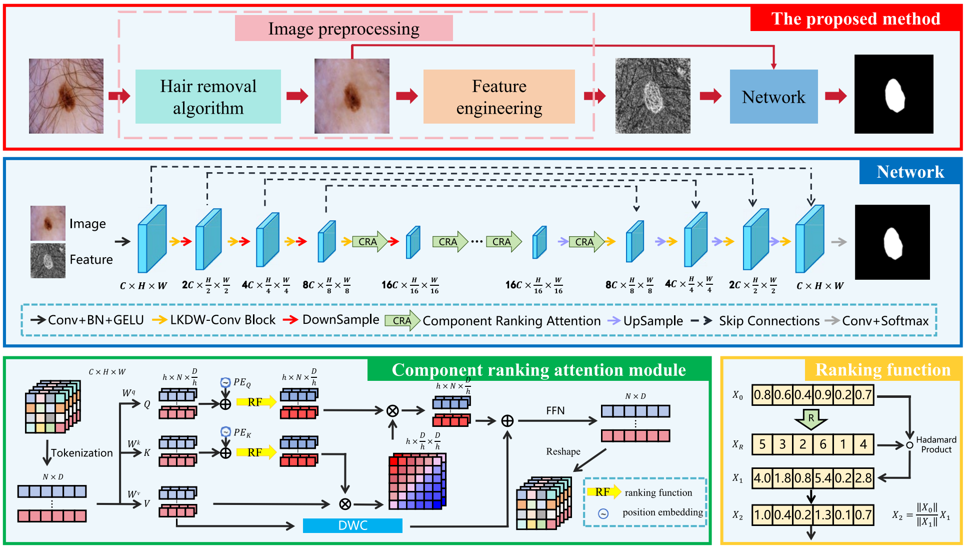

The proposed method
Image preprocessing
Hair removal algorithm
Feature engineering
Network
Network
Image
Feature
$C \times H \times W$
$2C \times \frac{H}{2} \times \frac{W}{2}$
$4C \times \frac{H}{4} \times \frac{W}{4}$
$8C \times \frac{H}{8} \times \frac{W}{8}$
CRA
$16C \times \frac{H}{16} \times \frac{W}{16}$
CRA
CRA
$16C \times \frac{H}{16} \times \frac{W}{16}$
CRA
$8C \times \frac{H}{8} \times \frac{W}{8}$
$4C \times \frac{H}{4} \times \frac{W}{4}$
$2C \times \frac{H}{2} \times \frac{W}{2}$
$C \times H \times W$
Conv+BN+GELU
LKDW-Conv Block
DownSample
CRA Component Ranking Attention
UpSample
Skip Connections
Conv+Softmax
Component ranking attention module
$C \times H \times W$
Tokenization
$N \times D$
W^q Q
W^k K
W^v V
$h \times N \times \frac{D}{h}$
PE_Q
PE_K
RF
RF
$h \times N \times \frac{D}{h}$
$h \times N \times \frac{D}{h}$
$h \times \frac{D}{h} \times \frac{D}{h}$
DWC
FFN
$N \times D$
Reshape
RF ranking function
position embedding
Ranking function
X_0 0.8 0.6 0.4 0.9 0.2 0.7
R
X_R 5 3 2 6 1 4
Hadamard Product
X_1 4.0 1.8 0.8 5.4 0.2 2.8
X_2 1.0 0.4 0.2 1.3 0.1 0.7
$X_2 = \frac{\|X_0\|}{\|X_1\|} X_1$

Fig. 3 An overview of the proposed skin lesion segmentation method. We first preprocess the input image using a joint DullRazor and Telea-based hair removal strategy, followed by Sobel-based edge enhancement to sharpen lesion boundaries. The result is fed into an improved U-Net integrating Component Ranking Attention (CRA) module and LKDW-Conv block. Within CRA, a ranking function enhances similarity among semantically relevant features, enabling accurate lesion segmentation

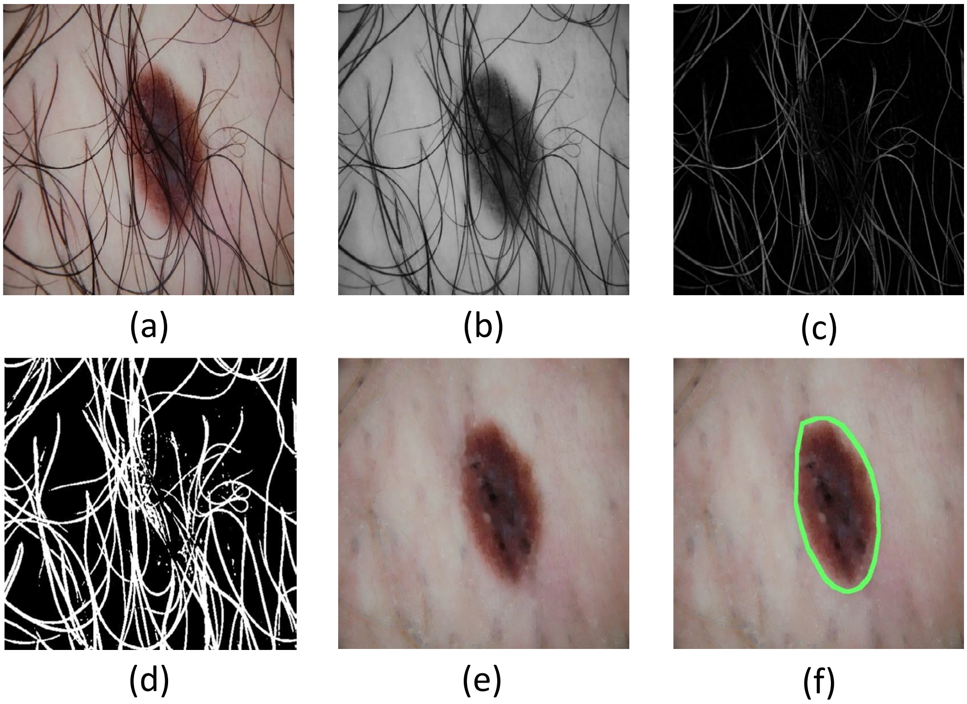
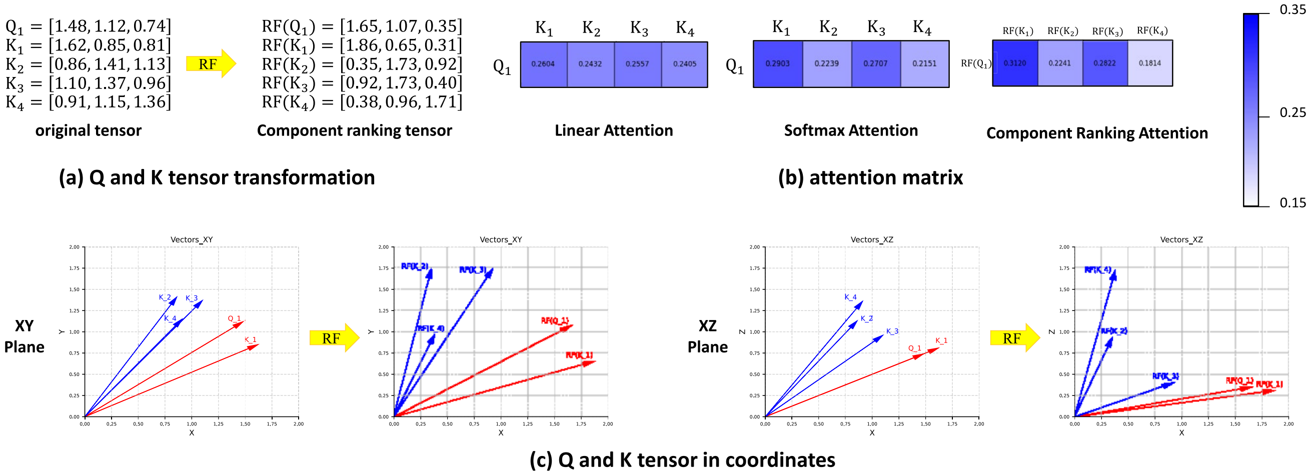


Fig. 4 Intermediate generated images. **a** Original image. **b** Grayscale image. **c** Black-hat transform image. **d** Binarized hair pixels. **e** Clean hairless images. **f** Ground truth


$Q_1 = [1.48, 1.12, 0.74]$
$K_1 = [1.62, 0.85, 0.81]$
$K_2 = [0.86, 1.41, 1.13]$
$K_3 = [1.10, 1.37, 0.96]$
$K_4 = [0.91, 1.15, 1.36]$
original tensor
RF
$RF(Q_1) = [1.65, 1.07, 0.35]$
$RF(K_1) = [1.86, 0.65, 0.31]$
$RF(K_2) = [0.35, 1.73, 0.92]$
$RF(K_3) = [0.92, 1.73, 0.40]$
$RF(K_4) = [0.38, 0.96, 1.71]$
Component ranking tensor
K_1 K_2 K_3 K_4
Q_1 0.2604 0.2432 0.2557 0.2405
Linear Attention
K_1 K_2 K_3 K_4
Q_1 0.2903 0.2239 0.2707 0.2151
Softmax Attention
$RF(K_1)$ $RF(K_2)$ $RF(K_3)$ $RF(K_4)$
$RF(Q_1)$ 0.3120 0.2241 0.2822 0.1814
Component Ranking Attention
0.35
0.25
0.15
XY Plane
Vectors_XY
K_2
K_3
K_4
Q_1
K_1
RF
Vectors_XY
RF(K_2)
RF(K_3)
RF(K_4)
RF(Q_1)
RF(K_1)
XZ Plane
Vectors_XZ
K_4
K_2
K_3
Q_1
K_1
RF
Vectors_XZ
RF(K_4)
RF(K_2)
RF(K_3)
RF(Q_1)
RF(K_1)
X
Y
Z

(a) Q and K tensor transformation

(b) attention matrix

(c) Q and K tensor in coordinates

Fig. 5 The role of ranking function in tensors similarity calculation. **a** Five tensors with L2 norm of 2 and dimensions of 3 are transformed by a ranking function. **b** Comparison of attention matrices for three different types of attention. **c** Visualization of the transformation of the five tensors in the coordinate system after undergoing ranking function conversion (red vectors represent high semantic correlation)

Dual position embedding Positional embeddings are introduced to enhance the module's capacity to capture spatial structure in image data. In conventional self-attention mechanisms, a single positional embedding is typically injected into the input features prior to linear projection into query (Q) and key (K) tensors. Because both Q and K derive from the same positionally enriched representation and undergo only linear transformations, their spatial correspondence remains intact, rendering a shared positional encoding sufficient for accurate positional awareness.

In contrast, our CR-Attention employs a non-linear ranking function $\mathrm{RF}(\cdot)$ that independently reorders and reweights the feature dimensions of Q and K based on their respective activations. While this operation strengthens semantic discrimination by adaptively emphasizing relevant components and suppressing irrelevant ones, it disrupts the original spatial alignment between Q and K. Specifically, tokens occupying identical spatial locations may—due to differences in their feature distributions—be mapped to divergent directions in the transformed space after ranking. Consequently, a shared positional embedding can no longer provide consistent spatial cues for both Q and K.

To mitigate this misalignment, we propose dual position embedding, two independent, learnable positional encoding

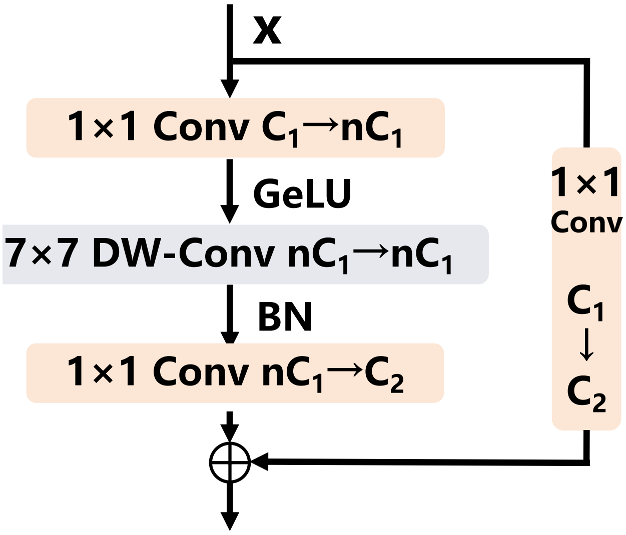


Fig. 6 Large-kernel depthwise separable convolution block

tensors, PE_Q and PE_K, are added to Q and K, respectively, after projection but before the ranking operation. Both of these positional encoding tensors are randomly initialized. This design allows each stream to develop its own spatial bias, enabling the model to reconstruct consistent relative positional relationships in the post-ranking feature space. As a result, query-key matching remains spatially coherent despite the structural perturbations introduced by the ranking function, thereby preserving the attention mechanism's capacity for precise spatial reasoning.

Feature diversity restoration In linear attention, the rank of the attention matrix is limited by the embedding dimension, reducing the model's capacity to capture diverse interactions and leading to similar output features [16, 17]. To address this limitation, a depthwise convolutional (DWC) module is applied to the value tensor, aiming to preserve the diversity of output features.

The proposed module is illustrated in the bottom-left panel of Fig. 3, with the formulation expressed as:

$$CRA = \mathrm{RF}(Q + PE_Q)\mathrm{RF}(K + PE_K)V + \mathrm{DWC}(V) \tag{9}$$

where PE_Q and $PE_K \in \mathbb{R}^{N\times D}$ represent learnable tensors.

3.3 Overall architecture of CRA-U

CRA-U adopts a U-shaped encoder-decoder architecture, as shown in Fig. 3. The preprocessed clean images and edge features are fed into the network, where multi-scale features are first extracted using the LKDW-Conv block. In the deep layers of the model, the CR-Attention module is employed to fuse shallow features and obtain refined representations. In the decoder, image resolution is recovered via upsampling. Finally, after convolution layer and Softmax operations, the lesion region $J \in \mathbb{R}^{1\times H\times W}$ is output to achieve skin lesion segmentation.

3.3.1 LKDW-Conv block

We designed an LKDW-Conv block to extract low-resolution image features, as shown in Fig. 6. In the block, a 7 × 7 convolutional kernel is employed to enlarge the receptive field, while fewer activation functions and normalization layers are used to ensure computational efficiency. In the residual path, a pointwise convolution is used to adjust the number of feature channels. It can be mathematically represented as:

$$X_1 = Conv_{1\times1}(X, C_1, nC_1) \tag{10}$$

$$X_2 = depth\text{-}Conv_{7\times7}(GELU(X_1), nC_1, nC_1) \tag{11}$$

$$Y = Conv_{1\times1}(BN(X_2), nC_1, C_2) + Res(X) \tag{12}$$

where X_1, X_2 represent intermediate features, Y represents the output feature, and C_1 and C_2 represent the number of input and output channels.

3.3.2 Details of CRA-U

In this design, a five-layer encoder-decoder architecture is adopted. Specifically, 2×2 max-pooling operations are used in the encoder to achieve downsampling, while linear interpolation is employed in the decoder for upsampling. Inspired by the UNeXt [34], the number of channels for each layer of the network is set to $(C_1, C_2, C_3, C_4, C_5) = (32, 64, 128, 160, 256)$.

3.4 Loss function

Medical image segmentation is a pixel-level binary classification task, where input data are classified into two classes [8]. Therefore, the binary cross-entropy loss function is adopted. Expressed as follows:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}\log(y_p) + (1-\hat{y})\log(1-y_p)] \tag{13}$$

where y_p represents the ground truth, $\hat{y}$ represents the predicted segmentation mask, and N is the number of pixels. More accurate segmentation results are also achieved by minimizing the Dice loss function. The Dice loss is expressed as follows:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \times \sum_{i=1}^{N} y_p \hat{y}}{\sum_{i=1}^{N} y_p + \sum_{i=1}^{N} \hat{y}} \tag{14}$$

To achieve a balance between per-pixel classification accuracy and segmentation quality, a combination of binary cross-entropy and Dice loss is employed to train CRA-U. The final loss function is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{BCE}(y_p, \hat{y}) + \lambda_2 \mathcal{L}_{Dice}(y_p, \hat{y}) \tag{15}$$

where the weights of $\mathcal{L}_{BCE}$ and $\mathcal{L}_{Dice}$ are determined by previous experimental results [8]. In our experiments, λ_1 is set to 0.5 and λ_2 is set to 1.

4 Experiments and results

In this section, the proposed method is compared with SOTA methods on four datasets of skin lesion segmentation. In addition, two typical medical image tasks are used to evaluate the versatility of CRA-U. Subsequently, ablation studies are performed to examine the specific contributions of each module in the proposed method.

4.1 Datasets

Four datasets for the skin lesion segmentation task, ISIC-2016 [15], ISIC-2017 [11], ISIC-2018 [10] and PH^2 [29], are selected to evaluate our method. In addition to the skin lesion task, the 2018 DSB [3] and BUSI [1] are also selected to demonstrate generalization. Details are as follows:

ISIC-2016 This series of datasets originates from the competition organized by the International Skin Imaging Collaboration and has become the main benchmark for evaluating skin lesion segmentation methods. ISIC-2016 contains a total of 1279 skin lesion images, with 900 used for training and 379 for testing.

ISIC-2017 ISIC-2017 contains 2000 training images, 150 validation images and 600 test images.

ISIC-2018 ISIC-2018 contains 2594 images. Three random splits are conducted, with the training and test sets comprising 80% and 20% respectively.

PH^2 The dataset includes 200 skin lesion images with a resolution of 768 × 560. Due to its small data size, we only perform cross-dataset evaluation with this dataset.

2018 DSB This dataset, comprising 670 CT images, is primarily designed for cell nucleus localization in medical images.

BUSI This dataset is used to address the breast ultrasound task, including 647 ultrasound images of benign and malignant breast lesions.

4.2 Experimental settings

Since original images vary in size, all images are resized to 256 × 256. To address overfitting due to limited training samples, four data augmentation techniques are applied during training: random horizontal flipping, random vertical flipping, random cropping, and random rotation within the range of $(-\pi/2, \pi/2)$.

All experiments are conducted with PyTorch on an NVIDIA GeForce RTX 2080Ti. The Adam optimizer is employed with a learning rate of 0.001 and momentum of 0.9. Moreover, a cosine annealing learning rate scheduler with a minimum learning rate of 0.00001 is used. The batch size is set to 8. A total of 400 epochs of training are carried out. To ensure fairness and follow previous data partitioning methods, the dataset is randomly split three times (random_ state was set to 0, 1, and 2, respectively) at a ratio of 80 - 20. The final results are reported as the mean and standard deviation of three independent runs.

4.3 Evaluation criteria

Five widely used segmentation metrics are chosen to evaluate performance, including IoU, Dice coefficient, Accuracy (ACC), Sensitivity (SE), and Specificity (SP). The formulas for these evaluation metrics are as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{16}$$

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{17}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{18}$$

$$SE = \frac{TP}{TP + FN} \tag{19}$$

$$SP = \frac{TN}{TN + FP} \tag{20}$$

where True Positive (TP) denotes the number of pixels correctly classified as the key region. False Positive (FP) refers to pixels incorrectly labeled as the key region but belonging to non-key regions. Correspondingly, True Negative (TN) represents pixels correctly classified as non-key regions, while False Negative (FN) is the number of key region pixels misclassified as non-key.

Table 1 Comparison with various SOTA methods on ISIC-2018 and PH^2 dataset

Method	Params (in M)	GFLOPs	ISIC-2018					ISIC 2018 → PH^2				
			IoU (%)	Dice (%)	ACC (%)	SE (%)	SP (%)	IoU (%)	Dice (%)	ACC (%)	SE (%)	SP (%)
U-Net [31]	31.13	55.84	81.38 ± 0.63	88.80 ± 0.51	95.78 ± 0.19	90.94 ± 0.52	96.88 ± 0.27	84.07	90.89	93.89	95.43	93.91
AttU-Net [30]	34.88	51.02	81.77 ± 0.64	88.83 ± 0.41	95.72 ± 0.24	90.83 ± 0.71	96.88 ± 0.51	86.84	92.58	95.49	97.82	92.77
UCTransNet [36]	66.43	32.93	82.66 ± 0.84	89.45 ± 0.63	96.02 ± 0.68	91.17 ± 0.66	96.92 ± 0.68	83.92	90.44	94.56	95.98	92.42
DAE-Former [2]	48.07	25.97	82.91 ± 0.67	89.62 ± 0.45	95.99 ± 0.32	91.11 ± 0.45	97.29 ± 0.50	86.07	92.14	94.94	96.27	94.47
EIU-Net [44]	14.15	18.91	83.6	90.2	96.7	90.7	96.7	84.7	91.6	94.4	96.9	93.7
UNeXt [34]	1.47	2.28	82.93 ± 0.16	90.21 ± 0.15	95.52 ± 0.09	91.70 ± 0.18	96.94 ± 0.28	84.26	91.39	93.83	95.73	92.96
MISSFormer [21]	42.46	7.18	82.77 ± 0.73	89.47 ± 0.58	95.92 ± 0.24	91.20 ± 0.78	96.77 ± 0.55	84.86	91.42	94.49	97.07	92.47
GFANet [45]	24.20	7.68	83.66 ± 0.10	90.13+0.13	96.29+0.10	90.75 ± 0.71	97.79 ± 0.36	86.17	92.28	95.47	96.86	**94.81**
ADF-Net [22]	36.21	8.29	84.52 ± 0.44	90.82 ± 0.30	**96.70 ± 0.13**	92.34 ± 0.34	97.41 ± 0.26	86.84	92.58	95.49	**97.82**	92.77
UConvNeXt [8]	1.76	2.44	83.93 ± 0.11	90.89 ± 0.11	96.62 ± 0.07	92.84 ± 0.16	96.31 ± 0.26	86.51	92.33	95.60	96.61	93.72
SLP-Net [42]	0.75	2.30	80.61 ± 0.41	88.21 ± 0.28	93.87 ± 0.14	89.30 ± ± 1.21	95.36 ± 0.82	79.17	87.20	91.62	92.16	93.14
I^{2}U-Net [12]	27.49	8.26	83.66 ± 0.33	90.10 ± 0.35	95.76 ± 0.21	92.51 ± 0.21	96.01 ± 0.33	85.70	91.83	95.77	96.57	93.58
UltraLight [37]	0.049	0.060	82.29 ± 0.52	90.28 ± 0.46	95.97 ± 0.41	89.96 ± 0.67	97.76 ± 0.25	84.29	91.47	94.20	96.40	93.15
Ours (CRA-U)	2.51	2.28	**84.82 ± 0.17**	**91.56 ± 0.21**	96.01 ± 0.18	**92.91 ± 0.26**	**98.15 ± 0.26**	**86.98**	**92.88**	**95.86**	96.77	93.98

4.4 Results on the ISIC 2018 and PH^2

Quantitative Analysis To validate the effectiveness of training and testing, experiments are performed on the ISIC 2018 dataset. The ISIC 2018-trained model is also tested on the PH^2 dataset. Table 1 compares our method with 12 SOTA approaches. The same setup as UNeXt is used for fairness, with other methods' results obtained from their papers. Observations reveal that our method not only has an advantage in computational complexity but also maintains competitive performance across various segmentation evaluation metrics. Compared to the classic U-Net, our approach improves computational efficiency 10× while achieving a nearly 3% improvement in all evaluated segmentation metrics.

Qualitative Analysis To visually compare the performance of methods, we conduct visual evaluations on five open-source approaches and select two sets of typical challenging examples for illustration. When applied to challenging cases, U-Net and U-Net++ struggle to accurately segment key regions with ambiguous boundaries, as shown in Fig. 7. Additionally, these methods misclassify some normal tissues as lesions. While AttU-Net and TransUNet incorporate advanced attention mechanisms to enhance feature representation, they still show minor performance gaps compared to our approach. Compared with the comparable parametric method UNeXt, CRA-U demonstrates superior detail preservation due to CR-Attention. After adding preprocessing, we obtain a more excellent processing effect because interferences such as hair are eliminated.

4.5 Results on the ISIC 2016 and ISIC 2017

Quantitative Analysis The effectiveness of the proposed method is verified on the ISIC 2016 and ISIC 2017 datasets. Table 2 compares the proposed method with 10 SOTA methods across multiple performance metrics. Compared with U-Net, our method achieves significant improvements of 1.95% in IoU, 1.63% in Dice Score, 0.73% in Accuracy, and 1.73% in Sensitivity. Moreover, the number of model parameters is reduced to 1/12 of that in U-Net, and the computational complexity is decreased by 90%. It is obvious that our method not only surpasses the SOTA method in performance, but also maintains an excellent inference efficiency.

Qualitative Analysis In Fig. 8, a visual comparison between ISIC 2016 and ISIC 2017 is presented. Compared with the other five competing methods, CRA-U demonstrates higher skin lesion segmentation accuracy. Even for challenging skin lesion images with different background colors and irregular shapes, the contours extracted by CRA-U closely match the ground truth. In addition, when

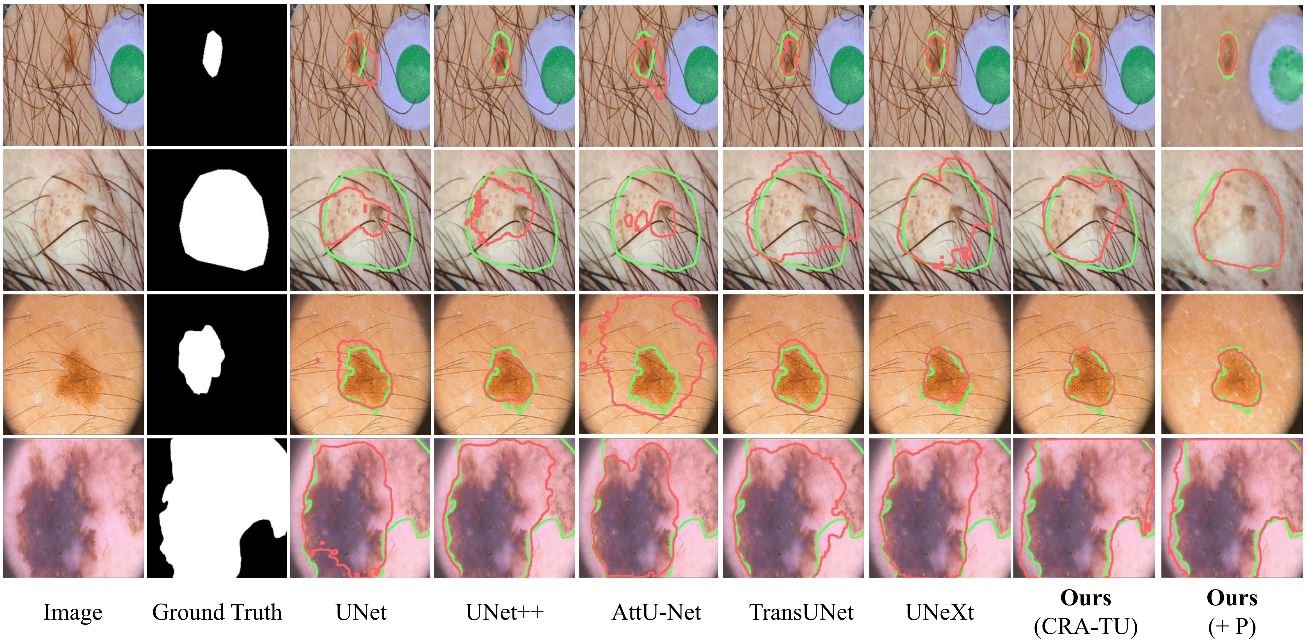


Fig. 7 Visual comparisons are conducted with SOTA methods on the ISIC-2018 and PH^2 dataset. The green contour represents the ground truth, while the red contour indicates the segmentation results of the respective methods (the upper two groups are from ISIC-2018, and the lower two groups are from PH^2) (P stands for preprocessing)

preprocessing is incorporated, detail preservation improves significantly.

4.6 Results on medical image dataset

Quantitative Analysis The generalization ability of CRA-U is verified by evaluating two additional benchmark datasets: 2018 DSB and BUSI. Tables 3 and 4 show the comparison results with SOTA methods on the two datasets. Compared to UNeXt with comparable parameter counts, IoU, Dice, ACC, and SE of our method improved by 1.24%, 0.85%, 0.46%, and 2.32% respectively on 2018 DSB. On BUSI, these metrics increased by 2.6%, 3.17%, 1.00% and 2.97% respectively.

Qualitative Analysis In Fig. 9, the upper two groups are from the cell nucleus segmentation task, while the lower two groups are from the breast ultrasound image segmentation task. Due to inductive bias of convolution operations, convolution-based methods may produce inaccurate predictions in locally similar texture regions. AttU-Net and TransUNet outperform traditional methods in detail handling. CRA-U significantly reduces segmentation errors in non-lesion regions through long-range dependency modeling. Even for tumors with different scales, irregular shapes and blurred boundaries, or cells with different background colors and irregular shapes, the contours extracted by CRA-U closely match the ground truth.

4.7 Ablation analysis

Ablation analysis is conducted on ISIC-2016 and ISIC-2018 to demonstrate the effectiveness of each component of the proposed method. Dice and IoU are used as evaluation metrics. Additionally, model parameters, computational complexity, and inference time for a single 256 × 256 image are presented.

Model architecture As shown in Table 5, we use a U-shaped encoder-decode model as the baseline, with only a 3 × 3 convolution module retained in each layer. When the LKDW-Conv module is used to replace 3 × 3 convolution blocks, the model parameters and computational complexity are significantly reduced. When incorporating the CR-Attention module into the baseline model, performance is improved by approximately 3%. When integrating all proposed modules into the baseline model, our final method achieves significant performance improvements compared to the baseline on the ISIC-2018 dataset: IoU is improved by 4.02% and Dice by 3.63%.

The role of CR-Attention We replace the CR-Attention module in CRA-U with three different attention mechanisms to demonstrate the effectiveness of CR-Attention, as shown in Table 6. In terms of processing efficiency, CR-Attention requires only 80% of the computational cost of self-attention, achieving a 10% improvement in inference speed. Meanwhile, it delivers superior performance compared to self-attention. While linear attention is more lightweight, its performance degrades. In contrast, the

Table 2 Comparison with various SOTA methods on ISIC-2016 and ISIC-2017 dataset

Method	Params (in M)	GFLOPs	ISIC-2016					ISIC 2017				
			IoU (%)	Dice (%)	ACC (%)	SE (%)	SP (%)	IoU (%)	Dice (%)	ACC (%)	SE (%)	SP (%)
U-Net [31]	31.13	55.84	84.78	91.21	95.52	93.22	96.05	75.61	84.2	92.79	81.73	97.56
AttU-Net [30]	34.88	51.02	85.23	91.37	95.47	91.63	96.4	75.75	84.29	93.04	81.2	**98.22**
DAE-Former [2]	48.07	25.97	85.76	91.80	95.85	93.29	96.55	77.29	85.46	93.62	83.22	97.82
MISSFormer [21]	42.46	7.18	85.39	91.45	95.65	93.06	95.89	76.11	84.46	93.04	82.13	97.4
UCTransNet [36]	66.43	32.93	85.74	91.64	95.78	92.98	95.61	76.54	84.96	93.18	83.14	98.07
UNeXt [34]	1.47	2.28	85.42	91.97	93.00	93.49	95.77	75.05	85.44	93.11	82.21	96.57
EIU-Net [44]	14.15	18.91	85.5	91.9	95.9	91.8	94.3	77.1	85.5	93.7	84.2	96.8
GFANet [45]	24.20	7.68	85.92	91.78	96.04	92.95	**97.25**	77.75	85.74	**93.97**	81.37	97.87
UConvNeXt [8]	1.76	2.44	86.02	91.93	95.55	92.48	96.75	77.39	85.59	93.08	83.02	97.49
BCF-CNN +LS [20]	10.78	22.39	85.1	91.2	95.3	93.8	94.4	77.0	85.5	93.0	**88.2**	94.9
Ours (CRA-U)	2.51	2.28	**86.73**	**92.84**	**96.25**	**95.01**	96.95	**77.92**	**86.31**	93.32	84.85	96.39

proposed CR-Attention maintains low computational overhead while achieving optimal performance. To verify that CRA-U can reduce the computational complexity of the attention module from quadratic complexity with respect to sequence length N to linear complexity, we measured the FLOPs of the attention module within the input resolution range from 128×128 to 1024×1024 (corresponding to $N = 64, 256, 1024, 4096$). The results are shown in Fig. 11. The FLOPs in standard self-attention is mainly dominated by two terms, the linear projection overhead from Q, K, and V transformations $3N(hd)^2$; and the quadratic attention computation overhead from QK^T and weighted summation $2N^2(hd)$. When N is small, the linear term accounts for a large proportion of the total FLOPs, resulting in a subquadratic growth characteristic (when N increases from 64 to 256, FLOPs only increase by about 6.5 times, far lower than the theoretical 16 times). As N increases, quadratic terms gradually become dominant, and the growth of FLOPs approaches N^2 (when N increases from 1024 to 4096, FLOPs increase by approximately 13.6 times). In contrast, CRA-U exhibits a FLOPs growth factor of approximately 4.0 times across all test intervals, consistent with the linear growth of N; and at $N = 4096$, its FLOPs are only 6.5% of standard self-attention. This confirms that CRA-U successfully compresses core attention computation to linear complexity.

The role of preprocessing After performing hair removal processing on images, the segmentation performance of the model demonstrated significant improvements across different datasets: on the ISIC-2018 dataset, the IoU improved by 0.6%, and the Dice coefficient increased by 0.57%; on the ISIC-2016 dataset, the IoU improved by 0.69%, and the Dice coefficient increased by 0.5%. Introducing edge features have a critical impact on training efficiency. As shown in Fig. 10, when edge features are fed into the model, its performance reaches its peak at approximately 180 training epochs and converges around 220 epochs. By contrast, the model without edge features requires approximately 250 epochs to reach its performance peak, and needs more than 100 additional epochs to converge compared with the former. This result indicates that providing additional edge features to the model, particularly for Transformer-based high-training-cost models, can effectively reduce training overhead and significantly enhance training efficiency.

The role of dual positional encoding We conduct an ablation study to evaluate the impact of different positional encoding strategies on segmentation performance. As shown in Table 7, removing positional encoding entirely (w/o PE) leads to a clear performance drop, confirming its necessity for spatial reasoning. While a shared positional encoding (Shared PE) offers marginal gains, and adding PE only to K yields moderate improvement, our proposed dual

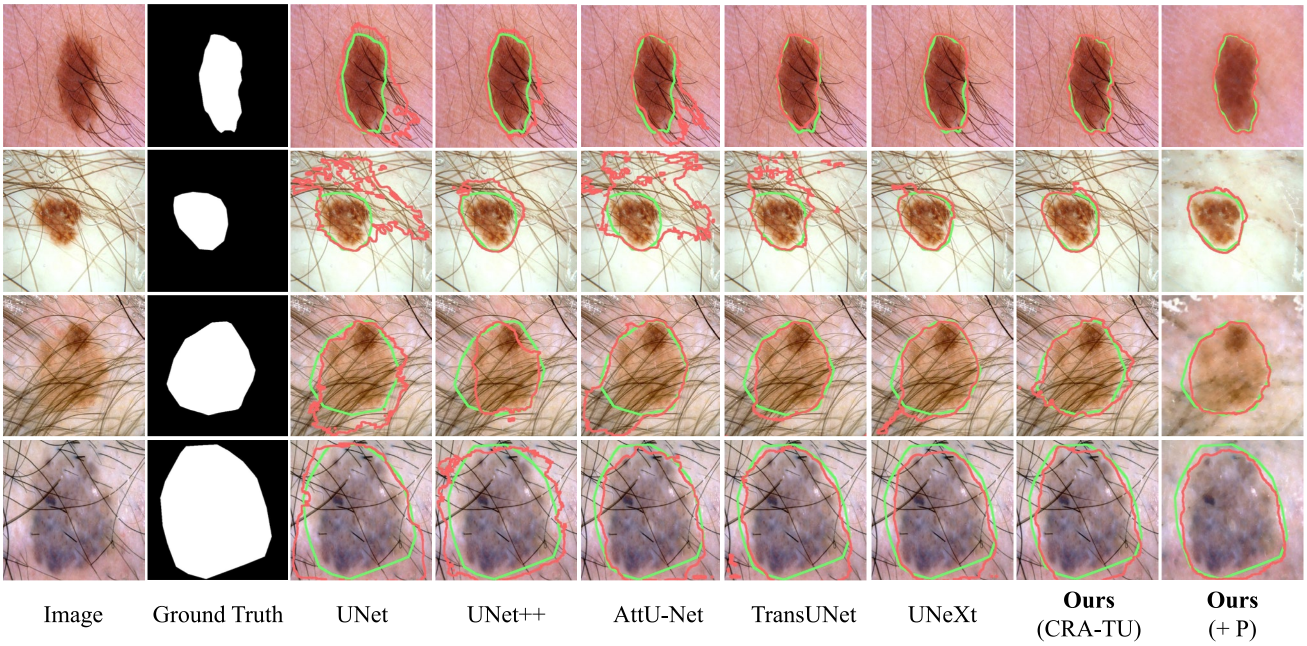


Fig. 8 Visual comparisons are conducted with SOTA methods on the ISIC-2016 and ISIC-2017 dataset (the upper two groups are from ISIC-2016, and the lower two groups are from ISIC-2017)

Table 3 Comparison with SOTA methods on the 2018 DSB

Method	Params (in M)	IoU (%)	Dice (%)	ACC (%)	SE (%)
U-Net [31]	31.13	80.8 ± 0.13	88.7 ± 0.09	95.5 ± 0.05	92.0 ± 0.11
AttU-Net [30]	34.88	81.6 ± 0.15	88.7 ± 0.14	95.3 ± 0.05	91.8 ± 0.14
TransUNet [7]	105.32	82.1 ± 0.14	89.5 ± 0.10	95.4 ± 0.05	90.6 ± 0.12
UNeXt [34]	1.48	84.83 ± 0.17	91.74 ± 0.14	97.38 ± 0.31	92.37 ± 0.21
DDA-Net [33]	6.83	84.52 ± 0.10	91.82 ± 0.06	–	91.39 ± 0.09
MSRF-Net [32]	18.38	85.34 ± 0.08	92.24 ± 0.05	–	94.02 ± 0.07
DAE-Former [2]	49.21	85.67 ± 0.12	92.07 ± 0.09	96.88 ± 0.49	92.86 ± 0.22
DCSAU-Net [41]	2.6	85.0 ± 0.11	91.4 ± 0.08	95.9 ± 0.05	92.4 ± 0.08
Ours (CRA-U)	2.51	**86.07 ± 0.27**	**92.59 ± 0.19**	**97.84 ± 0.30**	**94.69 ± 0.29**

Table 4 Comparison with SOTA methods on the BUSI

Method	Params (in M)	IoU (%)	Dice (%)	ACC (%)	SE (%)
U-Net [31]	31.13	65.36 ± 1.81	74.35 ± 1.72	95.22 ± 0.29	77.86 ± 2.24
Unet++ [46]	9.163	63.34 ± 3.29	76.40 ± 2.52	95.39 ± 0.86	77.51 ± 2.97
AttU-Net [30]	34.88	62.65 ± 2.74	75.51 ± 2.66	95.49 ± 0.49	73.68 ± 4.65
TransUNet [7]	105.32	71.47 ± 0.98	79.00 ± 0.79	96.28 ± 0.51	80.78 ± 1.63
UNeXt [34]	1.48	69.07 ± 0.42	79.47 ± 0.23	95.48 ± 0.62	79.47 ± 0.23
AAU-Net [6]	29.65	68.82 ± 0.44	77.51 ± 0.68	–	81.10 ± 0.52
DAE-Former [2]	49.21	68.87 ± 0.32	77.97 ± 0.39	96.28 ± 0.50	78.82 ± 0.32
ATFE-Net [28]	24.96	69.73	82.46	96.32	**82.78**
Ours (CRA-U)	2.51	**71.67 ± 0.62**	**82.64 ± 0.42**	**96.48 ± 0.22**	82.44 ± 0.52

PE design–assigning independent learnable embeddings to Q and K–achieves the best results across all metrics. Specifically, it improves IoU by + **1.29%**, Dice by + **0.55%**, ACC by + **0.33%**, SE by + **1.84%**, and SP by + **1.36%** compared to the strongest single-PE variant. This gain validates our hypothesis: the non-linear sorting operation in CR-Attention disrupts the spatial alignment between Q and K, making a single shared encoding insufficient. By decoupling positional information into branch-specific embeddings, our model better adapts to the transformed feature geometry, enabling more accurate attention computation and improved segmentation fidelity–particularly critical for fine-grained medical image analysis.

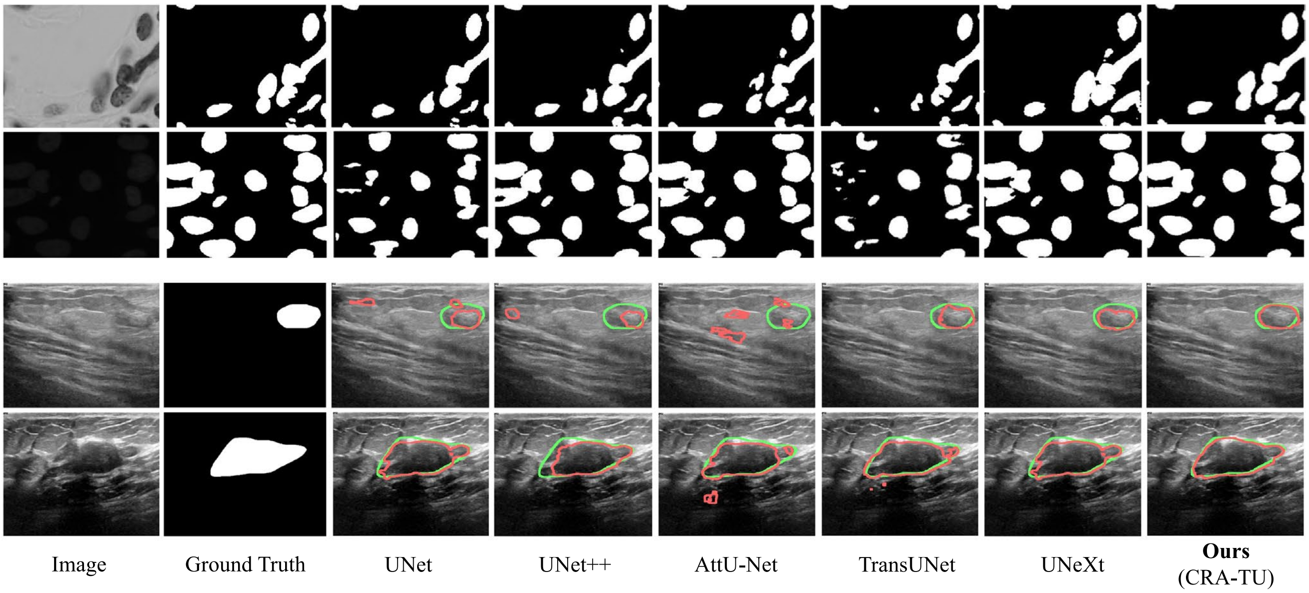


Fig. 9 Visual comparisons are conducted with SOTA methods on the 2018 DSB and BUSI (the upper two groups are from 2018 DSB, and the lower two groups are from BUSI)

Table 5 Ablation analysis on the ISIC-2018 and ISIC-2016

Method	Params (in M)	GFLOPs	Inference speed (ms)	ISIC-2018		ISIC-2016	
				IoU (%)	Dice (%)	IoU (%)	Dice (%)
Conv stage (baseline)	1.42	1.53	5.56	80.75	87.93	83.98	90.62
Conv → LKDW-Conv	0.99	1.19	7.57	80.57 (– 0.18%)	87.51 (– 0.42%)	83.84 (– 0.14%)	90.52 (– 0.14%)
Baseline + CR-Attention	2.93	1.76	10.77	83.93 (+ 3.36%)	90.89 (+ 2.96%)	85.74 (+ 1.76%)	92.02 (+ 1.40%)
CRA-U(not preprocessed)	2.51	2.28	11.83	84.22 (+ 3.47%)	91.09 (+ 3.16%)	86.04 (+ 2.06%)	92.33 (+ 1.71%)
CRA-U	2.51	2.28	11.83	**84.82 (+ 4.02%)**	**91.56 (+ 3.63%)**	**86.73 (+ 2.75%)**	**92.84 (+ 2.22%)**

Table 6 Advantages of CR-Attention over other attention mechanisms

Method	Params (in M)	GFLOPs	Inference speed (ms)	ISIC-2018		ISIC-2016	
				IoU (%)	Dice (%)	IoU (%)	Dice (%)
CRA-U → self-attention	2.4	2.83	13.13	83.99	90.82	85.84	92.12
CRA-U → linear attention	2.4	1.96	11.5	83.48 (– 0.51%)	90.57 (– 0.25%)	85.43 (– 0.41%)	91.79(–0.3%)
CRA-U → windows multi-head self-attention	2.4	2.43	12.86	84.06 (+ 0.07%)	90.93 (+ 0.11%)	85.93 (+ 0.09%)	92.19 (+ 0.07%)
CRA-U (not preprocessed)	2.51	2.28	11.83	**84.22 (+ 0.23%)**	**91.09 (+ 0.27%)**	**86.04 (+ 0.20%)**	**92.33 (+ 0.21%)**

5 Conclusion

In this work, we present a skin lesion segmentation method named CRA-U. We first use a two-stage image preprocessing strategy to enhance the quality of input images. In the initial stage, we propose LKDW-Conv to extract features. In the latent stage, we propose CR-Attention to capture global features. This module not only inherits the low-complexity design of linear attention but also integrates a non-linear reweighting mechanism. The limitation of linear attention's focusing capability is addressed via a ranking function. We evaluate the proposed method on four publicly available skin lesion datasets, and experimental results demonstrate its superiority over existing SOTA methods. Despite the good performance of the CRA-U model, several limitations warrant discussion. First, while the two-stage preprocessing workflow effectively improves the quality of the input image, it relies on hand-designed operations. This may limit its generalization ability when dealing with skin lesions with highly diverse imaging conditions or containing artifacts not covered by the algorithm design. Second, at the model architecture level, although the proposed Component Ranking Attention (CRA) effectively enhances semantic relevance and improves attention focus by introducing a

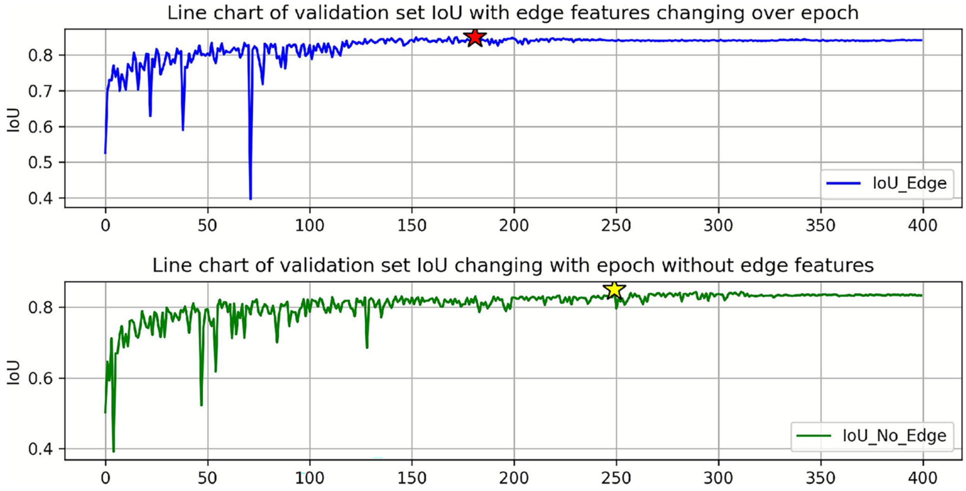


Fig. 10 Comparison of the impact of edge features on model training

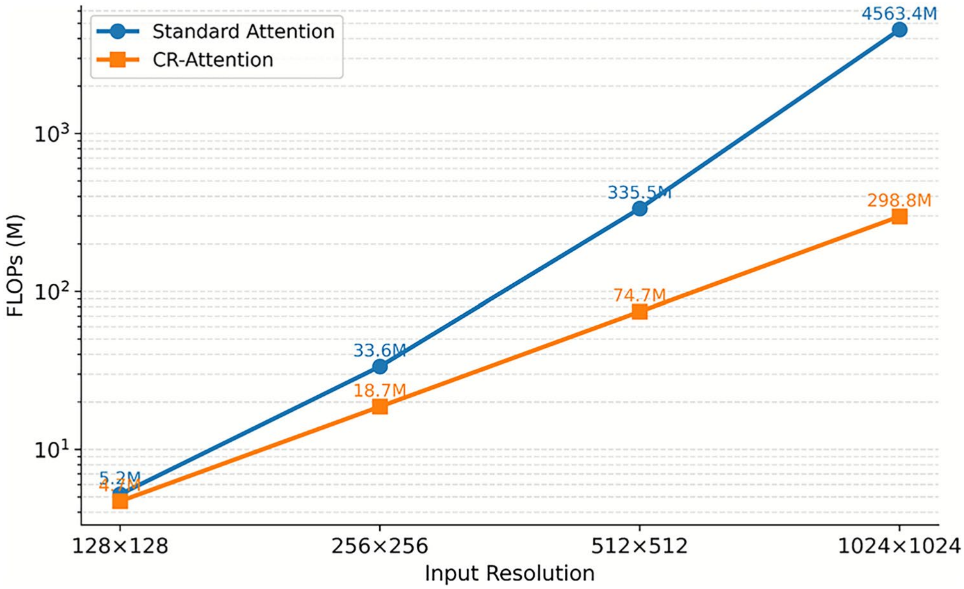


Fig. 11 FLOPs comparison between Standard Attention and CR-Attention across increasing input resolutions

Table 7 Ablation study on different positional encoding strategies in the proposed CR-Attention module

Method	IoU (%)	Dice (%)	ACC (%)	SE (%)
w/o PE	83.53	90.92	95.56	90.48
Shared PE	83.63 (+ 0.10%)	91.01 (+ 0.09%)	95.60 (+ 0.04%)	91.06 (+ 0.58%)
PE on K only	83.96 (+ 0.43%)	91.21 (+ 0.29%)	95.67 (+ 0.11%)	90.47 (−0.01%)
Dual PE (ours)	84.82 (+ 1.29%)	91.56 (+ 0.64%)	96.01 (+ 0.45%)	92.91 (+ 2.43%)

ranking function, this function applies independent ranking operations to the Query and Key tensors, disrupting their collaborative structure in the original feature space. Even with the introduction of dual learnable positional encoding to compensate for spatial alignment, this design may still lead to the model's inability to fully model the subtle spatial dependencies relied upon by traditional attention mechanisms.

Author contributions Y.X designed the framework and supervised the study. Y.C conducted experiments, data analysis, and manuscript writing. Z.J assisted in implementation and setup. X.X and C.X guided the experiments and contributed to manuscript proofreading.

Funding This work was supported by the Science and Technology Commission of Shanghai Municipality (21DZ2203100) and National Natural Science Foundation of China (62006150).

Data availability The code and datasets used in this article can be obtained from https://github.com/jizhanpeng/CRA-U.

Declarations

Conflict of 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.

References

1. Al-Dhabyani W, Gomaa M, Khaled H et al (2020) Dataset of breast ultrasound images. In: Data in brief, p 104863. https://doi.org/10.1016/j.dib.2019.104863
2. Azad R, Arimond R, Aghdam EK et al (2023) DAE-Former: dual attention-guided efficient transformer for medical image segmentation, pp 83–95
3. Caicedo J, Goodman A, Karhohs K et al (2019) Nucleus segmentation across imaging experiments: the 2018 data science bowl. Nat Methods 16(12):1247–1253. https://doi.org/10.1038/s41592-019-0612-7
4. Cao H, Wang Y, Chen J et al (2022) SWIN-UNet: UNet-like pure transformer for medical image segmentation, pp 205–218
5. Chang S, Wang P, Lin M et al (2023) Making vision transformers efficient from a token sparsification view, pp 6195–6205
6. Chen G, Li L, Dai Y et al (2023) AAU-Net: an adaptive attention U-Net for breast lesions segmentation in ultrasound images. IEEE Trans Med Imaging 42(5):1289–1300. https://doi.org/10.1109/TMI.2022.3226268
7. Chen J, Lu Y, Yu Q et al (2021) TransUNet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306
8. Chen YX, Xiong YJ, Qiu XH et al (2024) Harmonious parameters and performance: lightweight convolutional stage and local feature weighted fusion MLP for medical image segmentation. Biomed Signal Process Control 98:106726. https://doi.org/10.1016/j.bspc.2024.106726
9. Choromanski K, Likhosherstov V, Dohan D et al (2020) Rethinking attention with performers
10. Codella N, Rotemberg V, Tschandl P et al (2019) Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368
11. Codella NC, Gutman D, Celebi ME 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). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, pp 168–172
12. Dai D, Dong C, Yan Q et al (2024) I2U-Net: a dual-path u-net with rich information interaction for medical image segmentation. Med Image Anal 97:103241. https://doi.org/10.1016/j.media.2024.103241
13. Dosovitskiy A, Beyer L, Kolesnikov A et al (2020) An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
14. Gao Q, Wang Y, Zhou F et al (2025) MSFM-UNET: enhancing medical image segmentation with multi-scale and multi-view frequency fusion. Pattern Anal Appl 28(1):17. https://doi.org/10.1007/s10044-024-01384-8
15. Gutman D, Codella NCF, Celebi E et al (2016) Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) arXiv preprint arXiv:1605.01397
16. Han D, Pan X, Han Y et al (2023) Flatten transformer: vision transformer using focused linear attention. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5961–5971
17. Han D, Ye T, Han Y et al (2023) Agent attention: on the integration of Softmax and linear attention. arXiv preprint arXiv:2312.08874. https://doi.org/10.48550/arXiv.2312.08874
18. Hassani A, Walton S, Li J et al (2023) Neighborhood attention transformer, pp 6185–6194
19. Hoang NK, Nguyen DH, Tran TT et al (2025) DermoMamba: a cross-scale mamba-based model with guide fusion loss for skin lesion segmentation in dermoscopy images. Pattern Anal Appl 28(3):128. https://doi.org/10.1007/s10044-025-01506-w
20. Huang L, Zhao YG, Yang TJ (2024) Skin lesion image segmentation by using backchannel filling CNN and level sets. Biomed Signal Process Control 87:105417. https://doi.org/10.1016/j.bspc.2023.105417
21. Huang X, Deng Z, Li D et al (2023) MISSFormer: an effective transformer for 2d medical image segmentation. IEEE Trans Med Imaging 42(5):1484–1494. https://doi.org/10.1109/TMI.2022.3230943
22. Huang Z, Deng H, Yin S et al (2024) ADF-Net: a novel adaptive dual-stream encoding and focal attention decoding network for skin lesion segmentation. Biomed Signal Process Control 91:105895. https://doi.org/10.1016/j.bspc.2023.105895
23. Katharopoulos A, Vyas A, Pappas N et al (2020) Transformers are RNNS: fast autoregressive transformers with linear attention, pp 5156–5165
24. Lin X, Yu L, Cheng KT et al (2023) The lighter the better: rethinking transformers in medical image segmentation through adaptive pruning. IEEE Trans Med Imaging 42(7):2325–2337. https://doi.org/10.1109/TMI.2023.3247814
25. Liu S, Wang P, Lin Y et al (2024) SMRU-Net: skin disease image segmentation using channel-space separate attention with depthwise separable convolutions. Pattern Anal Appl 27(3):93. https://doi.org/10.1007/s10044-024-01307-7
26. Liu Z, Lin Y, Cao Y et al (2021) Swin transformer: hierarchical vision transformer using shifted windows, pp 10012–10022
27. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation, pp 3431–3440
28. Ma Z, Qi Y, Xu C et al (2023) ATFE-Net: axial transformer and feature enhancement-based CNN for ultrasound breast mass segmentation. Comput Biol Med 153:1289–1300. https://doi.org/10.1016/j.compbiomed.2022.106533
29. Mendonça T, Ferreira PM, Marques JS et al (2013) Ph 2-a dermoscopic image database for research and benchmarking, pp 5437–5440
30. Oktay O, Schlemper J, Folgoc LL et al (2018) Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999
31. Ronneberger O, Fischer P, Brox T (2015) U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015. Springer, Cham, pp 234–241
32. Srivastava A, Jha D, Chanda S et al (2021) MSRF-Net: a multi-scale residual fusion network for biomedical image segmentation. IEEE J Biomed Health Inform 26(5):2252–2263. https://doi.org/10.1109/JBHI.2021.3138024
33. Tomar NK, Jha D, Ali S et al (2021) DDANet: dual decoder attention network for automatic polyp segmentation, pp 307–314
34. Valanarasu JMJ, Patel VM (2022) UNEXT: MLP-based rapid medical image segmentation network. In: Medical image computing and computer assisted intervention—MICCAI 2022. Springer, Cham, pp 23–33
35. Vaswani A, Ramachandran P, Srinivas A et al (2021) Scaling local self-attention for parameter efficient visual backbones, pp 12894–12904
36. Wang H, Cao P, Wang J et al (2022) UCTransNet: rethinking the skip connections in U-Net from a channel-wise perspective with transformer, pp 2441–2449
37. Wu R, Liu Y, Ning G et al (2024) Ultralight VM-UNET: parallel vision mamba significantly reduces parameters for skin lesion segmentation. Patterns 6:101298

38. Wu R, Liu Y, Liang P et al (2025) H-VMUNET: high-order vision mamba UNet for medical image segmentation. Neurocomputing 624:129447
39. Wu R, Pan L, Liang P et al (2025) SK-VM++: mamba assists skip-connections for medical image segmentation. Biomed Signal Process Control 105:107646
40. Xiong Y, Zeng Z, Chakraborty R et al (2021) Nyströmformer: a nyström-based algorithm for approximating self-attention, pp 14138–14148
41. Xu Q, Ma Z, He N et al (2023) DCSAU-Net: a deeper and more compact split-attention U-Net for medical image segmentation. Comput Biol Med 154:106626. https://doi.org/10.1016/j.compbiomed.2023.106626
42. Yang B, Zhang R, Peng H et al (2024) SLP-Net: an efficient lightweight network for segmentation of skin lesions. Biomed Signal Process Control 101:107242. https://doi.org/10.1016/j.bspc.2024.107242
43. You H, Xiong Y, Dai X et al (2023) Castling-ViT: compressing self-attention via switching towards linear-angular attention at vision transformer inference, pp 14431–14442
44. Yu Z, Yu L, Zheng W et al (2023) EIU-Net: enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation. https://doi.org/10.1016/j.compbiomed.2023.107081
45. Zhang Y, Zhang J (2023) GFANet: group fusion aggregation network for real time stereo matching. IEEE Robot Autom Lett 8(7):4251–4258. https://doi.org/10.1109/LRA.2023.3280818
46. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N et al (2018) UNet++: a nested U-Net architecture for medical image segmentation, pp 3–11
47. Zhu M, Tang Y, Han K (2021) Vision transformer pruning. arXiv preprint arXiv:2104.08500

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.