scientific reports

OPEN

ScaleFusionNet: transformer-guided multi-scale feature fusion for skin lesion segmentation

Saqib Qamar[1,2]✉, Syed Furqan Qadri[3], Roobaea Alroobaea[4], Goram Mufarah Alshmrani[5], Mohd Fazil[6] & Richard Jiang[5]✉

Melanoma is a malignant tumor that originates from skin cell lesions. Accurate and efficient segmentation of skin lesions is essential for quantitative analysis but remains a challenge owing to blurred lesion boundaries, gradual color changes, and irregular shapes. To address this, we propose ScaleFusionNet, a hybrid model that integrates a Cross-Attention Transformer Module (CATM) and adaptive fusion block (AFB) to enhance feature extraction and fusion by capturing both local and global features. We introduce CATM, which utilizes Swin transformer blocks and Cross Attention Fusion (CAF) to adaptively refine feature fusion and reduce semantic gaps in the encoder-decoder to improve segmentation accuracy. Additionally, the AFB uses Swin Transformer-based attention and deformable convolution-based adaptive feature extraction to help the model gather local and global contextual information through parallel pathways. This enhancement refines the lesion boundaries and preserves fine-grained details. ScaleFusionNet achieves Dice scores of 92.94%, 91.80%, and 95.37% on the ISIC-2016, ISIC-2018, and HAM10000 datasets, respectively, demonstrating its effectiveness in skin lesion analysis. Simultaneously, independent validation experiments were conducted on the PH^2 dataset using the pretrained model weights. The results show that ScaleFusionNet demonstrates significant performance improvements compared with other state-of-the-art methods. Our code implementation is publicly available at https://github.com/sqbqamar/ScaleFusionNet.

Keywords Transformer, Skin Lesion, Image Segmentation, Information fusion, feature enhancement

The incidence of melanoma has risen significantly in recent decades due to increasing environmental pollution and ultraviolet radiation[1]. This trend has attracted significant attention from the global medical community. Early detection is essential for improving treatment outcomes and patient survival rates. Traditional diagnostic methods, such as clinical observation and tissue biopsy, are limited by subjectivity and invasiveness, making them unsuitable for large-scale screening. Medical image segmentation offers a noninvasive and high-precision alternative that provides clinicians with more detailed and accurate information. This technology has the potential to significantly enhance the early detection and treatment of skin cancer.

With the rapid advancement of deep learning, convolutional neural networks (CNNs) have achieved notable success in medical image segmentation tasks such as skin lesion segmentation. Among these, the U-Net model[2] stands out as a pioneering framework. U-Net has demonstrated remarkable performance in medical image segmentation and established a U-shaped architectural paradigm. However, CNN-based methods often struggle to capture global contextual information effectively due to the inherent limitations of convolutional operations. This limitation is particularly evident in medical image segmentation tasks with significant inter-sample variability, such as skin lesion segmentation. To address this challenge, researchers have explored various strategies, including the use of large kernel convolutions, dilated convolutions, and other techniques aimed at expanding receptive fields[3–7]. For example, Hu et al.[8] improved receptive fields by using self-attention, while Tang et al.[9] proposed a model that used large convolutional kernels and fusion to achieve promising results in tasks such as breast nodule ultrasound image segmentation. Inspired by the ConvNeXt model[10], Han et al.[11]

[1]Division of Robotics, Perception and Learning(RPL), Department of Intelligent Systems, KTH Royal Institute of Technology, 10044 Stockholm, Sweden. [2]Faculty of Computing and Information Technology (FCIT), Sohar University, 311 Sohar, Oman. [3]BGI Research, Hangzhou 310030, China. [4]Department of Computer Science, College of Computer and Information Technology, Taif University, P. O. Box 11099, 21944 Taif, Kingdom of Saudi Arabia. [5]School of Computing and Commutations, Lancaster University, Lancaster LA1 4YW, UK. [6]Department of Information Technology, College of Computer and Information Sciences, IMSIU, Riyadh, Kingdom of Saudi Arabia. ✉email: sqamar@su.edu.om; r.jiang2@lancaster.ac.uk

developed a method for medical image segmentation using large kernel convolutions and they successfully implemented it for tasks such as retinal vessel segmentation. Despite these advancements, simply increasing the size of the convolutional kernels may not fully resolve the challenge of modelling global features because the fundamental constraints of the receptive field remain.

Recently, Transformers[12] have achieved notable success in both the natural language processing and computer vision domains by using global contextual information in feature extraction. Cai et al.[13] unveiled the BiADATU-Net that combined Transformer and feature adaptation modules, which resulted in promising outcomes in several publicly accessible skin lesion segmentation datasets. Zhang et al.[14] developed DAE-Former, a pure Transformer U-shaped medical image segmentation model, harnessing efficient Transformers steered by dual attention, similar to Swin-Unet[15], and exhibited commendable performance in diverse image segmentation datasets, including ISIC-2018[16]. However, because Vision Transformers can only output single-scale feature representations, they lack the ability to capture multi-scale information in two-dimensional images[17,18]. Consequently, transformer-based medical image segmentation models may struggle to seamlessly integrate multi-scale information, leading to insufficient attention to lesion regions and incomplete decoding of feature details. Additionally, there is a significant issue with medical image segmentation models based on the U-Net design architecture. Although skip connections in U-Net transmit multi-scale information between different stages to the decoder, a semantic gap issue may arise when there is a considerable semantic difference between encoder and decoder. To address this, some studies have attempted to mitigate this issue by improving skip connections. For instance, UNet++[19] and MISSFormer[20] aim to achieve the fusion of multi-scale information between different stages through dense skip connections and contextual bridges. Nevertheless, this study argues that the differently sized feature maps transmitted through skip connections represent macroscopic multi-scale information that is easily observable, and such methods have limited effectiveness in enhancing the model's ability to integrate multi-scale information. In particular, in the skin lesion segmentation task, the lesion edges are often irregular, with colors gradually fading from the center, and the progressive compression of feature maps leads to the loss of fine details, retaining only a macro-level focus. This can affect the model performance to some extent.

To address the challenges of skin lesion segmentation, this study proposes ScaleFusionNet, a model that integrates an AFB and CATM for enhanced feature extraction and fusion. ScaleFusionNet employs a hierarchical Swin Transformer-based encoder, where patch embedding and Swin Transformer blocks[21] extract the multi-scale features. The decoder utilizes AFBs, which combine Swin Transformer and deformable convolution features to refine feature integration and improve lesion boundary preservation. To bridge the semantic gap between the encoder and decoder features, the CATM uses cross-attention, which allows high-level decoder features to guide low-level skip connections. The experimental results demonstrate that ScaleFusionNet achieves highly competitive performance in skin lesion segmentation. The key contributions of this study are as follows:

- We have proposed ScaleFusionNet for skin lesion segmentation, based on a hybrid architecture combining CNNs and Transformers, which outperforms other state-of-the-art methods.
- We have introduced AFB, which integrates both Swin Transformer-based and deformable convolution-based feature extraction, enabling the model to capture both local and global contextual information.
- We developed the CATM to effectively reduce the semantic gap and enhance the interaction between the encoder and the decoder.

Related work

CNNs for medical image segmentation

Recently, CNNs have achieved success in different domains due to their powerful feature extraction capabilities. This success is particularly evident in medical image segmentation. In 2015, Ronneberger et al. introduced U-Net, a CNN-based model designed specifically for medical image segmentation, which has become foundational in this field. Zhou et al.[22] proposed UNet++, which introduces nested dense skip connections to address the semantic gap between the encoder and decoder. Li et al.[23] developed an H-DenseUNet, a U-shaped model that enhances intra-slice and inter-slice representations through hybrid dense connections, demonstrating effective performance in liver tumor segmentation tasks. Saqib et al.[24] proposed multi-scaled architecture using separable convolution for brain tumor segmentation. Oktay et al.[25] developed Attention U-Net, a model that focuses on important areas by using attention gates, which helps reduce the differences between the encoder and decoder. Furthermore, UNet3+[26] advanced the skip connection by adding full-scale skip connections and deep supervision, achieving better results in segmentation tasks. Xie et al.[27] designed a feature-steered network to learn the more distinctive features, which is built on a scale-adaptive module and cross path fusion (CPF) module. Wang et al.[28] used attention-based UNet to enhance the completeness of representation with the fusion of edge and body features. Katar et al.[29] introduced a mixed model that combines ConvNeXt blocks with self-attention methods to improve skin lesion segmentation. Despite the success of U-Net and its variants, CNN-based models are limited by their inability to capture long-range dependencies. This limitation arises from the inherent nature of convolution operations, which struggle to model global contextual information. Additionally, dense skip connections based on simple summation offer limited solutions for addressing the semantic gap, particularly in tasks where fine-grained detail and global context are crucial, such as skin lesion segmentation.

Transformers for medical image segmentation

Transformers, which are adept at capturing long-range dependencies, offer an effective alternative to CNNs. The Vision Transformer[30], the first application of Transformer to computer vision, partitions input images into a sequence of patches for embedding and encoding using Transformer blocks. This innovative approach inspired Transformer-based U-shaped models for medical image segmentation tasks. TransUnet[31] integrates transformer blocks with U-Net, leveraging their global feature modelling capabilities. In parallel, Swin-Unet[15] drew on the Swin Transformer to propose a fully Transformer-based method that applies Swin Transformer blocks to both the encoder and decoder. For 3D medical image segmentation, nnFormer[32] combines local and global attention for multi-organ segmentation, demonstrating impressive performance. However, transformer models like TransUnet have problems because they have a lot of parameters and are complicated to compute, and the Swin transformer's shifting window method can create rough edges in some situations. Furthermore, transformers inherently operate on single-scale outputs, limiting their ability to fully utilize multi-scale information, which hampers performance, especially in segmentation tasks requiring precise boundary delineation.

Deformable convolution network

The Deformable Convolutional Network (DCN)[33] is an extension of traditional convolutional networks designed to improve the adaptability of convolutional kernels to object deformations. DCNs introduce learned offsets to control the sampling positions of kernels on input feature maps. This dynamic adjustment allows the model to better capture the features of targets with varying shapes and positions. Compared to traditional convolution, deformable convolutions offer more flexibility in capturing features of objects with diverse shapes and scales, which is crucial in medical image segmentation. DCNs have proven effective in addressing the complexities of various segmentation tasks, where object deformations are a significant concern[34]. Xin et al.[35] used deformable convolution to build a feature extraction module, which enhances the modeling ability of the model for deformation. However, their method shows limited generalization on complex structures. On the other hand, deformable convolutions are critical in skin lesion segmentation due to challenges such as jagged, fuzzy, or occluded melanoma boundaries. Fixed-grid convolutions struggle to adapt, while deformable convolutions learn dynamic sampling offsets to align kernel coverage with skin lesion geometry.

Recently, Ma et al.[36] presented U-Mamba for general-purpose biomedical image segmentation, which integrates the advantages of local pattern recognition from CNNs and global context understanding from Mamba. Compared to CNNs, U-Mamba's SSM-based latent states are less interpretable than layer-wise CNN feature visualizations, and unlike Transformers, it lacks explicit attention maps for global context analysis, complicating debugging and trust in clinical settings. Li et al.[37] presented a dual-path network that uses two parallel CNNs for feature extraction from different modalities. It integrates CNNs and Transformers with a feature-level fusion strategy that uses the local attention aggregation block to focus on region-of-interest features and suppress invalid regions. However, it does not address the semantic gap between features. To achieve better accuracy in medical image segmentation, especially for skin lesions, ScaleFusionNet combines multi-scales of feature extraction with a hybrid of Transformer and CNN designs to improve how well it segments images. Traditional CNN-based models like U-Net struggle with fine-grained details and global context, prompting the need for improved architectures. ScaleFusionNet addresses these issues by using AFBs that incorporate swin transformers and deformable convolutions to collect features at various scales, which helps to refine the edges of lesions while preserving small details. Additionally, the CATM enhances encoder-decoder feature fusion, reducing semantic gaps through guided attention mechanisms. By combining Swin transformer blocks[21] for global context and adaptive multi-scale fusion for local detail refinement, ScaleFusionNet achieves superior segmentation accuracy, demonstrating strong generalization in the skin lesion segmentation task.

Methods

Figure 1 illustrates the architecture of ScaleFusionNet, which follows the U-Net design and consists of three primary components: an encoder, CATM, and AFB. The encoder employs a hybrid approach that integrates convolutional layers and Swin transformer blocks to effectively capture local and global features. By combining convolutional locality with self-attention mechanisms, the encoder enhances feature representation, ensuring robust extraction of hierarchical information. The CATM is introduced at skip connections to refine encoder-decoder feature fusion. It utilizes Swin transformer blocks and a CAF to mitigate the semantic gap and dynamically align hierarchical features. By using cross-attention mechanisms, the CATM enhances the integration of skip connection features with decoder information, ensuring improved contextual understanding. The AdaptiveFusionBlock further enhances multi-scale feature extraction and fusion by integrating deformable convolutions with Swin transformer-based attention. This fusion process refines the lesion boundaries and preserves fine-grained details, which are essential for accurate segmentation. Given an input image $I \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels, respectively, the encoder progressively extracts the multi-scale features. The CATM is positioned at skip connection points, receiving features X_{Skip} from the encoder and X_{Decoder} from a lower-level decoder. The extracted features are dynamically refined using cross-attention mechanisms, allowing for precise alignment and feature enhancement. The AFB processes these refined features by fusing multi-scale information through deformable convolutions and Swin transformer-based attentions. The final segmentation result is obtained after feature fusion and upsampling operations in the decoder, ensuring a high-resolution and well-defined segmentation mask.

CATM

In U-Net, skip connections serve to provide information supplementation. During encoding, continuous compression of feature maps leads to a significant loss of spatial detail. Using low-level semantic features from the encoder to supplement the decoder feature restoration is an effective strategy. However, a fundamental issue remains: the semantic gap between the encoder and decoder features. Only concatenating features at different semantic levels can result in performance degradation owing to this misalignment.

To address this, we introduced the CATM to refine the encoder-decoder feature fusion. Unlike conventional skip connections, the CATM employs Swin transformer blocks and cross-attention fusion to adaptively align hierarchical features. By using self-attention and cross-attention mechanisms, the CATM ensures the effective

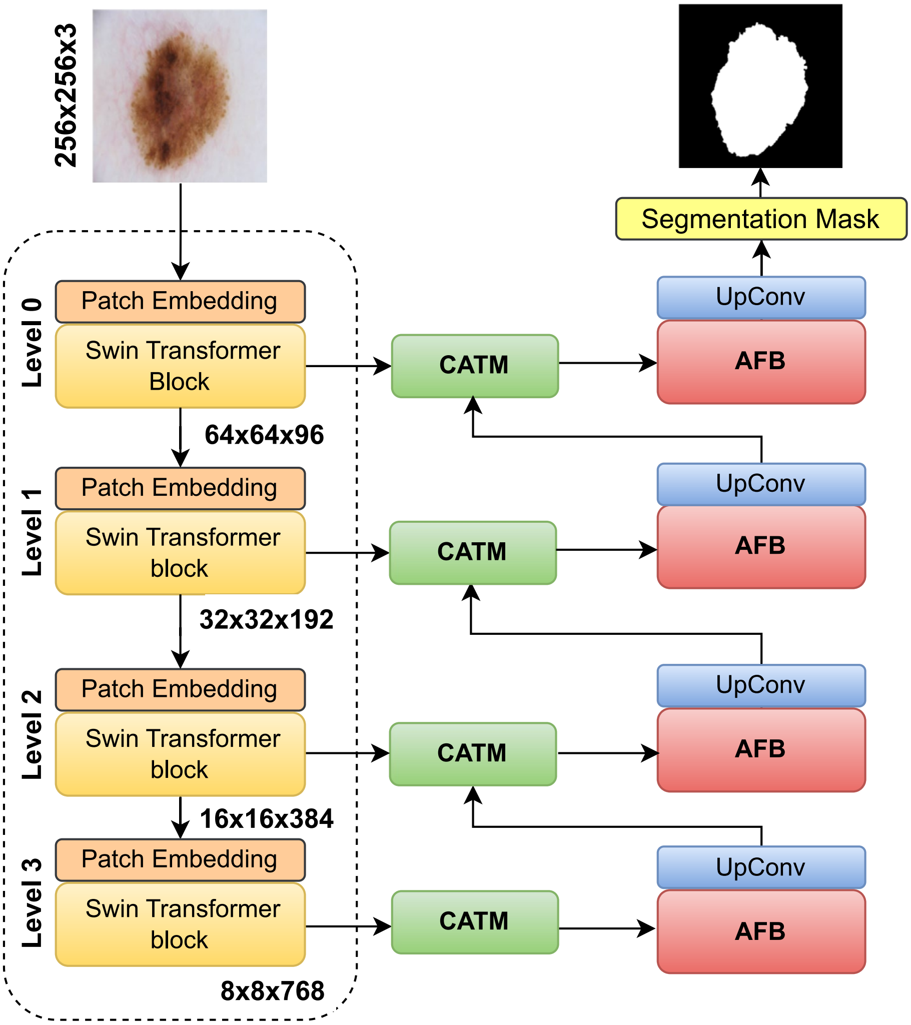

Fig. 1. Architecture of ScaleFusionNet with a U-Net design, consisting of an encoder, CATM, and AFB. The encoder, utilizing convolutional layers and Swin transformer blocks, extracts multi-scale features at resolutions $64 \times 64 \times 96$, $32 \times 32 \times 192$, $16 \times 16 \times 384$, and $8 \times 8 \times 768$. The CATM refines feature fusion at skip connections using cross-attention, while the AFB enhances multi-scale fusion with deformable convolutions and Swin transformer-based attention to preserve fine-grained details for accurate segmentation.

transfer of relevant spatial and contextual information across the network. As shown in Fig. 2, given the encoder features X_{Skip} and decoder features X_{Decoder}, the CATM dynamically refines the feature alignment using a learnable attention mechanism. This process is formulated as follows.

$$Q, K, V = \text{Swin Transformer Block}(X_{\text{Decoder}})$$

$$\text{V'} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V$$

$$X'_{\text{Skip}} = \text{CAF}(\text{V'}, X_{\text{Skip}})$$

here, the Swin transformer block extracts query (Q), key (K), and value (V) representations from the decoder features. Q represents the decoder's high-level semantic features X_{Decoder}. It "queries" the encoder's spatial details and guides the attention toward relevant regions in X_{Skip}. K and V are projections of X_{Decoder} through Swin Transformer blocks. K computes similarity scores with Q to weight the importance of encoder features. V aggregates contextual information from X_{Skip} based on attention weights. V' computes a weighted sum of values V based on the attention scores between Q and K, where d_k is the dimension of K. after that, CAF integrates these with the encoder features X_{Skip}, to reduce the semantic gap between encoder and decoder. After passing through the CAF, the features X'_{Skip} undergo a full-stage parameter-shared spatial attention mechanism to achieve unified feature attention. SharedSA computes a spatial attention map $A \in \mathbb{R}^{H \times W}$ shared across all stages on the given input feature map $X'_{\text{Skip}} \in \mathbb{R}^{H \times W \times C}$ from CAF:

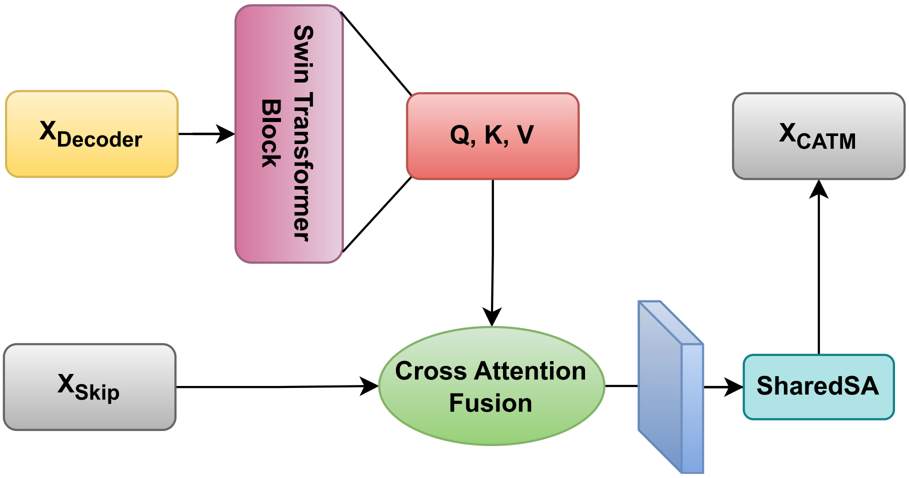

Fig. 2. Schematic diagram of CATM. The decoder features X_{Decoder} generate query, key, and value representations, which are fused with the encoder features X_{Skip} via CAF. The refined features are then processed with SharedSA to produce the final aligned features X_{CATM}, preserving both fine details and high-level semantics.

$$A = \sigma\left(\text{Conv}_{1\times1}\left([X'_{\text{Skip}} \oplus \text{AvgPool}(X'_{\text{Skip}}) \oplus \text{MaxPool}(X'_{\text{Skip}})]\right)\right)$$

where σ denotes the sigmoid activation function, while $\text{Conv}_{1\times1}$ represents a pointwise convolution. The operations AvgPool and MaxPool refer to global average and max pooling operations, respectively. Additionally, $\oplus$ indicates channel-wise concatenation, where the features are concatenated along the channel dimension. The refined output X_{CATM} is then obtained by:

$$X_{\text{CATM}} = X'_{\text{Skip}} \otimes A$$

here, $\otimes$ denotes element-wise multiplication that emphasizes spatially salient regions uniformly across all decoder levels. The refined features X_{CATM} preserve both fine-grained details and high-level contextual information.

```
Require: $X_{\text{Decoder}}, X_{\text{Skip}} \in \mathbb{R}^{H\times W\times C}$
Ensure: $X_{\text{CATM}} \in \mathbb{R}^{H\times W\times C}$
$Q, K, V = \text{SwinTransformerBlock}(X_{\text{Decoder}})$        ▷ Obtain Query, Key, and Value
$X'_{\text{Skip}} = \text{CAF}(X_{\text{Skip}}, Q, K, V)$        ▷ Feature Alignment
$X_{\text{CATM}} = \text{SharedSA}(X'_{\text{Skip}})$        ▷ Refined Feature Fusion
return $X_{\text{CATM}}$        ▷ Return
```

Algorithm 1. CATM for ScaleFusionNet

AFB

The decoder plays a critical role in feature decompression and mask generation for medical image segmentation. A key challenge is the accurate restoration of boundary details and enhancement of attention toward target regions. Conventional decoder designs, whether convolution- or transformer-based, often struggle to effectively capture fine-scale information, leading to imprecise lesion localization and segmentation. To address these limitations, we introduced the AFB, which integrates adaptive multi-scale feature fusion to refine segmentation. This block is built with Swin transformer-based attention and deformable convolution-based adaptive feature extraction as it learns dynamic sampling offsets to align kernel coverage with skin lesion geometry, allowing the model to capture both local and global contextual information through parallel pathways. Figure S1 shows the GradCAM heatmap of deformable convolutional.

As shown in Fig. 3, given the input feature (X) from the decoder, the AFB processes it through three parallel branches to extract complementary representations: Swin transformer, deformable convolution, and identity branches. In the Swin transformer branch, the model employs a resolution-aware adaptation strategy to balance efficiency and richness of features. The Tiny variant of Swin transformer is used to capture long-range dependencies, but its processing is dynamically adjusted based on the encoder level and spatial resolution. The Swin transformer branch employs a resolution-aware adaptation strategy to accommodate varying spatial resolutions during the decoding process. Specifically, for higher-resolution inputs (Level 0: 64 × 64), only

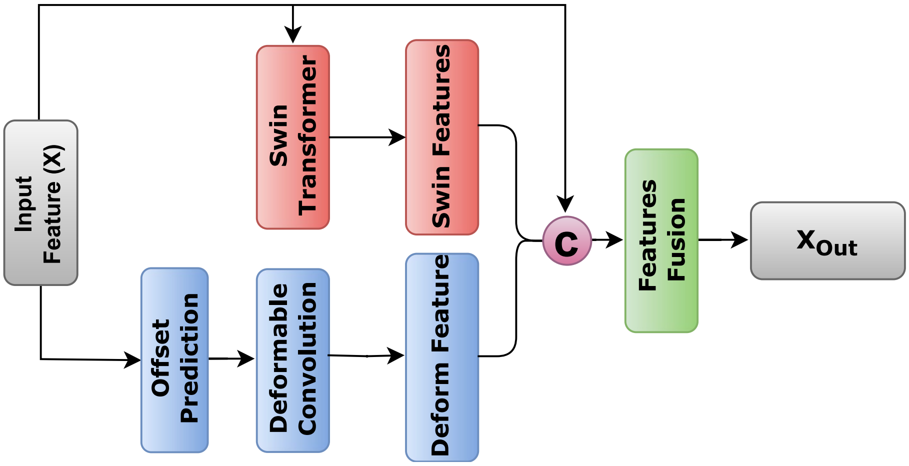

Fig. 3. Schematic diagram of AFB for multi-scale feature fusion in the decoder. The input feature X is processed through three parallel branches: the Swin transformer branch X'_{Swin}, deformable convolution branch X'_{Deform}, and identity branch. The outputs are concatenated X_{Combined} and passed through a 1x1 convolution to produce the final output X_{Out}, refining lesion boundaries and improving segmentation accuracy.

the first two Swin Transformer stages are utilized to preserve fine-grained, detailed features. At intermediate resolutions (Level 1: 32 × 32), the model processes the features through the first three Swin stages, providing a balance between the feature richness and computational cost. For deeper levels with lower resolutions (Level 2: 16 × 16), all four Swin Transformer stages are employed; however, the embedding dimensions are reduced to mitigate potential memory bottlenecks. At the deepest level (Level 3: 8 × 8), the Swin Transformer is omitted altogether, and a simple convolutional operation is applied instead because the spatial resolution at this stage is too limited to benefit from self-attention mechanisms. This adaptive approach ensures computational efficiency while preserving the benefits of hierarchical feature learning.

$$X'_{\text{swin}} = \begin{cases} \text{SwinTransformer}(X), & \text{if encoder\_level} < 3 \\ \text{Conv}(X), & \text{otherwise} \end{cases}$$

The deformable convolution branch performs spatially adaptive feature extraction with offset prediction as follows:

$$\text{Offset} = \text{Conv2D}_{(3\times3)}(X)$$

$$X'\text{Deform} = \text{DeformConv}(X, \text{Offset})$$

The identity branch preserves the original features X to maintain low-level information for the model stability. The outputs from all three branches are fused through channel-wise concatenation, followed by feature reduction:

$$X_{\text{Combined}} = \text{Concat}(X, X'_{\text{Swin}}, X'_{\text{Deform}})$$

$$X_{\text{Out}} = \text{Conv2D}_{(1\times1)}(X_{\text{Combined}})$$

This three-way fusion mechanism enhances the refinement of the lesion boundary while preserving fine-grained details through complementary feature representations. The deformable convolution adapts to irregular lesion shapes through learnable spatial offsets, the Swin transformer provides a global contextual understanding, and identity mapping maintains essential low-level features. By integrating AFB into ScaleFusionNet, we achieved improved segmentation performance by effectively combining multiple feature extraction strategies in parallel, leading to better lesion delineation and generalization across datasets.

```
Require: $X \in \mathbb{R}^{H \times W \times C}$
Ensure: $X_{\text{Out}} \in \mathbb{R}^{H \times W \times C}$
$X'_{\text{Swin}} = \text{SwinTransformer}(X)$    ▷ Extract features using Swin Transformer
$\text{Offset} = \text{Conv2D}_{3\times3}(X)$    ▷ Predict offset for deformable convolution
$X'_{\text{Deform}} = \text{DeformConv}(X, \text{Offset})$    ▷ Extract features using deformable convolution
$X_{\text{Combined}} = \text{Concat}(X, X'_{\text{Swin}}, X'_{\text{Deform}})$    ▷ Concatenate original and extracted features
$X_{\text{Out}} = \text{Conv2D}_{1\times1}(X_{\text{Combined}})$    ▷ Fuse features using 1x1 convolution
return $X_{\text{Out}}$    ▷ Return the fused output features
```

Algorithm 2. AFB for ScaleFusionNet

Experiments

Datasets

ISIC-2016: This dataset is derived from the skin lesion analysis for the melanoma detection challenge in 2016, comprising 1250 images meticulously annotated by professional experts with high-quality standard labels[38]. Among these, 900 images are designated as training data, and 350 images are allocated for validation.

ISIC-2018: The ISIC-2018 dataset, also collected by ISIC in 2018, consists of 2594 images and corresponding labels. The resolutions of the images ranged from 720×540 to 6708×4439 pixels[16]. Among these, 2594 images are randomly divided into training, validation, and test sets at a ratio of 8:1:1.

HAM10000: HAM10000[39] is the largest public skin lesion dataset with 10,015 dermoscopic images (600 × 450 pixels) covering seven pigmented lesion types including actinic keratosis, basal cell carcinoma, and melanoma. More than 50% of cases are histopathologically confirmed, with remaining cases validated by expert.

PH ^{2}**:** The dermoscopic images used in this study were acquired from Hospital Pedro Hispano in Matosinhos, Portugal, using the Tuebinger Mole Analyzer system set at 20x magnification. These images are formatted as 8-bit RGB color files with dimensions of $768{\times}560$ pixels. The dataset consists of 200 dermoscopic images featuring various melanocytic lesions[40].

Evaluation metrics

The main evaluation metrics used the Dice coefficient (DSC), Intersection over Union (IOU), Sensitivity (SE), Specificity (SP), and Accuracy (ACC). The Dice coefficient measures the overlap between the predicted segmentation mask and the ground truth mask. It is defined as:

$$DSC = \frac{2|A \cap B|}{|A| + |B|}$$

where A is the predicted mask and B is the ground truth mask. A higher Dice score indicates better segmentation performance. Intersection over Union evaluates segmentation accuracy by computing the ratio of the intersection to the union between the predicted and actual masks:

$$IOU = \frac{|A \cap B|}{|A \cup B|}$$

SE measures the proportion of actual positive samples correctly identified by the model.

$$\text{SE} = \frac{|A \cap B|}{|B|}$$

SP measures the proportion of actual negative samples that the model correctly identifies.

$$\text{SP} = \frac{|\overline{A} \cap \overline{B}|}{|\overline{B}|}$$

where $\overline{A}$ and $\overline{B}$ are compelments of A and B. Universal set U is the total number of pixels in the image. ACC measures the proportion of correct predictions made by a model out of total predictions.

$$\text{ACC} = \frac{|A \cap B| + |\overline{A} \cap \overline{B}|}{|U|}$$

Implementation details

All experiments in this paper were conducted using the PyTorch 1.12.0 framework. The experiments were performed on a computer equipped with an Ubuntu 18.04 operating system, Intel Core i9-13900K CPU, Nvidia

RTX 4060 GPU, and 1TB solid-state drive. For all experiments involving ScaleFusionNet, the AdamW optimizer was utilized with a learning rate and weight decay set to 1e−4. We used a combination of BCE and IOU losses to form our loss function, along with random rotation and random flipping for data augmentation. For comparison, we referenced the experimental results disclosed in relevant papers for similar methods. For outstanding models that did not perform skin lesion segmentation tasks, we retrained them using publicly available official execution codes. To ensure fairness, we kept parameters that did not affect the model learning capacity, such as epochs and batch size, consistent with ScaleFusionNet, setting epochs to 200 and batch size to 8. The input size for the network was 256 × 256. The experiments focused on the ISIC-2016 and ISIC-2018 datasets,and external testing was conducted on the PH ^{2} dataset based on the trained weights.

This experimental setup ensured a robust and fair evaluation of ScaleFusionNet's performance by utilizing state-of-the-art hardware and software configurations to achieve accurate and reproducible results. The use of a combined loss function and data augmentation techniques further enhances the model's ability to generalize and perform well on diverse skin lesion segmentation tasks.

Experimental results

Results on ISIC-2016

We selected 13 prominent models for comparison with the proposed ScaleFusionNet. In all 13 models, we also included SAM2-UNet[41] and U-Mamba[36] for skin lesion segmentation. The SAM2-UNet is an emerging vision foundation model that continuously achieves good performance on various tasks. UMamba is a general-purpose network inspired by State Space Sequence Models (SSMs), a new family of deep sequence models known for their strong capability in handling long sequences. Table 1 shows that ScaleFusionNet achieved a DSC score of 92.94% and an IOU score of 87.35% on the ISIC-2016 dataset, which are the average results of five-fold experiments, demonstrating outstanding performance. Compared with the Swin-Unet model, ScaleFusionNet improved by 2.82% in the DSC metric and 4.14% in the IOU metric. Compared with MISSFormer, ScaleFusionNet exhibited enhancements of 2.48% and 3.43% in the DSC and IOU metrics, respectively. Against the D-LKA model, ScaleFusionNet still showed improvements of 0.11% and 0.20% in the DSC and IOU metrics, respectively. This indicates that ScaleFusionNet is more accurate than D-LKA in detecting the refined boundary of skin lesions. Although the performance gains may not be significant, from the perspective of parameter count and computational complexity, ScaleFusionNet reduces the number of parameters by 37.7% and decreases the computational load by 22.5% compared with D-LKA. In terms of memory usage, ScaleFusionNet reduces memory consumption by 25% compared to D-LKA. This indicates that ScaleFusionNet consumes fewer hardware resources than D-LKA while maintaining the model size and computational complexity. SAM2Unet and U-Mamba are also behind ScaleFusionNet in terms of DSC and IOU. Compared to other methods such as U-Net, although ScaleFusionNet employs a more complex architecture to address issues in U-shaped medical image segmentation models, we find this approach justified given the 5.13% performance improvement and reduced computational load. In clinical applications, a faster and lighter model can support a broader range of compatible use cases, which is crucial for hospitals and organizations with limited computational resources.

To enhance the assessment of model performance, we selected 10 high-performing models for qualitative scrutiny of the experimental outcomes, elucidating the differences among them. The red areas in the illustrations represent ground-truth labels meticulously annotated by experts, reflecting the diagnostic preferences of clinical doctors in real-world scenarios. A larger red area indicates lower model accuracy and poorer discrimination of the affected regions. In contrast, the green areas represent the predicted labels obtained during the model-testing phase. A larger green area suggests that the model has mistakenly segmented healthy skin, which could mislead doctors into treating non-affected areas, especially with destructive procedures, such as lasers or cryotherapy. The yellow areas indicate the overlap between the predicted and ground truth labels; a larger yellow area indicates a more accurate identification of the lesion region by the model. In summary, from the perspective of clinical

Methods	Params(M)	FLOPs(G)	GPU Mem(GB)	DSC	IOU
U-Net[2]	34.53	124	4.1	87.81 ± 0.41	80.25 ± 0.50
Att-Unet[25]	34.88	126.1	4.3	87.43 ± 0.47	79.70 ± 0.62
nnU-Net[42]	–	–	–	90.45 ± 0.35	84.52 ± 0.53
SwinUNet[15]	27.17	6.16	3.8	90.12 ± 0.39	83.21 ± 0.57
MISSFormer[20]	42.46	9.89	5.2	90.46 ± 0.33	83.92 ± 0.45
DAEFormer[14]	48.07	27.89	5.9	91.19 ± 0.31	85.40 ± 0.43
HiFormer[43]	25.51	8.05	3.9	91.48 ± 0.28	85.15 ± 0.40
TransFuse[44]	26.25	8.82	4.0	92.03 ± 0.26	86.19 ± 0.38
D-LKA[45]	101.64	19.92	7.1	92.83 ± 0.23	87.15 ± 0.34
SU-Net[46]	20.9	4.58	3.2	92.33 ± 0.24	86.58 ± 0.36
U-Mamba[36]	16.40	3.51	2.9	91.77 ± 0.27	86.16 ± 0.42
SAM2-UNet[41]	21.09	5.65	3.5	91.52 ± 0.29	85.88 ± 0.44
MSCA-Net[47]	27.09	12.88	4.1	91.35 ± 0.32	85.59 ± 0.48
ScaleFusionNet (Ours)	62.91	15.45	5.3	**92.94 ± 0.21**	**87.35 ± 0.30**

Table 1. Performance comparison on ISIC-2016 dataset. Singificance values are in bold.

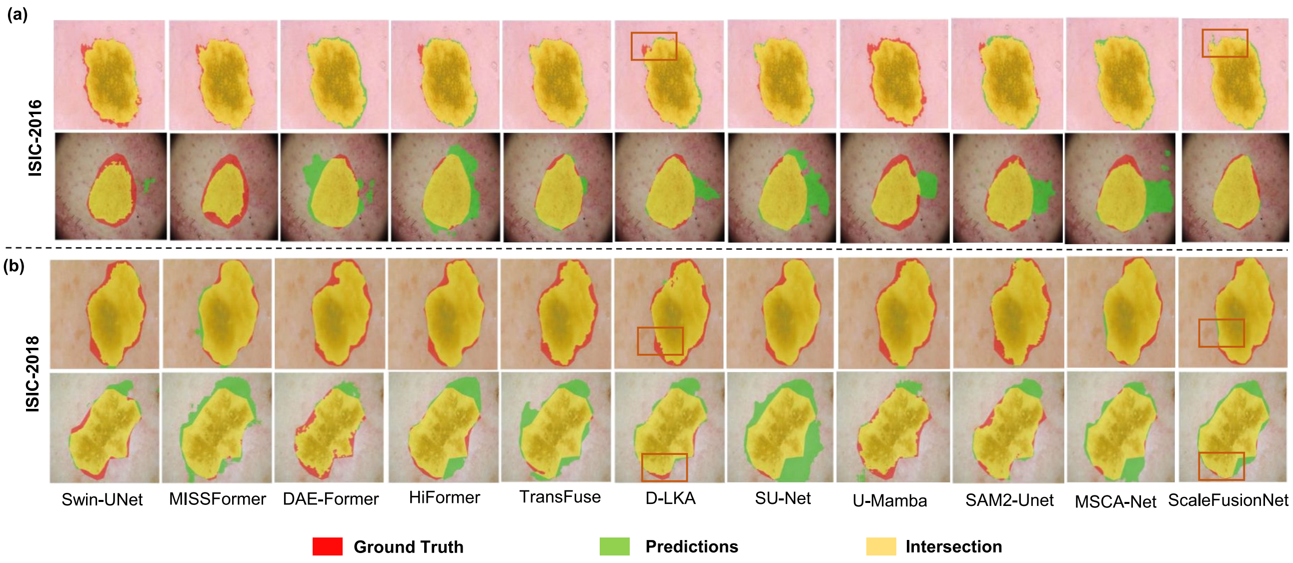

Fig. 4. Qualitative comparison of 10 skin lesion segmentation models on ISIC-2016 and ISIC-2018 datasets. Yellow denotes correct prediction, red indicates missed regions (ground truth only), and green highlights false positives(predicted healthy skin as lesion). ScaleFusionNet shows superior overlap with minimal errors. The red square regions illustrate edge refinement capability among the models.

Methods	DSC	IOU	SE	SP	ACC
U-Net[2]	85.45 ± 0.45	77.33 ± 0.58	88.00 ± 0.62	96.97 ± 0.35	94.04 ± 0.42
Att-Unet[25]	85.66 ± 0.48	77.64 ± 0.61	86.74 ± 0.67	98.63 ± 0.28	93.76 ± 0.44
nnU-Net[42]	89.03 ± 0.38	82.02 ± 0.49	91.02 ± 0.54	97.55 ± 0.32	96.40 ± 0.37
Swin-Unet[15]	89.31 ± 0.40	82.14 ± 0.51	90.99 ± 0.56	97.20 ± 0.33	95.99 ± 0.39
MISSFormer[20]	89.44 ± 0.37	82.41 ± 0.47	90.79 ± 0.53	96.92 ± 0.34	96.04 ± 0.38
DAEFormer[14]	89.89 ± 0.35	83.21 ± 0.44	90.52 ± 0.50	97.33 ± 0.31	96.13 ± 0.36
HiFormer[43]	90.55 ± 0.33	83.81 ± 0.42	92.02 ± 0.48	96.43 ± 0.32	96.55 ± 0.34
TransFuse[44]	91.08 ± 0.31	84.65 ± 0.39	91.39 ± 0.46	97.80 ± 0.29	96.66 ± 0.32
D-LKA[45]	91.64 ± 0.28	85.64 ± 0.36	**91.94 ± 0.43**	**98.20 ± 0.26**	96.89 ± 0.30
SU-Net[46]	90.90 ± 0.32	84.49 ± 0.40	90.76 ± 0.47	97.64 ± 0.30	96.66 ± 0.33
TranSiam[20]	90.44 ± 0.37	83.45 ± 0.49	90.83 ± 0.55	96.92 ± 0.34	95.04 ± 0.39
U-Mamba[36]	89.74 ± 0.36	83.16 ± 0.45	90.83 ± 0.52	97.33 ± 0.31	97.75 ± 0.35
SAM2-UNet[41]	89.52 ± 0.38	83.07 ± 0.46	90.75 ± 0.54	98.13 ± 0.28	97.54 ± 0.37
MSCA-Net[47]	89.31 ± 0.39	83.28 ± 0.45	90.37 ± 0.55	97.53 ± 0.31	97.24 ± 0.38
ScaleFusionNet (Ours)	**91.80 ± 0.26**	**85.57** ± 0.34	90.88 ± 0.41	97.67 ± 0.25	**98.24 ± 0.28**

Table 2. Performance comparison on ISIC-2018 dataset. Values are presented as mean ± standard deviation. Significance values are in bold.

diagnosis and treatment, smaller green and red areas and a larger yellow area indicate better model performance, enabling more effective segmentation of skin lesions to assist in diagnosis and treatment decisions. The results of the qualitative analysis of the 10 models in the ISIC-2016 dataset are shown in Fig. 4a. From the first two rows, it is evident that ScaleFusionNet exhibits a broader yellow region than the D-LKA model, indicating that ScaleFusionNet is better at identifying affected areas. Furthermore, ScaleFusionNet's predictions show fewer green and red regions, which reduces the likelihood of misdiagnosis and missed diagnoses in the clinical setting. This difference was even more pronounced in the latter two rows. Although TransFuse had the largest yellow region, it also had the largest green area, indicating a misdiagnosis of healthy regions. Although it covers areas of injury, this misdiagnosis can have serious implications for clinical diagnosis. In comparison with D-LKA, ScaleFusionNet has a similarly sized yellow region but with a smaller misdiagnosis area, aligning better with clinical diagnostic needs.

These results highlight the superior ability of ScaleFusionNet to accurately segment skin lesions while minimizing errors, making it a highly effective tool for clinical applications. Its combination of high accuracy and strong generalization ability positions it as a leading solution for skin lesion segmentation tasks. The model's ability to preserve fine-grained details and refine the limits of the injury further underscores its potential to improve melanoma diagnosis and treatment in real-world clinical settings.

Results on ISIC-2018

In the experiments conducted on the ISIC-2018 dataset, 14 mainstream medical image segmentation models were selected for comparison with ScaleFusionNet, and more relevant parameters were disclosed. The quantitative analysis results of the ISIC-2018 comparison experiments are listed in Table 2. ScaleFusionNet continued to exhibit competitive results compared to the other 14 medical image segmentation methods, achieving the best results in terms of the DSC metric on ISIC-2018, with most other metrics ranking in the top two. Compared with TransFuse, ScaleFusionNet showed a 0.72% improvement in the DSC metric while maintaining a consistent IOU level of 0.92%. Compared with D-LKA, ScaleFusionNet exhibited a 0.06% improvement in the DSC metric, although it slightly lagged behind D-LKA in the IOU metric. Other metrics, namely, Sensitivity, Specificity, and Accuracy, are also presented in Table 2, where we can see the performance of ScaleFusionNet compared to other models. Finally, ScaleFusionNet yielded competitive results with other methods, demonstrating strong performance in skin lesion segmentation. Figure S3 have shown the ROC curve of model performance. The curve demonstrates outstanding discriminative capability, with an AUC of 0.9985, indicating near-perfect separation between lesion and non-lesion pixels across varying thresholds. This aligns with the high sensitivity 90.88% and specificity 97.67% reported in Table 2, confirming that the model consistently achieves high true positive rates while maintaining a very low false positive rate. The curve's proximity to the top-left corner further reflects the robustness and reliability of ScaleFusionNet in medical image segmentation tasks. The Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD) were also calculated to further assess the model's performance. The proposed ScaleFusionNet achieved a mean HD of 8.1783 and a mean ASSD of 0.1093 on the ISIC-2018 test set. The low ASSD value indicates that the predicted lesion boundaries are, on average, highly consistent with the ground truth, while the moderate HD reflects minimal occurrences of outlier boundary deviations. These results are consistent with the high DSC, IOU, sensitivity, and specificity reported in Table 2.

The qualitative analysis results of ISIC-2018 are shown in Fig. 4b. Visual inspection of the results in the last two rows reveals that, compared with D-LKA, ScaleFusionNet has a larger yellow region and a smaller red region, indicating a higher prediction accuracy, even though it performs slightly worse on the IOU metric. In comparison to TransFuse, while the yellow regions were nearly identical in size, ScaleFusionNet exhibited a smaller range of green areas, demonstrating its superior ability to identify skin lesion regions and reduce the likelihood of misdiagnosis. Similarly, the results in the last two rows further highlight the overall superior performance of the ScaleFusionNet. Based on the experiments using the ISIC-2018 dataset, the D-LKA model and ScaleFusionNet achieved the first and second best performances in terms of the DSC and IOU metrics, respectively. In terms of overall performance, ScaleFusionNet demonstrated better accuracy and boundary fitting. The qualitative analysis of the results, as illustrated in Fig. 4b, shows that ScaleFusionNet has a smaller green region than D-LKA, indicating a reduced likelihood of misdiagnosis. This further underscores the ability of ScaleFusionNet to accurately segment skin lesions while minimizing errors. These results highlight the superior performance of ScaleFusionNet on the ISIC-2018 dataset, achieving high accuracy in lesion segmentation while maintaining efficiency in terms of parameter count and computational complexity. Its ability to reduce misdiagnosis and improve lesion boundary delineation makes it a highly effective tool for skin lesion analysis, particularly in clinical applications in which precision and efficiency are critical.

Results on HAM10000

To evaluate our model's performance on large-scale datasets, we selected 14 high-performing models for comparison using the HAM10000 dataset. These 14 models had not reported their performance on HAM10000 in their original papers, so we retrained them using their publicly available implementations. To ensure fair

Methods	DSC	IOU	SE	SP	ACC
U-Net	92.95	88.04	93.87	96.97	96.04
Att-Unet	91.59	87.64	93.74	97.63	95.76
nnU-Net	92.03	88.12	94.02	97.95	96.65
Swin-Unet	93.31	88.14	93.99	97.20	95.97
MISSFormer	93.44	89.41	94.79	96.92	96.14
DAEFormer	93.89	89.21	95.52	97.33	96.83
HiFormer	94.55	89.81	95.02	97.43	96.91
TransFuse	94.38	90.65	94.39	98.24	97.16
D-LKA	95.11	91.14	95.44	**98.20**	97.17
SU-Net	94.70	90.49	94.76	97.94	96.66
TranSiam	92.44	87.45	93.83	96.92	95.04
U-Mamba	92.74	88.16	92.83	97.33	95.75
SAM2-UNet	91.52	88.07	93.75	98.13	95.54
ScaleFusionNet (Ours)	**95.37**	**91.74**	**95.71**	98.01	**97.26**

Table 3. Performance domparison on HAM10000 dataset. Singificance values are in bold.

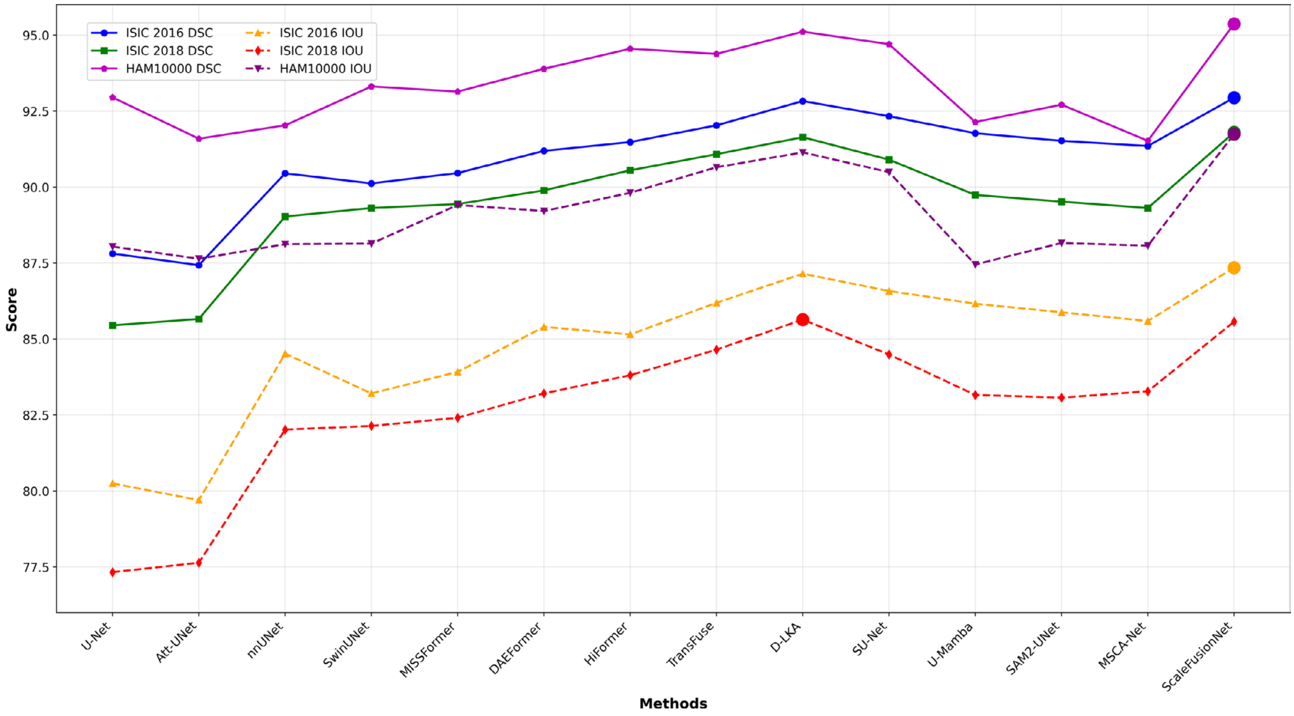

Fig. 5. Performance comparison of segmentation models on the ISIC-2016, ISIC-2018, and HAM10000 datasets. The DSC (solid lines with markers) and IOU (dashed lines with markers) scores are shown for various models. ScaleFusionNet outperforms the other models across all datasets, achieving the highest accuracy in both DSC and IOU.

Methods	PH^2 (test)				
	DSC	SE	SP	ACC	IOU
Swin-UNet	90.92 ± 0.45	96.97 ± 0.32	91.42 ± 0.58	93.39 ± 0.41	84.09 ± 0.67
MISSFormer	91.9 ± 0.38	97.1 ± 0.29	92.82 ± 0.45	94.24 ± 0.33	85.49 ± 0.54
DAEFormer	90.28 ± 0.52	97.41 ± 0.26	90.02 ± 0.61	92.99 ± 0.48	83.37 ± 0.73
HiFormer	92.03 ± 0.41	96.6 ± 0.35	93.46 ± 0.39	94.45 ± 0.37	85.88 ± 0.58
D-LKA	92.17 ± 0.36	97.3 ± 0.28	93.53 ± 0.42	94.52 ± 0.34	86.14 ± 0.51
SUnet	92.32 ± 0.33	98.14 ± 0.21	92.19 ± 0.47	**94.86 ± 0.31**	86.23 ± 0.49
U-Mamba	92.1 ± 0.39	97.7 ± 0.25	93.34 ± 0.43	94.51 ± 0.35	85.66 ± 0.55
SAM2-UNet	91.83 ± 0.44	98.02 ± 0.22	92.46 ± 0.46	94.52 ± 0.32	85.48 ± 0.62
MSCA-Net	91.76 ± 0.47	96.97 ± 0.34	**93.79 ± 0.38**	94.27 ± 0.29	85.28 ± 0.66
ScaleFusionNet(Ours)	**92.37 ± 0.31**	**98.23 ± 0.19**	92.44 ± 0.45	94.73 ± 0.28	**87.10 ± 0.44**

Table 4. Performance comparison on PH^2 dataset. Values are presented as mean ± standard deviation. Significance values are in bold.

comparison, we only adjusted the epoch and batch size parameters while keeping all other configurations identical to the official implementations. To further validate the generalization ability of each model, we conducted external testing using the PH2 dataset. We treated PH2 as an external test set and evaluated each model using weights trained on the HAM10000 dataset. The results are presented in Table 3.

As shown in Table 3, when the dataset size reaches a certain scale, the performance gap between different models significantly narrows. This highlights the importance of data volume in model performance. Compared to U-Net, ScaleFusionNet achieved a 2.4% improvement in the DSC metric on the HAM10000 dataset. Additionally, ScaleFusionNet outperformed the second-best model, D-LKA, in most metrics.

Figure 5 presents the performance of various segmentation models on the ISIC-2016, ISIC-2018, and HAM10000 datasets using both DSC and IOU metrics. ScaleFusionNet consistently outperforms the other models across all datasets. On the ISIC-2016 dataset, it achieves the highest DSC score, as highlighted by the blue line, while also leading in IOU performance (yellow dashed line). For ISIC-2018, ScaleFusionNet, represented by the green line (DSC) and red dashed line (IOU), again ranks among the top performers, closely followed by D-LKA and TransFuse. On the HAM10000 dataset, represented by magenta (DSC) and purple (IOU),

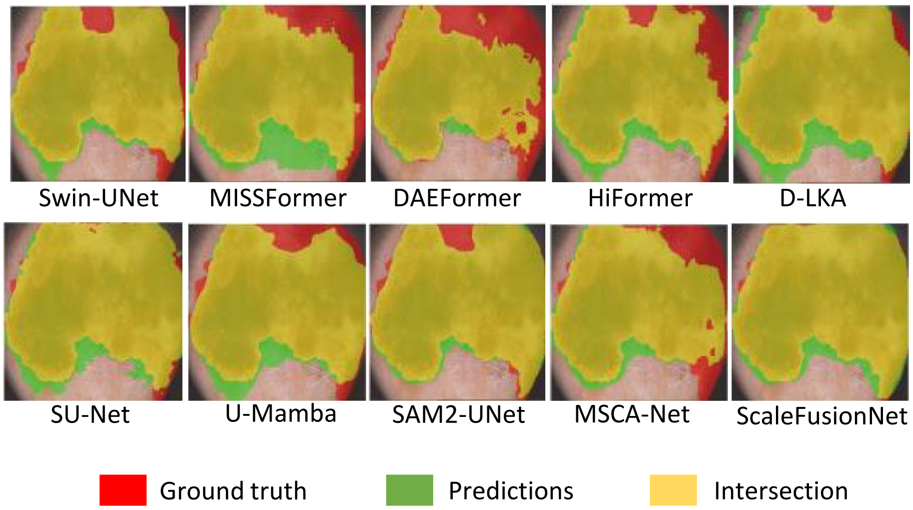

Fig. 6. Visual comparison of 10 models on the PH^2 dataset using ISIC-2018 pretrained weights. Red shows ground truth, green shows incorrect predictions, and yellow shows accurate segmentation. ScaleFusionNet and SU-Net perform best, with ScaleFusionNet showing better boundary accuracy and fewer false positives.

Methods	Swin transformer block	CATM	AFB	SharedSA	DSC↑
0	✓				91.76
1	✓		✓		92.01
2	✓	✓		✓	92.24
3	✓	✓	✓		92.48
ScaleFusionNet	✓	✓	✓	✓	92.94

Table 5. Structural ablation results of ScaleFusionNet on ISIC-2016. The Swin Transformer Block, CATM, AFB, and SharedSA were tested individually and in combination.

ScaleFusionNet maintains its superior performance, achieving the best results among all compared methods. Overall, the figure demonstrates that ScaleFusionNet achieves consistently high DSC and IOU values across multiple datasets, confirming its robustness and generalization ability in skin lesion segmentation.

External validation with ISIC-2018

Independent validation experiments were conducted on the PH^2 dataset using the model weights pretrained on ISIC-2018. The results show that ScaleFusionNet demonstrates significant performance improvements compared to the SU-UNet model, with a DSC increase of 0.05% and an IOU increase of 0.87%, indicating an improvement over the experiments on the ISIC-2018 dataset. From Table 4, it can be observed that ScalefusionNet continues to exhibit competitive results compared to the other 9 medical image segmentation methods, achieving the best results in terms of the DSC metric on the PH^2 dataset, with most other metrics ranking in the top two. Based on the external validation experiments using the ISIC2018 dataset, the SUnet model and ScaleFusionNet achieved the first and second best performances in terms of the DSC and IOU metrics, respectively. The qualitative analysis of the external validation experiments is shown in Fig. 6. Visual inspection of the results from the external validation experiments using the ISIC-2018 trained weights on the PH^2 dataset reveals that the yellow region of the SUnet model was larger, indicating a better ability to accurately predict lesions. In contrast, compared with ScaleFusionNet, the green region in the bottom-left corner is also larger for SUnet, suggesting that ScaleFusionNet performs better in terms of accuracy and fitting to boundaries. Overall, both SUnet and ScaleFusionNet outperformed the other selected comparison models in terms of visualized results based on ISIC-2018 trained weights on the PH^2 dataset. This conclusion demonstrates the excellent performance of the ScaleFusionNet skin lesion segmentation method proposed in this paper.

Ablation study

To validate the effectiveness of the proposed ScaleFusionNet, a structural ablation study was conducted on the ISIC-2016 dataset, with the DSC used as the primary evaluation metric. The data from the structural ablation studies are listed in Table 5. Method 0 represents the segmentation results obtained using only the hybrid architecture and Method 1 represents the experimental results with the hybrid architecture and AFB. Method 2 represents the experimental results obtained using the hybrid architecture and the CATM with SharedSA, and

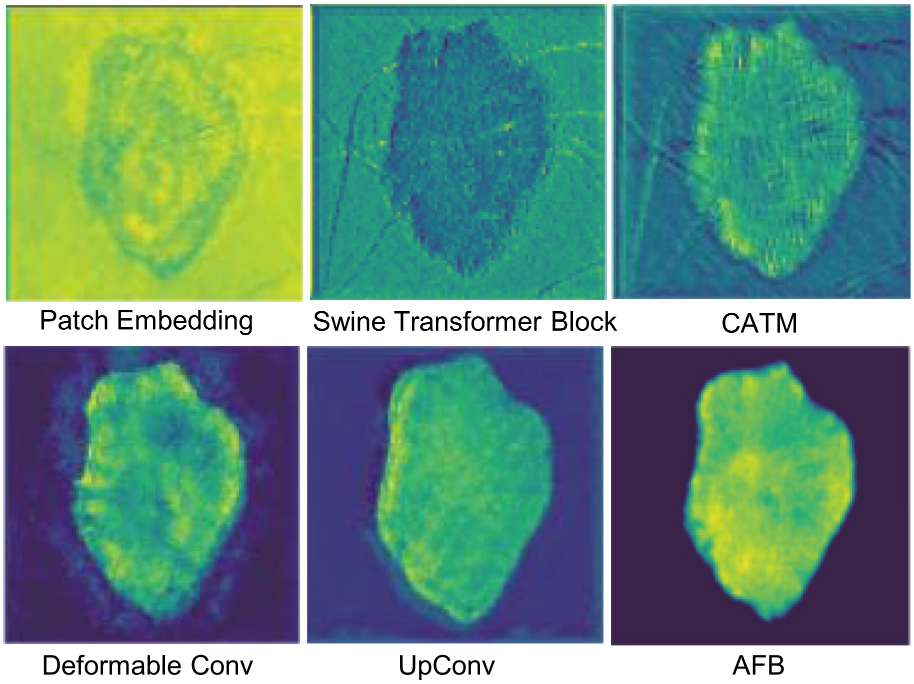

Fig. 7. Feature map visualization of different layers in ScaleFusionNet. Each module, including Patch Embedding, Swin Transformer Block, CATM, Deformable Convolution, UpConv, and AFB captures distinct features of the lesion.

Configuration	DSC (%)	IOU (%)
Q = Decoder, K/V = Skip	91.82	85.71
Q/K/V=Decoder	92.94	87.35
Q = Skip, K/V = Decoder	91.95	86.33
Q/K = Decoder, V = Skip	92.12	86.59

Table 6. Ablation study on the ISIC-2016 dataset to evaluate the impact of different Q/K/V configurations.

Method 3 represents the experimental results obtained using the hybrid architecture, AFB, and CATM without SharedSA.

From Table 5, it can be observed that using only the hybrid architecture results in a DSC of 91.76% on ISIC-2016, which is still superior to most models compared in Table 1. Method 1 consists of AFB to enhance multi scale fusion in the hybrid architecture, achieves a DSC of 92.01% on ISIC-2016. Method 2, which incorporates the CATM to enhance the skip connection features on top of the hybrid architecture, achieves a DSC of 92.24% on ISIC-2016. This performance surpasses that of the D-LKA method at 91.64%. Method 3, which introduces the AFB on top of the hybrid architecture and CATM without SharedSA, achieves a DSC performance of 92.48%. Finally, ScaleFusionNet, which combines the hybrid architecture, CATM, and AFB, achieves a DSC of 92.94% on ISIC-2016. The excellent performance of ScaleFusionNet is propelled by integrating these three proposed improvement methods. Furthermore, we visualize the features of Stage 0 and the corresponding modules within the same layer to intuitively observe the role of each structure. As shown in Fig. 7, it's clear that the focus on the target area becomes much stronger after it goes through the CATM, following the output of the hybrid architecture. Additionally, each feature map of the four multi-scale branches highlights different attention areas. This observation underscores the emphasis on micro-scale multi-resolution in this study. The micro-scale multi-resolution enables the features passed through the AFB to focus highly on the target area and exhibit excellent fitting to the target boundaries. Figure S2 presents the GradCAM analysis of the last layer for skin lesion segmentation.

These ablation experiments show how important the CATM and AFB are for ScaleFusionNet's performance, and they also highlight the need to improve the encoder's design. The results highlight the model's ability to achieve high accuracy in skin lesion segmentation while maintaining an efficient and balanced architecture.

To validate the design of our CAF mechanism, we systematically evaluated different configurations of Q, K, and V on the ISIC-2016 dataset, as presented in Table 6. The baseline configuration (Q/K/V = Decoder) achieved optimal performance (92.94% DSC, 87.35% IoU), demonstrating that using decoder features to guide all attention components best aligns encoder-decoder semantics. Alternative configurations revealed critical insights: using encoder skip connections for K/V (Q = Decoder, K/V = Skip) caused semantic misalignment (91.82% DSC), while reversing the roles (Q = Skip, K/V = Decoder) preserved semantics but lost spatial details (91.95% DSC). A hybrid approach (Q/K = Decoder, V = Skip) showed partial improvement (92.12% DSC) but underperformed the baseline, confirming that decoder-derived features dominate both attention weighting (K) and value aggregation (V) to mitigate the semantic gap. These results justify our CAF design, where decoder features comprehensively guide the attention process while selectively incorporating spatial details from encoder skip connections.

Conclusion and future work

This study presents a medical image segmentation model called ScaleFusionNet, which incorporates the CATM and AFB to learn and extract complex features from medical images. Experiments conducted on publicly available datasets demonstrate that ScaleFusionNet achieved competitive results in skin lesion segmentation. These innovative approaches positively impact diagnostic accuracy, guide treatment decisions, and promote further research in the field. However, its computational complexity is higher compared to some methods. This problem mainly comes from using multi-scale feature extraction and cross-attention mechanisms, where the input from different areas differs. One way to solve this is by better feature allocation, like self-selective routing. Also, deformable convolutions might worsen the issue, so simpler methods should be explored in future work. From a clinical perspective, a reliable medical image segmentation method must not only provide high-quality segmentation results but also deliver corresponding uncertainty metrics. ScaleFusionNet is an important advancement in skin lesion segmentation, as it uses adaptive multi-scale fusion and cross-attention mechanism to achieve accurate and strong results. However, future work should address computational efficiency and incorporate uncertainty quantification to further enhance its clinical applicability and reliability.

Data availibility

No datasets were generated or analysed during the current study.

Code availability

The implementation code can be found by clicking on this link.

Received: 16 April 2025; Accepted: 22 August 2025
Published online: 02 October 2025

References

1. Storelvmo, T. et al. Assessing the robustness and implications of econometric estimates of climate sensitivity. *Environ. Res. Lett.* **20**(2), 024055 (2025).
2. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In: . (eds. Navab, N., Hornegger, J., Wells, W.M. & Frangi, A.F.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* vol. 9351, pp. 234–241 (Springer, Cham, 2015). Lecture Notes in Computer Science
3. Cui, H., Wang, Y., Zheng, F., Li, Y., Zhang, Y. & Xia, Y. P2TC: A Lightweight Pyramid Pooling Transformer-CNN Network for Accurate 3D Whole Heart Segmentation. *J. Biomed. Health Inform.* (2025). Accessed 26 Feb 2025
4. Singh, M.K., Saini, I. & Sood, N. Less Complex U-Net (LCU-Net): A segmentation module in artificial intelligence-based prediction model of breast cancer. In: *Artificial Intelligence Techniques for Sustainable Development* 417–437 CRC Press
5. Li, M., Liu, W., Kang, Z. & Huang, X. Single image dehazing via multi-scale large kernel convolutional neural networks. In: *2024 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, Yokohama, Japan, 2024)
6. Hu, M. et al. LAMFFNet: Lightweight adaptive multi-layer feature fusion network for medical image segmentation. *Biomed. Signal Process. Control* **103**, 107456 (2025).
7. Qamar, S., Ahmad, P. & Shen, L. Dense encoder-decoder-based architecture for skin lesion segmentation. *Cogn. Comput.* **13**(2), 583–594 (2021).
8. Hu, B., Ye, Z., Wei, Z., Snezhko, E., Kovalev, V. & Ye, M. MLDA-Net: Multi-level deep aggregation network for 3D nuclei instance segmentation. *IEEE J. Biomed. Health Inform.* (2025).
9. Tang, F., Ding, J., Quan, Q., Wang, L., Ning, C. & Zhou, S.K. Cmunext: An efficient medical image segmentation network based on large kernel and skip fusion. In: *International Symposium on Biomedical Imaging (ISBI)* 1–5 (IEEE, Athens, Greece, 2024)
10. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T. & Xie, S. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 11976–11986 (2022)
11. Han, Z., Jian, M. & Wang, G.-G. ConvUNeXt: An efficient convolution neural network for medical image segmentation. *Knowl.-Based Syst.* **253**, 109512 (2022) (**Publisher: Elsevier**).
12. Thirunavukarasu, R. & Kotei, E. A comprehensive review on transformer network for natural and medical image analysis. *Comput. Sci. Rev.* **53**, 100648 (2024) (**Publisher: Elsevier**).
13. Cai, L., Hou, K. & Zhou, S. Intelligent skin lesion segmentation using deformable attention Transformer U-Net with bidirectional attention mechanism in skin cancer images. *Skin Res. Technol.* **30**(8), 13783 (2024).
14. Zhang, M. et al. Dual-attention transformer-based hybrid network for multi-modal medical image segmentation. *Sci. Rep.* **14**(1), 25704 (2024) (**Nature Publishing Group**).
15. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: (eds. Karlinsky, L., Michaeli, T. & Nishino, K. eds.) *ECCV 2022 Workshops* vol. 13803, pp. 205–218 (Springer, Cham, 2023). Lecture Notes in Computer Science
16. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H. & Halpern, A. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv (2019)
17. Broedermann, T., Sakaridis, C., Dai, D. & Van Gool, L. HRFuser: A multi-resolution sensor fusion architecture for 2D object detection. In: *International Conference on Intelligent Transportation Systems (ITSC)* 4159–4166 (IEEE, Bilbao, Spain, 2023)
18. Wang, J. et al. Xbound-former: Toward cross-scale boundary modeling in transformers. *IEEE Trans. Med. Imaging* **42**(6), 1735–1745 (2023) (**IEEE**).

19. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: (eds. Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R.S., Bradley, A., Papa, J.P., Belagiannis, V., Nascimento, J.C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H. & Madabhushi, A.) *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* vol. 11045, pp. 3–11 (Springer, Cham, 2018). Lecture Notes in Computer Science
20. Huang, X., Deng, Z., Li, D. & Yuan, X. MISSFormer: An Effective Medical Image Segmentation Transformer. arXiv (2021)
21. Islam, M. R., Qaraqe, M. & Serpedin, E. CoST-UNet: Convolution and swin transformer based deep learning architecture for cardiac segmentation. *Biomed. Signal Process. Control* **96**, 106633 (2024) (**Elsevier**).
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4* 3–11 (Springer, 2018)
23. Li, X. et al. H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018) (**IEEE**).
24. Qamar, S., Ahmad, P. & Shen, L. Hi-net: Hyperdense inception 3 d unet for brain tumor segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II, 6* 50–57 (Springer, 2021)
25. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B. & Rueckert, D. Attention U-Net: Learning Where to Look for the Pancreas. arXiv (2018)
26. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen Y.-W. & Wu, J. Unet 3+: A full-scale connected UNET for medical image segmentation. In *International Conference on Acoustics* 1055–1059 (IEEE, 2020).
27. Xie, X. et al. Discriminative features pyramid network for medical image segmentation. *Biocybern. Biomed. Eng.* **44**(2), 327–340 (2024).
28. Wang, G. et al. A skin lesion segmentation network with edge and body fusion. *Appl. Soft Comput.* **170**, 112683 (2025).
29. Katar, C., Eryilmaz, O. & Eksioglu, E. Att-next for skin lesion segmentation with topological awareness. *Expert Syst. Appl.* 127637 (2025)
30. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J. & Feichtenhofer, C. Multiscale vision transformers. In: *International Conference on Computer Vision (ICCV)* 6824–6835 (2021)
31. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. & Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv (2021)
32. Zhou, H., Guo, J., Zhang, Y., Yu, L., Wang, L. & Yu, Y. nnFormer: Interleaved Transformer for Volumetric Segmentation. arXiv (2022)
33. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H. & Wei, Y. Deformable convolutional networks. In: I*nternational Conference on Computer Vision (ICCV)* 764–773 (2017)
34. Mu, T., He, K. & Xu, D. Deformable coordinate kernel attention-based network for medical image segmentation. In: *Sixteenth International Conference on Graphics and Image Processing (ICGIP 2024)* vol. 13539, pp. 181–190 (SPIE, 2025)
35. Yang, X., Li, Z., Guo, Y. & Zhou, D. Dcu-net: A deformable convolutional neural network based on cascade u-net for retinal vessel segmentation. *Multimedia Tools Appl.* **81**(11), 15593–15607 (2022).
36. Ma, J., Li, F. & Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. Preprint arXiv:2401.04722 (2024)
37. Li, X. et al. Transiam: Aggregating multi-modal visual features with locality for medical image segmentation. *Expert Syst. Appl.* **237**, 121574 (2024).
38. Gutman, D., Codella, N.C.F., Celebi, E., Helba, B., Marchetti, M., Mishra, N. & Halpern, A. 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)
39. 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**(1), 180161 (2018).
40. Mendonça, T., Ferreira, P.M., Marques, J.S., Marçal, A.R. & Rozeira, J. Ph2-a dermoscopic image database for research and benchmarking. In: *35th Annual International Conference of Engineering in Medicine and Biology Society (EMBC)* 5437–5440 (IEEE, 2013).
41. Xiong, X., Wu, Z., Tan, S., Li, W., Tang, F., Chen, Y., Li, S., Ma, J. & Li, G. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. Preprint arXiv:2408.08870 (2024)
42. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021) (**Nature Publishing Group**).
43. Heidari, M., Kazerouni, A., Soltany, M., Azad, R., Aghdam, E.K., Cohen-Adad, J. & Merhof, D. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In: *IEEE Winter Conference on Applications of Computer Vision* 6202–6212 (2023)
44. Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In: (eds De Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y. & Essert, C.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021* vol. 12901, pp. 14–24 (Springer, 2021). Lecture Notes in Computer Science
45. Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E.K., Velichko, Y., Bagci, U. & Merhof, D. Beyond self-attention: Deformable large kernel attention for medical image segmentation. In: *IEEE Winter Conference on Applications of Computer Vision* 1287–1297 (2024)
46. Li, X. et al. SUnet: A multi-organ segmentation network based on multiple attention. *Comput. Biol. Med.* **167**, 107596 (2023).
47. Sun, Y. et al. Msca-net: Multi-scale contextual attention network for skin lesion segmentation. *Pattern Recogn.* **139**, 109524 (2023).

Acknowledgements

This work was supported in part by the UK EPSRC under Grant EP/P009727/2, and the Leverhulme Trust under Grant RF-2019-492. The authors also express their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-17).

Author contributions

The contributions of the authors are as follows: Saqib Qamar: Data curation, model development, and manuscript writing. Syed Furqan Qadri: Validation and manuscript writing. Roobaea Alroobaea: Provided resources for implementation and result validation. Goram Mufarah M Alshmrani: Assisted in figure making and validation. Richard Jiang: Supervised and validated all aspects of the manuscript. Mohd Fazil: Helped in revising the manuscript and validated all aspects of the manuscript.

Funding

This research was funded by Taif University, Taif, Saudi Arabia project number (TU-DSPP-2024-17).

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This study did not involve animals or human participants.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-17300-x.

Correspondence and requests for materials should be addressed to S.Q. or R.J.

Reprints and permissions information is available at www.nature.com/reprints.

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