IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 October 2023, accepted 26 November 2023, date of publication 28 November 2023,
date of current version 5 December 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3337528*

RESEARCH ARTICLE

# DEU-Net: Dual-Encoder U-Net for Automated Skin Lesion Segmentation

**ALI KARIMI[1], KARIM FAEZ[1], (Life Member, IEEE), AND SOHEILA NAZARI[2]**
[1]Department of Electrical Engineering, Amirkabir University of Technology, Tehran 15875-4413, Iran
[2]Faculty of Electrical Engineering, Shahid Beheshti University, Tehran 19839-69411, Iran

Corresponding author: Soheila Nazari (soheilanazari@aut.ac.ir)

**ABSTRACT** The computer-aided diagnosis (CAD) of skin diseases relies heavily on automated skin lesion segmentation, albeit presenting considerable challenges due to lesion diversity in shape, size, color, and texture, as well as potential blurry boundaries with surrounding tissues. Traditional Convolutional Neural Networks (CNN) typically underperform in this domain, given their inherent constraints in global context information capture. In the present study, we present a new U-shaped network, Dual-Encoder U-Net (DEU-Net), which is based on an encoder-decoder architecture. DEU-Net integrates a dual-encoder branch comprising a convolutional encoder and a transformer encoder, thereby facilitating the concurrent extraction of local features and global contextual information. Additionally, in order to enhance the performance of DEU-Net, we employ an integrated test-time augmentation technique. To ascertain the efficiency and superiority of our proposed methodology, we performed comprehensive experiments across four widely accessible skin lesion datasets, namely ISIC 2016, ISIC 2017, ISIC 2018, and PH2. The Dice coefficients achieved on these datasets were 92.90%, 87.16%, 90.81%, and 95.65%, respectively. These results demonstrate superior performance compared to most current state-of-the-art methods. The source code is released at https://github.com/alikm6/DEU-Net.

**INDEX TERMS** Convolutional neural networks, dermoscopy images, skin lesion segmentation, transformer.

## I. INTRODUCTION

Skin cancer is a widespread and potentially lethal disease with four main types: basal cell carcinoma, typically caused by sun exposure and radiation therapy, which grows slowly and rarely spreads; squamous cell carcinoma, often a result of sun exposure or skin damage, with a 2-5% likelihood of spreading; Merkel cell cancer, a rare, highly aggressive type that originates in hormone-producing cells beneath the skin; and melanoma, the most aggressive type, originating from melanocytes and responsible for the majority of skin cancer-related deaths, despite accounting for just 1% of cases. Projections by the American Cancer Society (ACS) for the year 2023 anticipate around 97,610 new melanoma cases in the United States alone, with approximately 7,990 fatalities resulting from the disease [1]. However, the prognosis can be drastically improved with early detection and diagnosis, allowing for the simple excision of the melanoma and ensuring a full recovery. The 5-year survival rate surpasses 99% with early diagnosis, while it plummets below 32% in cases of late detection [2]. These statistics underscore the vital role of precise medical image analysis in the timely diagnosis and treatment of skin ailments.

Dermoscopy, a non-invasive dermatological imaging modality [3], enhances the visibility and clarity of skin lesions by providing magnification and illumination. Applying specific materials to the skin reduces light reflection of the skin surface, making visual features more discernable. Clinical examinations leveraging dermoscopic imaging are significantly more accurate than diagnoses based solely on unaided observation, boosting diagnostic sensitivity indices by 10-27% [4].

Historically, dermatologists have visually identified malignant melanoma from dermoscopy images. However, this approach is often time-consuming and monotonous [5] and can result in diagnostic inaccuracies or inconsistencies,

The associate editor coordinating the review of this manuscript and approving it for publication was Huiyu Zhou.

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

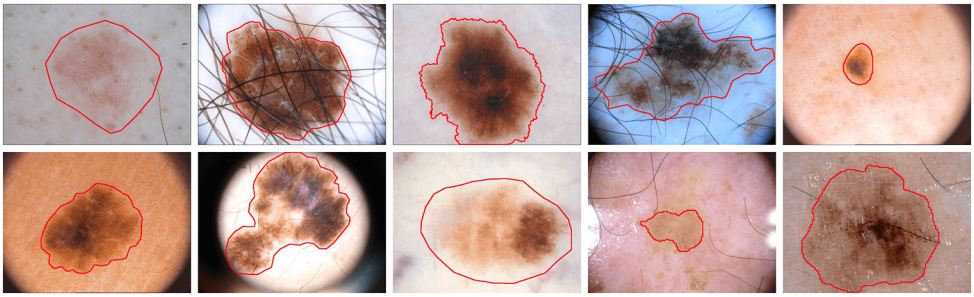


**FIGURE 1. Examples of skin lesions obtained from the ISIC 2017 dataset [12] highlight challenges in automated segmentation due to variations in skin color, texture, lesion size, site shape, contrast with the surrounding skin, and presence of artifacts.**

given its reliance on individual expertise [6]. With the advent of computer vision, skin lesion segmentation has become crucial in computer-aided diagnosis (CAD) of skin diseases. This advancement aids clinicians in quickly and accurately interpreting dermoscopic images while providing insightful medical image analyses [7]. Studies have confirmed that accurate skin lesion area segmentation and subsequent background noise reduction enhance the diagnostic accuracy of both dermatological and computational methods [8], [9], [10].

However, automated segmentation of skin lesions in dermoscopic images, aiming to differentiate them from the healthy surrounding skin, is a complex and challenging task. This difficulty arises from the diverse range of patient-specific factors, including skin color, texture, lesion size, lesion site shape, contrast between the lesion and non-lesion areas, and the existence of multiple artifacts such as shadows, reflections, uneven lighting, body hair, and air bubbles [11]. Fig. 1 illustrates several skin lesions for which these challenges hinder accurate segmentation. Therefore, deep learning algorithms must achieve high accuracy to tackle skin lesion segmentation tasks effectively.

Traditional skin lesion segmentation methods typically employ hand-crafted feature-based techniques to distinguish lesion borders from the surrounding skin. These techniques include thresholding methods [13], region-based methods [14], clustering-based methods [15], and others. However, these methods generally lack stability and robustness, resulting in less-than-ideal segmentation outcomes, especially when dealing with lesions with significant variations. Additionally, these conventional techniques typically necessitate the extraction of pre-defined image features. Deep learning methodologies have been developed to improve upon these limitations, leveraging convolutional neural networks (CNNs) to learn image features, thereby enhancing segmentation performance.

Over the past several years, a range of deep convolutional neural networks, such as the Fully Convolutional Network (FCN) [16] and U-Net [17], have found extensive use across multiple domains, especially in the realm of medical image segmentation. U-Net, in particular, has emerged as a commonly employed network architecture for medical image segmentation composed of encoding and decoding pathways. Numerous U-Net variants, including U-Net++ [18], 3D U-Net [19], V-Net [20], and others, have demonstrated exceptional performance across a range of medical image segmentation tasks employing various imaging techniques. Nonetheless, these approaches often overlook the critical global contextual information, which is imperative for accurate skin lesion localization. In essence, the semantic segmentation of pixels involves long-range dependencies that hold substantial significance in medical imagery, particularly for delineating boundary pixels. Consequently, enriching feature maps with global context information and understanding long-range dependencies among pixels within medical images could contribute to more precise localization and boundary demarcation of skin lesions, thereby improving segmentation performance.

The practical utility of U-Net in elevating the outcomes of numerous medical segmentation tasks is mainly attributable to the skip connections bridging the encoder and decoder. This encoder-decoder framework, strengthened by skip connections, enables U-Net to facilitate the effective extraction of input data's low-level and high-level features. However, during sequential sampling processes, the loss of spatial and global contextual information may restrict improvements to segmentation accuracy. Additionally, the successive up-sampling in the decoding phase, which relies on feature maps of a higher level, often overlooks the intricate spatial information embedded within feature maps of a lower level. Consequently, acquiring more global contextual information is critical to improving segmentation performance [5]. Researchers have proposed various strategies to enlarge receptive fields inspired by advancements in dilated convolution [21], [22]. Lee et al. [23] utilized dilated convolution throughout their network to remedy the issue of ambiguous boundaries, enabling the prediction of boundary key-point maps to steer the attention module. Furthermore, Wang et al. [24] have implemented non-local interaction modeling to calculate the response at a specific location through a weighted summation of features across all locations within the given feature maps, aiming to comprehend

long-range dependencies. The non-local attention mechanism can be regarded as a basic form of self-attention, given its ability to compute the interrelations amongst all pair-wise positions present within the input feature maps. In contemporary times, the transformer model [25] effectively extracts long-range dependencies by utilizing self-attention mechanisms, which have proven beneficial in natural language processing and computer vision. In contrast to non-local neural networks, the Vision Transformer (ViT) [26] can capture long-range dependencies with multiple parallel attention heads. Additionally, the Swin transformer [27] uses shifted windows and hierarchical feature fusion, effectively handling long-range dependencies within the data. Furthermore, the MaxViT [28], by introducing the multi-axis self-attention (Max-SA) block, reduced the computational complexity of ViT from quadratic to linear without losing non-locality.

In this study, we introduce Dual-Encoder U-Net (DEU-Net), an innovative segmentation network derivative of the U-Net [17], specifically crafted to tackle the complex task of skin lesion segmentation. Drawing inspiration from the pioneering FAT-Net [5], our approach utilizes a dual encoder comprising convolutional and transformer branches. This dual framework allows us to extract local features and global contextual information concurrently, a vital component in skin lesion segmentation. To optimize the fusion of features derived from the final layer of both the convolutional and transformer encoders, we adopt the use of the squeeze and excitation (SE) module [29]. The SE module effectively activates the more efficient channels and suppresses the less useful ones by adjusting the channel weights within the feature map. Furthermore, inspired by PCANet [30] and recognizing the proven success of data distillation [31] and model distillation [32] methods, we have incorporated the integrated test-time augmentation technique in our network testing phase. This method synthesizes insights from several models and transformations at the testing stage, enhancing model robustness and improving performance. Finally, to assess the effectiveness of our proposed approach, we have conducted tests on four separate datasets: ISIC 2016 [33], ISIC 2017 [12], ISIC 2018 [34], [35], and PH2 [36]. The results confirm that our novel approach yields promising results. Our research can be encapsulated within the following main contributions:

- Introducing a novel network, DEU-Net, amalgamates the strengths of convolutional and transformer networks for superior skin lesion segmentation. By replacing the single branch encoder characteristic of traditional U-Net architectures with a dual encoder in our DEU-Net, we can capture rich global contextual information for skin lesion segmentation alongside local features.
- The application of the integrated test-time augmentation technique, which consolidates the predictions of several models and various transformations at the test stage. This method leads to enhanced robustness of the model and superior results.
- A comprehensive comparison of our approach with pre-existing approaches using ISIC 2016, ISIC 2017, ISIC 2018, and PH2 datasets. The results of our experiments demonstrate superior accuracy with our model. The visual results also validate the effectiveness of our approach in detailed segmentation.

This paper is structured as follows: Section II presents related works, Section III describes the methodology, Section IV presents the validation and results, Section V presents the discussion and limitations, and Section VI provides the conclusion.

## II. RELATED WORKS

### A. SKIN LESION SEGMENTATION NETWORK

In the realm of skin lesion segmentation, conventional approaches primarily relied on hand-crafted low-level features extracted from the images. Celebi et al. [37] proposed a novel unsupervised approach that utilizes the statistical region merging algorithm for detecting boundaries in dermoscopic images. Peruch et al. [38] developed a skin lesion segmentation system that mimics the process followed by dermatologists, involving feature detection, dimensionality reduction, noise reduction, clustering, and post-processing. However, these feature-based methods face challenges in selecting discriminative features and determining appropriate hyperparameters, which limits their development.

In the past few years, there has been a notable advancement in image processing by applying deep learning techniques. Specifically, segmentation approaches based on convolutional neural networks (CNNs) have proven highly effective in skin lesion segmentation, delivering exceptional outcomes. In contrast to conventional feature-based approaches, CNN-based approaches for skin lesion segmentation do not rely on explicit feature definitions. Yuan et al. [39] introduced a 19-layer deep CNN with a novel loss function and a set of training strategies for fully automatic skin lesion segmentation. Jha et al. [40] introduced the DoubleU-Net, which stacks two U-Net architectures in parallel to capture multi-scale image features and improve accuracy. Hasan et al. [41] employed depth-wise separable convolutions in their model to optimize parameter count and enhance network efficiency. Dong et al. [42] proposed the FAC-Net, which consists of a feedback fusion block and an attention block to extract critical features and achieve superior segmentation performance effectively. Dai et al. [43] presented the Ms RED network, which utilizes multi-scale residual encoding and decoding to handle challenging cases and achieve robustness. Bi et al. [44] developed a novel semi-automatic segmentation approach utilizing fully convolutional networks (FCN), which mitigates information loss by combining user inputs and image features in multiple steps. Lei et al. [45] introduced a novel approach for skin lesion segmentation by employing a generative adversarial network (GAN) and leveraging joint learning to enhance the decision-making process of the discriminative module. Despite their substantial contributions to the progress of skin

lesion segmentation, these methods often overlook the importance of extracting global contextual information, which is crucial for accurate segmentation.

### B. TRANSFORMER-BASED NETWORK

The Transformer [25] model, originally derived from natural language processing, has recently gained significant attention in image classification, semantic segmentation, and object recognition. Dosovitskiy et al. [26] introduced the Vision Transformer (ViT), which divides the input image into patches of size $16 \times 16$, flattens them into a sequence, and employs the vanilla transformer encoder for classification. Nevertheless, the computational cost of training these models is significantly elevated due to the substantial parameter count and intricate computations involved. To tackle this problem, Liu et al. [27] proposed a hierarchical Swin Transformer that enables the exchange of information across windows and reduces computational complexity through a sliding window strategy while maintaining high performance. Tu et al. [28] developed MaxViT, which preserves non-locality while reducing the computational complexity of ViT from quadratic to linear by utilizing the multi-axis self-aware (Max-SA) block.

The applicability of the Transformer model in semantic segmentation tasks has garnered attention from researchers due to its remarkable capabilities. Several studies have investigated the utilization of the vanilla Transformer in semantic segmentation. Cao et al. [46] introduced the Swin-Unet model, where Swin Transformer blocks are employed in the U-Net architecture without any convolutional operations. Lin et al. [47] introduced DS-TransUNet, which utilizes input patches of different scales and two parallel Swin Transformers as encoders to capture richer information. The results demonstrated that Transformers significantly aid in capturing global contextual information, which can compensate for the limitations of CNNs. This has led to various hybrid models that integrate both CNNs and Transformers. Chen et al. [48] presented TransUNet, which integrates Transformer blocks between the encoder and decoder to model low-level CNN features globally. Zhang et al. [49] employed both CNN and Transformer in parallel to process the input image. To merge these processing results, they incorporated the BiFusion module. Wu et al. [5] presented a dual encoder system, which effectively captures local and global contextual information by combining CNN and Transformer architectures within a novel feature-adaptive transformer network (FAT-Net). Wang et al. [50] proposed the boundary-aware transformer (BAT) network, employing the boundary-wise attention gate (BAG) to leverage prior knowledge of boundaries.

## III. METHODS

### A. NETWORK

The proposed DEU-Net (Dual-Encoder U-Net) network takes inspiration from U-Net [17], FAT-Net [5], EfficientNet [51], MaxViT [28], and SENet [29] and targets skin lesion segmentation. As illustrated in Fig. 2, DEU-Net's architecture closely follows the encoder-decoder pattern seen in the U-Net network, enhanced by skip connections for optimal segmentation performance. It comprises three sections: an encoder, a center, and a decoder. Uniquely, our encoder integrates two separate encoders - an EfficientNet convolutional encoder and a MaxViT transformer encoder - working in parallel to facilitate extracting local features and global contextual information correspondingly. These features are crucial for accurately segmenting skin lesions. The feature maps generated by these two encoders are subsequently merged in the center part of the network and fed into the decoder. The decoder performs the up-sampling operation progressively until a dense prediction map is generated. It is noteworthy that the skip connections only leverage the feature map of the convolutional encoder.

#### 1) ENCODER

The sophisticated task of skin lesion segmentation, characterized by their ambiguous boundaries, irregular patterns, and variations in shape, necessitates excellent feature extraction. This extraction ranges from local to global long-range dependencies, pivotal for differentiating lesion and background pixels. Inspired by FAT-Net [5], our U-Net network employs a dual encoder, illustrated in Fig. 2, that utilizes a convolutional encoder to derive local features and a transformer encoder to capture global long-range dependencies.

The convolutional encoder harnesses the power of the EfficientNet network, a high-performing image model known for its efficient use of computing resources. The heart of this network is the Mobile Inverted Bottleneck Convolution (MBConv), detailed in Fig. 2. MBConv consists of four layers; the first layer expands the input into a higher dimensional space using a $1 \times 1$ convolution; the second employs depthwise convolution, a more computationally efficient alternative to traditional convolutions, as it operates on each input channel individually; the third layer deploys the squeeze and excitation (SE) module [29], which enables adaptive reweighting of channel-wise feature responses, and the fourth "compresses" the features into a lower dimensional space using a $1 \times 1$ convolution, thereby retaining only essential features. To aid rapid convergence and prevent the gradient vanishing problem, a residual connection similar to ResNet [52] is utilized. In summary, the MBConv module is formulated as follows:

$$X' = \mathrm{Conv}_{1\times1}\left(\mathrm{SE}\left(\mathrm{DepthwiseConv}\left(\mathrm{Conv}_{1\times1}\left(X\right)\right)\right)\right) \quad (1)$$

$$Y = X + X' \quad (2)$$

where $X$ represents the input feature map, $Y$ represents the output feature map, and $\mathrm{SE}(\cdot)$ represents the squeeze and excitation operation.

Nevertheless, this CNN-based encoder suffers from a limited effective receptive field that prevents it from capturing

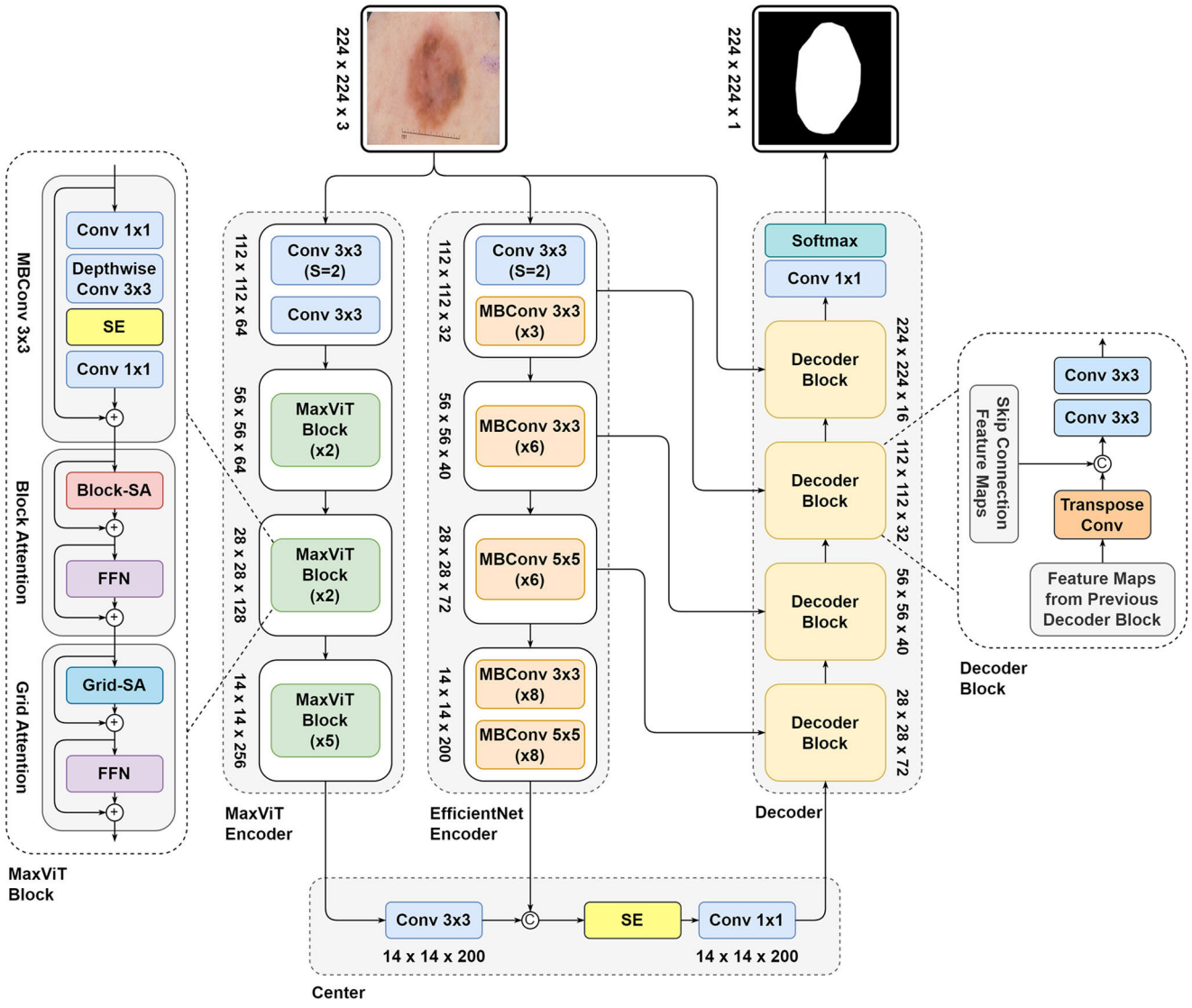



**FIGURE 2.** Overview of the proposed DEU-Net. The architecture, based on U-Net, incorporates a dual-encoder configuration. The symbols' +' and 'C' represent element-wise summation and concatenation, respectively. For simplification, the normalization and activation layers are excluded.

global contextual information, leading to potential inaccuracies in skin lesion segmentation.

Recognizing the successes of transformers in computer vision tasks [26], our encoder integrates an additional transformer-based branch to capture global features. In contrast to convolutional layers, which extract information in the vicinity of pixels and gradually build long-range dependencies through layer stacking, transformer layers capture global context information directly. We use the MaxViT [28] network for the encoder's second branch, a convolution and transformer hybrid. It introduces a novel attention module, the multi-axis self-attention (Max-SA), visually shown in Fig. 3. It decomposes the fully dense attention mechanisms into two more lightweight variants – block attention and grid attention. This restructuring effectively mitigates the quadratic complexity associated with vanilla attention, resulting in a linear complexity, all while retaining the crucial non-locality aspect. In the MaxViT framework, Relative self-attention [27], [53], [54], [55] has been chosen over standard self-attention [25], [26] to introduce a relative learned bias to the attention weights. This strategy has demonstrated superior performance on multiple vision tasks compared to the original attention mechanism [27], [53], [54].

The advantage of self-attention over local convolution is its capability for global interaction. However, due to the quadratic complexity of the attention operator, it is computationally untenable to apply attention across the entire space. The MaxViT network proposes a multi-axis approach to tackle this issue, decomposing full-size attention into local and global forms through spatial axes decomposition.

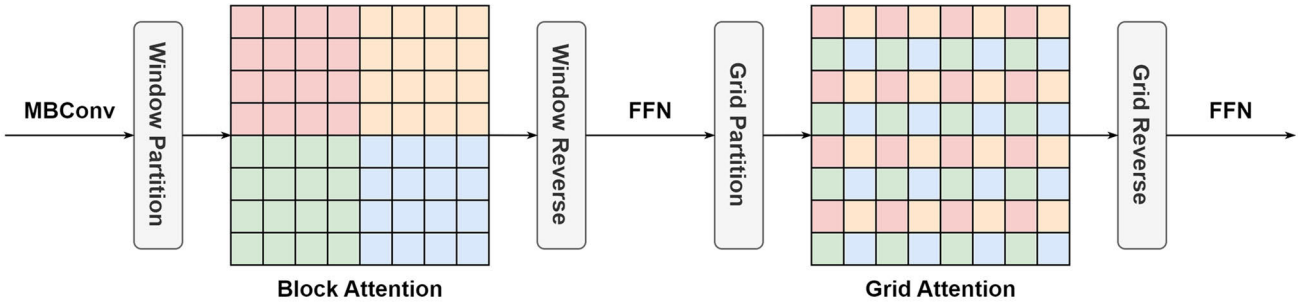



**FIGURE 3.** The visual representation of the multi-axis self-attention (Max-SA) featuring two modules: block-attention, which operates within specific windows, and grid-attention, which globally attends to pixels using a sparse, uniform grid spanning the entire 2D space; both exhibit linear complexity relative to input size due to the utilization of consistent attention footage. [28].

Consider an input feature map represented as $X \in \mathbb{R}^{H\times W\times C}$. Rather than deploying attention to the flattened spatial dimension HW, the feature map is partitioned into a tensor conforming to the shape $\left(\frac{H}{P} \times \frac{W}{P}, P \times P, C\right)$. This operation results in non-overlapping windows, each of which is $P \times P$ in size. Employing self-attention on this local spatial dimension, specifically $P \times P$, is equipollent to focus within a restricted window. This process is colloquially known as ''block attention'' and is instrumental in fostering local interactions.

Local-attention models, while avoiding the hefty computation of full self-attention, tend to underfit on large-scale datasets. The concept of grid attention is introduced to remedy this. Inspired by block attention, grid attention eschews fixed window size partitioning in favor of gridding the tensor into the shape $\left(G \times G, \frac{H}{G} \times \frac{W}{G}, C\right)$ with a fixed $G \times G$ uniform grid. This results in adaptive-size windows $\frac{H}{G} \times \frac{W}{G}$, and applying self-attention on the subdivided grid axis, denoted by $G \times G$, is equivalent to a dilated, global fusion of tokens in the spatial domain. Maintaining consistent window and grid dimensions (where $P = G = 7$) harmonizes the computational load among local and global processes, each demonstrating a linear complexity concerning spatial dimensions or sequence length.

The MaxViT Block, seen in Fig. 2, combines MBConv, block attention, and grid attention modules, empowering the network to capture local and global features from shallow to deep stages. The combined use of MBConv and attention increases network generalizability and trainability. An additional benefit arises from placing MBConv layers before attention, wherein depth-wise convolutions function as conditional position encoding (CPE) [56], eliminating the requirement for separate positional encoding layers within the model.

### 2) CENTER

In the center part of the network, features from the convolutional and transformer encoders are merged and forwarded to the network decoder. Initially, a $3 \times 3$ convolution is applied to the transformer encoder's feature map, equalizing its channel count to the convolutional encoder's feature map. Subsequently, these feature maps are concatenated, doubling the channel count compared to a single encoder. Such an increase could potentially compromise memory efficiency. Therefore, the features must be compacted. Although combining local and global contexts through simple convolution reduces channel numbers, it often fails to capture crucial feature correlations among channels, inhibiting improvements in segmentation accuracy. The squeeze and excitation (SE) module [29] addresses this issue. Initially, global average pooling is employed to squeeze the input feature maps $f \in \mathbb{R}^{H\times W\times C}$ into channel-wise statistics $s \in \mathbb{R}^{1\times 1\times C}$. Afterward, utilizing a sigmoid activation function, a simple gating mechanism effectively adapts channel-wise statistics into input feature maps by capturing channel-wise dependencies. Finally, the channels of the dual encoder's feature maps were reweighted via the SE module, selectively activating beneficial channels and suppressing redundant ones, leading to improved fusion of local and global contexts. Following the SE module, a $1 \times 1$ convolution halves the number of feature map channels. Finally, this processed feature map is delivered to the network decoder for further operations. The computation in the center part can be summarized as follows:

$$F'_{\text{tra}} = \text{Conv}_{3\times 3}\left(F_{\text{tra}}\right) \tag{3}$$

$$F_{\text{center}} = \text{Conv}_{1\times 1}\left(\text{SE}\left(\text{Concat}\left(F_{\text{cnn}}, F'_{\text{tra}}\right)\right)\right) \tag{4}$$

where $F_{\text{tra}}$ represents the feature map obtained from the last layer of the transformer encoder, $F_{\text{cnn}}$ represents the feature map obtained from the last layer of the convolutional encoder, and SE $(\cdot)$ represents the squeeze and excitation operation.

### 3) DECODER

The decoder part of the network conducts up-sampling layer by layer to generate pixel-level prediction results. As depicted in Fig. 2, the feature map procured from the antecedent decoder layer and the corresponding convolutional encoder layer's feature map is amalgamated within each decoder layer. Initially, the feature map from the previous decoder layer is up-sampled via a transposed convolution, effectively doubling its dimensions while halving its channel count. Subsequently, this updated feature map is concatenated with the feature map obtained from the corresponding convolutional encoder layer. Two $3 \times 3$ convolutional layers are

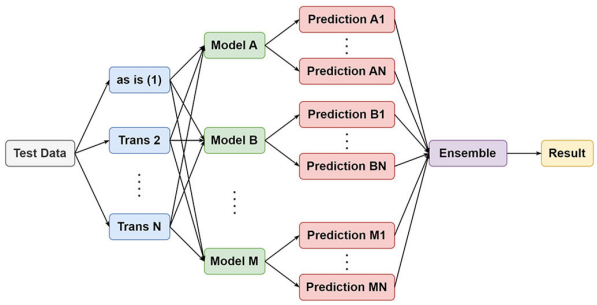



**FIGURE 4.** Diagram of the integrated test-time augmentation (ITTA) technique.

then utilized to integrate these two feature maps. In the final decoding layer, after applying transposed convolution, the feature map from the preceding decoder layer matches the input image's dimensions. This feature map is concatenated with the input image and forwarded through two $3 \times 3$ convolution layers. The above process can be summarized as follows:

$$F'_{\text{decoder}_{(i)}} = \begin{cases} \text{Concat}\left(F_{\text{cnn}_{(i-1)}}, \text{Deconv}\left(F_{\text{center}}\right)\right) & i = 4 \\ \text{Concat}\left(F_{\text{cnn}_{(i-1)}}, \text{Deconv}\left(F_{\text{decoder}_{(i+1)}}\right)\right) & i = 3, 2 \\ \text{Concat}\left(\text{Input}, \text{Deconv}\left(F_{\text{decoder}_{(i+1)}}\right)\right) & i = 1 \end{cases} \tag{5}$$

$$F_{\text{decoder}_{(i)}} = \text{Conv}_{3\times3}\left(\text{Conv}_{3\times3}\left(F'_{\text{decoder}_{(i)}}\right)\right) \quad i = 1, 2, 3, 4 \tag{6}$$

where Input represents the input image, $F_{\text{cnn}_{(i)}}$ represents the feature map obtained from the i-th layer of the convolutional encoder, $F_{\text{center}}$ represents the feature map obtained from the center part, $F_{\text{decoder}_{(i)}}$ represents the output feature map of the i-th layer of the decoder, and Deconv $(\cdot)$ represents the transposed convolution operation.

The final step towards pixel-level prediction employs a $1 \times 1$ convolution, which diminishes the number of feature map channels to two. These channels are subjected to a softmax function, generating the ultimate segmentation result. The above process can be summarized as follows:

$$F_{\text{final}} = \text{Softmax}\left(\text{Conv}_{1\times1}\left(F_{\text{decoder}}\right)\right) \tag{7}$$

where $F_{\text{final}}$ represents the network output and $F_{\text{decoder}}$ represents the feature map obtained from the final layer of the decoder.

### B. INTEGRATED TEST-TIME AUGMENTATION

We adopted the integrated test-time augmentation (ITTA) technique inspired by PCANet [30] to boost our network's test phase accuracy. The procedural diagram can be found in Fig. 4. The ITTA concept hinges on two primary principles. First, applying multiple transformations to test images and subsequent aggregation of results often lead to improved final predictions compared to single-transformation inferences [31]. Second, using an ensemble of models usually yields superior performance than any single model [32]. These principles highlight the merit of combining diverse models and transformations, as they provide complementary information-enhancing test outcomes.

The ITTA methodology follows a four-step process:

- Model Construction and Training: We build and train a group of base models on the training set, allowing each model to learn the dataset's unique features.
- Test Data Transformation: Test data is subjected to multiple transformations, creating distinct versions of each test image.
- Prediction of Transformed Images: Each transformed image is inputted into the trained models, generating unique predictions.
- Aggregation of Predictions: We combine all predictions from the several models and various transformations by averaging all outputs, leading to the final prediction result.

We leverage the synergistic potential of multiple models and transformations through this approach, thereby improving our model's skin lesion segmentation capability.

### C. LOSS FUNCTION

The design of our loss function is based on a hybrid approach incorporating both the Dice loss [20] and Cross-Entropy loss, each chosen due to their respective attributes that contribute significantly to the model's performance.

The Dice loss function, named after the Dice coefficient (a measure of the overlap between two samples), is notably effective in improving convergence speed and mitigating overfitting. The definition of the Dice loss function is as follows:

$$L_{\text{dice}}\left(p, y\right) = 1 - \frac{2\sum_i^N p_i y_i + \epsilon}{\sum_i^N p_i^2 + \sum_i^N y_i^2 + \epsilon} \tag{8}$$

Here, $p_i$ denotes the predicted probability of a pixel being part of the lesion class, $y_i$ is the corresponding ground truth label, $N$ is the total number of pixels in the image, and $\epsilon$ is a small constant to ensure numerical stability.

In the context of pixel-level classification tasks, the Cross-Entropy loss function demonstrates its suitability by efficiently classifying individual pixels into specific classes. The definition of the Cross-Entropy loss function is as follows:

$$L_{\text{CE}}\left(p, y\right) = -\sum_i^N \left[y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right)\right] \tag{9}$$

Here, $p_i$ denotes the predicted probability of a pixel being part of the lesion class, $y_i$ is the corresponding ground truth label, and $N$ is the total number of pixels in the image.

The combined loss function $L$ is expressed as a weighted sum of the Dice and Cross-Entropy losses, calculated as follows:

$$L\left(p, y\right) = 0.4 * L_{\text{dice}}\left(p, y\right) + 0.6 * L_{\text{CE}}\left(p, y\right) \tag{10}$$

**TABLE 1. Details of the ISIC 2016, ISIC 2017, ISIC 2018, and PH2 datasets.**

| Dataset | Total Images | Train/Validation/Test (Specified by the provider) | Train/Validation/Test (Used in implementation) | Resolution (pixel) |
|---|---|---|---|---|
| ISIC 2016 | 1279 | 900/0/379 | 788/112/379 | 566 × 679 to 2848 × 4288 |
| ISIC 2017 | 2750 | 2000/150/600 | 2000/150/600 | 566 × 679 to 4499 × 6748 |
| ISIC 2018 | 3694 | 2594/100/1000 | 1815/259/520 | 480 × 640 to 4519 × 6808 |
| PH2 | 200 | - | 140/20/40 | 560 × 768 |

The amalgamation of the Dice and Cross-Entropy losses ensures accurate pixel classification (a strength of the Cross-Entropy loss) and the creation of cohesive, continuous segments (an advantage of the Dice loss). The resulting combined loss function exploits the strengths of both loss mechanisms, compensating for their limitations, thereby enhancing overall model performance.

## IV. EXPERIMENTS

### A. DATASETS

Our approach underwent rigorous testing across four distinct skin lesion segmentation datasets, all publicly accessible and widely recognized for their significance in this field. The details of these datasets are outlined in Table 1. The International Skin Imaging Collaboration (ISIC) supplied the first three datasets, namely ISIC 2016 [33], ISIC 2017 [12], and ISIC 2018 [34], [35]. The ISIC's commitment to providing globally accessible, annotated skin lesion image datasets is instrumental in advancing computer-aided diagnosis (CAD) techniques for melanoma and other skin diseases, fostering the progression of automated diagnostic procedures [12]. The fourth dataset, PH2 [36], was contributed by the Dermatology Department at the Pedro Hispano Hospital in Matosinhos, Portugal. Each of these datasets has been previously used to evaluate various methodologies. Let us delve into the specifics of each dataset:

- **ISIC 2016:** The ISIC 2016 dataset includes 1279 skin lesion images, with 900 designated for training and 379 for testing. Without a pre-defined validation set, we randomly selected 112 images from the training dataset as validation data, leaving 788 images for training.
- **ISIC 2017:** The ISIC 2017 dataset features 2000 images for training, and 150 for validation, with an additional 600 images set aside for testing.
- **ISIC 2018:** The ISIC 2018 dataset includes 2594 training images, 100 validation images, and 1000 test images. As the ground truths for the test data are not publicly available, we randomly divided the training images into three groups: 1815 images (70%) for training, 259 (10%) for validation, and the remaining 520 (20%) for testing.
- **PH2:** The PH2 dataset has 200 skin lesion images. We distributed these randomly, assigning 140 (70%) images for training, 20 (10%) for validation, and the remaining 40 (20%) for testing.

### B. IMPLEMENTATION DETAILS

Our proposed network has been developed using the PyTorch platform on a computer equipped with an AMD Ryzen 9 5900HX CPU, 16GB DDR4 RAM, and an Nvidia GeForce RTX 3080 Laptop GPU. All training and testing were carried out under identical hardware conditions. The computational environment was configured with Python 3.9 as the programming language, utilizing PyTorch 2.0.0 as the fundamental framework for building the neural network structure and aiding in model debugging. Throughout the experiment, the resolution of all images utilized for training, validation, and testing was adjusted to a size of 224 × 224.

In order to optimize the initialization of model weights, we utilized the pre-trained EfficientNet-B6 [51] and MaxViT-T [28] weights from the ImageNet 2012 dataset [57] as the network encoder's foundation.

The Adam optimizer was utilized to conduct end-to-end training of the network. Each network underwent a 50-epoch training cycle, utilizing an initial learning rate of 0.001, momentum parameters set at $b1 = 0.5$ and $b2 = 0.999$, no weight decay, and batch size 16. After every epoch, evaluation metrics were calculated on the validation set, and if the Dice score improved, the model weights were saved for evaluation on the test set.

To diversify image samples, we adopted an augmentation approach inspired by the strategy delineated in FAT-Net [5]. Five distinct augmentation techniques were employed, such as vertical and horizontal rotations, angular rotations ranging from −15 to 15 degrees, random adjustments to the contrast and brightness within a −3% to 3% ratio, and random alterations to the hue, value, and saturation within the same ratio. While extensive augmentations risk distorting the distribution of color or brightness and compromising the integrity of original image information, the experiments demonstrated that any such corruption was negligible due to the minimal augmentation ratio. Crucially, this strategy of sample enrichment enhances our proposed segmentation network's generalization capability, facilitating a more effective capture of global contextual information and local features [5].

### C. EVALUATION METRICS

The effectiveness of the proposed network was gauged through widely accepted performance metrics used in skin lesion segmentation, namely Accuracy (ACC), Intersection over Union (IoU), Dice coefficient (Dice), Sensitivity (SE), and Specificity (SP).

**TABLE 2. Distribution of TP, TN, FP, and FN based on segmentation result and ground truth mask.**

| | Ground truth = skin lesion | Ground truth = background |
|---|---|---|
| Segmentation result = skin lesion | TP | FP |
| Segmentation result = background | FN | TN |

**TABLE 3. Performance comparison between single-encoder and dual-encoder architectures. The first value indicates the performance in the test set, and the second value indicates the performance in the validation set.**

| Backbone | Configuration Mode | ACC (%) | IoU (%) | Dice (%) | Number of Parameters |
|---|---|---|---|---|---|
| EfficientNet-B6 | All Layers | 93.94 / 95.48 | 77.77 / 79.29 | 85.89 / 87.05 | 54.7 |
| EfficientNet-B6 | Without Last Layer | 93.89 / 95.36 | 77.08 / 79.01 | 85.37 / 86.83 | 8.3 |
| MaxViT-T | All Layers | 93.55 / 95.12 | 76.90 / 78.50 | 85.19 / 86.52 | 33.4 |
| MaxViT-T | Without Last Layer | 93.65 / 95.34 | 76.51 / 77.49 | 84.67 / 85.71 | 13.5 |
| EfficientNet-B6 + MaxViT-T (proposed model) | Without Last Layer | **93.99** / **95.76** | **78.50** / **79.85** | **86.45** / **87.54** | 21.6 |

Accuracy is a comprehensive measure of the overall efficacy of the lesion image segmentation process. IoU, also known as the Jaccard index, assesses the intersection ratio between the derived segmentation outcomes and the actual ground truth mask. The Dice coefficient is another critical metric used to gauge similarity; it defines the extent of overlap between the predicted results and the ground truth in image segmentation. Sensitivity provides insight into the precision of skin lesion pixel segmentation, whereas Specificity quantifies the accurate segmentation of lesion-free regions.

The definitions for these metrics used to appraise segmentation outcomes are provided below:

$$\text{ACC} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{11}$$

$$\text{Dice} = \frac{2.\text{TP}}{2.\text{TP} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{IoU} = \text{JAC} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{15}$$

In these definitions, TP (True Positive) denotes the count of accurately segmented skin lesion pixels, while TN (True Negative) signifies the count of accurately segmented background pixels. Conversely, FP (False Positive) represents the misclassification of background pixels as skin lesion pixels, and FN (False Negative) indicates the incorrect prediction of skin lesion pixels as background pixels. A detailed breakdown of TP, TN, FP, and FN is provided in Table 2.

### D. ABLATION STUDIES

In this section, we provided an extensive ablation analysis to assess the individual contributions of different components to our proposed dual-encoder architecture. Our experiments exclusively utilized the ISIC 2017 dataset [12], and the other implementation settings remained consistent across all trials.

#### 1) SINGLE VERSUS DUAL-ENCODER ARCHITECTURE

In our first set of experiments, we investigated the impact of the dual-encoder architecture. To this end, we constructed two typical U-Net networks, each with a single encoder: one with an EfficientNet-B6 [51] encoder and another with a MaxViT-T [28] encoder. It is crucial to note that all backbones utilized in these networks were pre-trained on the ImageNet-1K dataset [57]. Two modes were considered for these single-encoder networks: one utilizing all five layers of the encoder backbone and another using only the first four layers. According to the results presented in Table 3, our dual-encoder architecture demonstrated superior performance compared to the single-encoder networks. Such a decrease occurs despite using all five layers of the backbone, which, although it increases the model parameter count, does not improve performance over the proposed dual-encoder configuration.

#### 2) CONVOLUTIONAL AND TRANSFORMER BACKBONES IN DUAL-ENCODER ARCHITECTURE

In our second series of experiments, we examined the influence of using different networks for the convolutional and transformer backbones in the dual-encoder architecture. We tested the ResNet-50 [52] and EfficientNet-B6 [51] as convolutional backbones and MaxViT-T [28] and Swin-V2-T [58] as transformer backbones, thus generating four distinct combinations. All these backbones were pre-trained on the ImageNet-1K dataset [57]. For each combination, similar to the previous experiment, we considered two modes: the first employing all layers of the backbones and the second omitting the last layer to create a lighter network.

Table 4 presents the outcomes of these experiments, demonstrating that the EfficientNet-B6 + MaxViT-T combination, utilizing only their initial four layers, yielded the most superior performance. In the earlier experiments, as seen in Table 3, increasing the layers in the encoding path in the single-encoder architecture enhanced network performance. However, as Table 4 reveals, this trend was reversed in the

**TABLE 4.** Performance comparison between different convolutional and transformer backbones in the dual-encoder architecture. The first value indicates the performance in the test set, and the second value indicates the performance in the validation set.

| Convolutional Backbone | Transformer Backbone | Configuration Mode | ACC (%) | IoU (%) | Dice (%) | Number of Parameters |
|---|---|---|---|---|---|---|
| ResNet-50 | MaxViT-T | All Layers | 93.14 / 95.62 | 76.08 / 78.23 | 84.77 / 86.15 | 125.7 |
| ResNet-50 | MaxViT-T | Without Last Layer | 93.65 / 95.61 | 77.78 / 79.08 | 85.81 / 87.04 | 38.9 |
| ResNet-50 | Swin-V2-T | All Layers | 92.75 / 94.94 | 74.98 / 76.83 | 83.51 / 85.07 | 127.8 |
| ResNet-50 | Swin-V2-T | Without Last Layer | 93.03 / 95.33 | 75.63 / 76.78 | 84.10 / 85.20 | 39.6 |
| EfficientNet-B6 | MaxViT-T | All Layers | 93.98 / **95.87** | 77.99 / 79.29 | 86.06 / 87.25 | 109.0 |
| EfficientNet-B6 | MaxViT-T | Without Last Layer | **93.99** / 95.76 | **78.50** / **79.85** | **86.45** / **87.54** | 21.6 |
| EfficientNet-B6 | Swin-V2-T | All Layers | 93.66 / 95.62 | 76.46 / 78.54 | 84.81 / 86.48 | 111.7 |
| EfficientNet-B6 | Swin-V2-T | Without Last Layer | 93.82 / 95.37 | 76.87 / 78.69 | 85.23 / 86.57 | 21.3 |

**TABLE 5.** Performance comparison of the models and the impact of integrated test-time augmentation technique. The first value indicates the performance in the test set, and the second value indicates the performance in the validation set.

| Model | Method | ACC (%) | IoU (%) | Dice (%) |
|---|---|---|---|---|
| M1 | Base | 93.99 / 95.76 | 78.50 / 79.85 | 86.45 / 87.54 |
| M2 | Base | 94.28 / 95.76 | 78.47 / 79.75 | 86.43 / 87.46 |
| M3 | Base | 94.33 / 96.02 | 78.64 / 79.63 | 86.39 / 87.32 |
| M1 | Common test-time augmentation | 94.10 / 95.77 | 78.95 / 80.39 | 86.75 / 87.92 |
| M2 | Common test-time augmentation | 94.33 / 95.94 | 78.70 / 80.17 | 86.65 / 87.59 |
| M3 | Common test-time augmentation | 94.40 / 96.06 | 79.00 / 80.24 | 86.66 / 87.92 |
| M1, M2, and M3 | Integrated test-time augmentation | **94.45 / 96.26** | **79.51 / 80.49** | **87.16 / 88.07** |

dual-encoder architecture, where adding more encoding layers diminished the model's performance. This observation can be rationalized by considering the substantial escalation in the model parameters (approximately 3 to 5 fold) when more layers are used in the dual-encoder setup. Given that the skin lesion segmentation task has a finite number of training images, the optimization of model parameters becomes challenging, inevitably leading to a decline in the final performance of the model.

In conclusion, our ablation studies substantiated the effectiveness of the dual-encoder architecture and the superiority of the EfficientNet-B6 + MaxViT-T combination for this task.

#### 3) INVESTIGATION OF INTEGRATED TEST-TIME AUGMENTATION TECHNIQUE

In the third part of our ablation study, we aimed to explore the effectiveness of the integrated test-time augmentation technique. We used the proposed model (EfficientNet-B6 + MaxViT-T Without Last Layer) as the baseline architecture for this experiment.

Initially, we trained three base models, M1, M2, and M3, on the ISIC 2017 training set. Owing to the stochastic nature of the training process, wherein a portion of the model weights are randomly initialized at each step, and the training images are provided in a randomized manner, the final weights for each model occupy unique spaces. Consequently, this leads to slight variances in the evaluation outcomes of the test data across these models. These performance metrics for M1, M2, and M3 are reported in the first three rows of Table 5.

Following this, we applied the common test-time augmentation technique. In this method, three new images are derived from the input test image using transformations: horizontal rotation, vertical rotation, and simultaneous horizontal and vertical rotation. These three transformed images, plus the original image are inputted into the model for prediction. The resulting outputs from these four images are then averaged to yield the final result. This process enhanced the performance of models M1, M2, and M3 during test time, as presented in rows four to six of Table 5.

In implementing the integrated test-time augmentation, we first performed the common test-time augmentation on the three base models. Following this, we calculated the average output from the ensemble of 12 heatmaps to generate the final result. As evident from the last row of Table 5, the integrated test-time augmentation technique improves the results and outperforms both the three base models and the common test-time augmentation technique.

### E. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, we comprehensively evaluated the effectiveness of our proposed approach compared to state-of-the-art methods. To establish a rigorous benchmark, we assessed the performance on multiple datasets: ISIC 2016, ISIC 2017, ISIC 2018, and PH2. The experimental results are presented below.

**TABLE 6.** Comparative analysis of various state-of-the-art segmentation methods on the ISIC 2016 test set.

| Method | Year | ACC (%) | IoU (%) | Dice (%) | SE (%) | SP (%) |
|---|---|---|---|---|---|---|
| U-Net [17] | 2015 | 94.37 | 81.53 | 88.57 | 91.76 | 95.71 |
| FAC-Net [42] | 2021 | 96.09 | 86.23 | 92.51 | 92.50 | **97.43** |
| Ms RED [43] | 2022 | 96.42 | 87.03 | 92.66 | - | - |
| FAT-Net [5] | 2022 | 96.04 | 85.30 | 91.59 | 92.59 | 96.02 |
| TC-Net [59] | 2022 | 96.06 | 86.68 | 92.82 | **93.17** | 97.12 |
| GFANet [60] | 2023 | 96.04 | 85.92 | 91.78 | 92.95 | 97.25 |
| EIU-Net [61] | 2023 | 95.90 | 85.50 | 91.90 | 91.80 | 94.30 |
| autoSMIM [62] | 2023 | 96.42 | 87.05 | 92.73 | - | - |
| Ours | - | **96.64** | **87.48** | **92.90** | 92.81 | 97.40 |

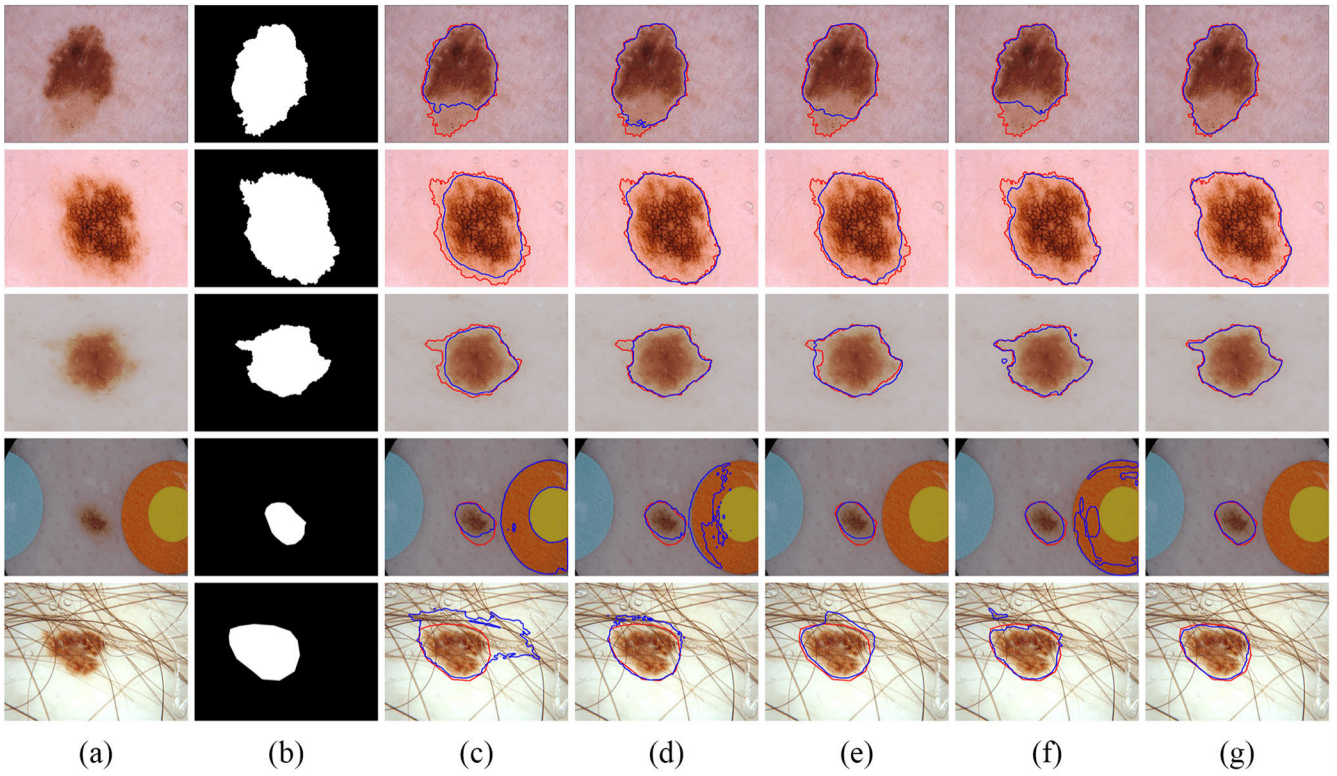



**FIGURE 5.** Comparative visualization of various state-of-the-art methods on the ISIC 2016 dataset. (a) Input images. (b) Ground truth. (c) U-Net [17]. (d) Ms RED [43]. (e) FAT-Net [5]. (f) EIU-Net [61]. (g) Ours. The ground truth and segmentation outcomes from various methods are represented by the red and blue lines correspondingly.

### 1) EVALUATION ON THE ISIC 2016 DATASET

We conducted a comparative analysis of the proposed DEU-Net with eight state-of-the-art approaches on the ISIC 2016 dataset, including U-Net [17], FAC-Net [42], Ms RED [43], FAT-Net [5], TC-Net [59], GFANet [60], EIU-Net [61], and autoSMIM [62]. While U-Net is a general model for medical image segmentation, the other approaches were developed specifically for skin lesion segmentation, among which FAT-Net and TC-Net leverage transformer-based methods for this task. As illustrated in Table 6, our DEU-Net demonstrated superior performance, scoring 96.64%, 87.48%, 92.90%, 92.81%, and 97.40% in ACC, IoU, Dice, SE, and SP metrics, respectively. It outperformed the other approaches in the critical metrics of ACC, IoU, and Dice, achieving marginal improvements of 0.22% in ACC over the second-ranked Ms RED and autoSMIM, 0.43% in IoU over the second-ranked autoSMIM, and 0.08% in Dice over the second-ranked TC-Net. Furthermore, our DEU-Net demonstrated considerable enhancements compared to U-Net, achieving improvements of 2.27%, 5.95%, and 4.33% in ACC, IoU, and Dice metrics, respectively.

Additionally, we visually compared the segmentation output from our proposed DEU-Net and four other approaches, namely U-Net, Ms RED, FAT-Net, and EIU-Net, in Fig. 5. These images depict five challenging skin lesion samples.

**TABLE 7. Comparative analysis of various state-of-the-art segmentation methods on the ISIC 2017 test set.**

| Method | Year | ACC (%) | IoU (%) | Dice (%) | SE (%) | SP (%) |
|---|---|---|---|---|---|---|
| U-Net [17] | 2015 | 92.67 | 74.40 | 83.27 | 81.74 | 97.47 |
| FAC-Net [42] | 2021 | 93.63 | 74.27 | 84.91 | 81.06 | 97.43 |
| Ms RED [43] | 2022 | 94.10 | 78.55 | 86.48 | - | - |
| FAT-Net [5] | 2022 | 93.26 | 76.53 | 85.00 | 83.92 | 97.25 |
| Act-AttSegNet [63] | 2022 | 93.50 | 79.20 | **87.20** | **89.70** | 96.80 |
| NCRNet [64] | 2022 | 94.01 | 78.62 | 86.55 | 86.89 | 95.88 |
| TC-Net [59] | 2022 | 93.68 | 74.55 | 85.20 | 81.45 | 97.79 |
| GFANet [60] | 2023 | 93.97 | 77.75 | 85.74 | 81.37 | 97.87 |
| CL-DCNN [65] | 2023 | 94.10 | 79.10 | 86.70 | 86.50 | 95.90 |
| RMMLP [66] | 2023 | - | 78.33 | 86.60 | - | - |
| EIU-Net [61] | 2023 | 93.70 | 77.10 | 85.50 | 84.20 | 96.80 |
| Ours | - | **94.45** | **79.51** | 87.16 | 84.50 | **98.25** |

Our proposed DEU-Net showed superior performance, generating results closer to the ground truth mask, particularly in cases involving ambiguous boundaries and artifacts such as patient hair and color calibration charts. This improved performance is attributed to the DEU-Net's ability to extract richer global context information compared to the other methods.

### 2) EVALUATION ON THE ISIC 2017 DATASET

In the case of the ISIC 2017 dataset, we conducted a comprehensive evaluation of our proposed approach in comparison to eleven state-of-the-art approaches, namely U-Net [17], FAC-Net [42], Ms RED [43], FAT-Net [5], Act-AttSegNet [63], NCRNet [64], TC-Net [59], GFANet [60], CL-DCNN [65], RMMLP [66], and EIU-Net [61]. Among them, FAT-Net and TC-Net stand out as transformer-based approaches for skin lesion segmentation. Table 7 shows the statistical outcomes regarding the segmentation of skin lesions on the ISIC 2017 dataset for different approaches. Our approach achieved superior performance across most metrics, with 94.45%, 79.51%, 87.16%, 84.50%, and 98.25% in ACC, IoU, Dice, SE, and SP metrics, respectively. Compared to U-Net, our approach achieved improvements of 1.78%, 5.11%, 3.89%, 2.76%, and 0.78% in ACC, IoU, Dice, SE, and SP metrics, respectively. Regarding the Dice score, our approach ranks second with a negligible difference of 0.04% compared to Act-AttSegNet. However, it outperforms Act-AttSegNet by 0.95% and 0.31% in the ACC and IoU metrics, which are crucial for accurate lesion segmentation.

Furthermore, a visual comparison is provided in Fig. 6, showcasing several examples of the segmentation results from our proposed approach and other methods. To facilitate this analysis, we have specifically chosen four approaches for evaluation: U-Net, Ms RED, FAT-Net, and EIU-Net. The comparative analysis of the visual results substantiates the superior performance of our approach against other participants, mainly when dealing with challenging cases involving ambiguous boundary situations. The illustrations showcase how our proposed approach successfully delineates lesion boundaries, even in the presence of hair and other artifacts, and in images with minimal differentiation between the skin and the lesion.

### 3) EVALUATION ON THE ISIC 2018 DATASET

We evaluated the proposed DEU-Net's performance on the ISIC 2018 dataset in comparison to ten state-of-the-art methods, including U-Net [17], Ms RED [43], FAT-Net [5], FTN [67], AS-Net [68], UNeXt [69], ICL-Net [70], GFANet [60], EIU-Net [61], and autoSMIM [62]. Among these methods, U-Net and UNeXt were designed for medical image segmentation, while the rest were tailored specifically for skin lesion segmentation. Notably, three of these methods, FAT-Net, FTN, and ICL-Net, are transformer-based. As depicted in Table 8, our DEU-Net exhibited competitive performance against these methods, securing scores of 96.57%, 84.44%, 90.81%, 92.40%, and 97.51% across the ACC, IoU, Dice, SE, and SP metrics, respectively. Our technique outperformed all others in the critical IoU and Dice metrics, enhancing the IoU and Dice scores by 0.48% and 0.41%, respectively, compared to the runner-up, autoSMIM. In the ACC metric, our method ranked third, trailing slightly by 0.53% and 0.13% to ICL-Net and EIU-Net, respectively. However, we surpassed these two techniques in the pivotal IoU and Dice metrics. Moreover, compared to U-Net, our approach demonstrated significant improvements, enhancing the ACC, IoU, and Dice metrics by 1.26%, 4.66%, and 3.59%, respectively.

To further illustrate the performance of DEU-Net, we showcased the segmentation results from five challenging ISIC 2018 dataset samples in Fig. 7, using our proposed method alongside four other approaches: U-Net, FAT-Net, UNeXt, and EIU-Net. It is apparent that our technique excelled in skin lesion segmentation compared to the other methodologies. Our method demonstrated robust

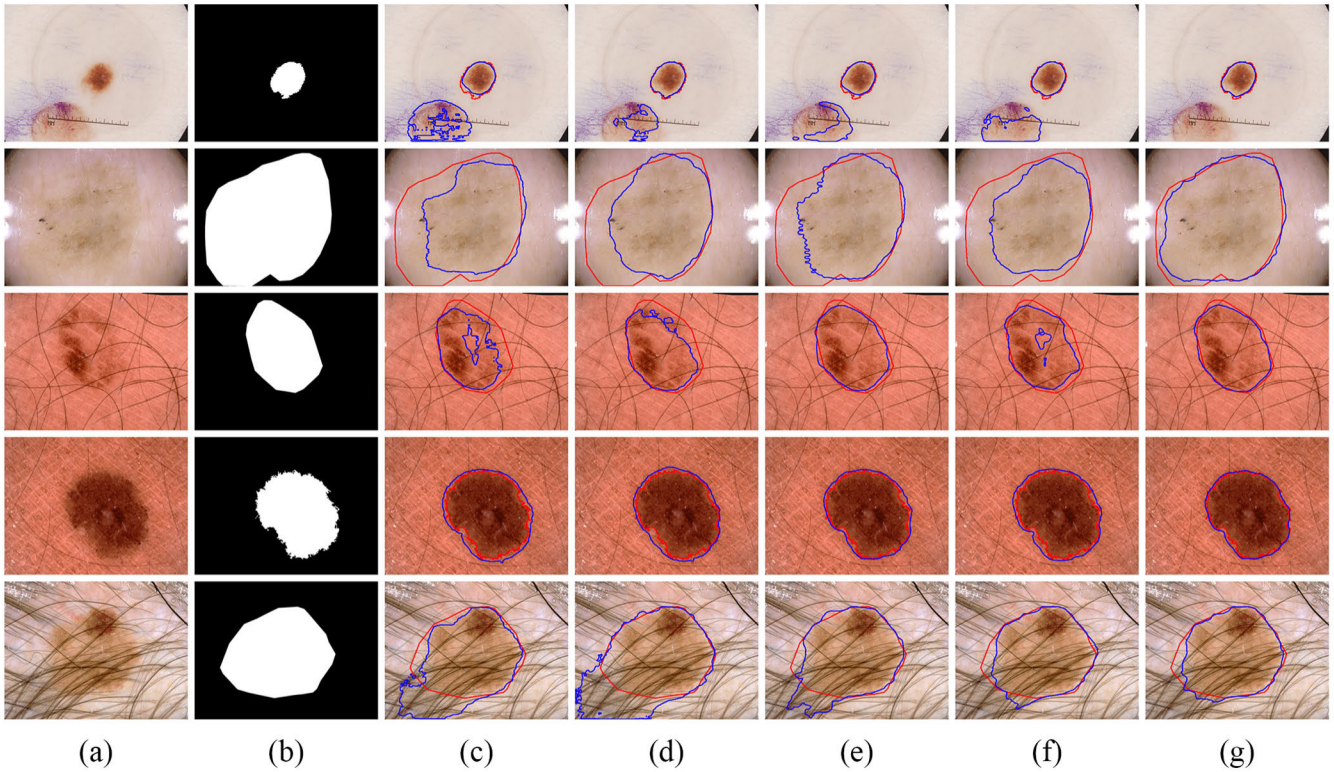



**FIGURE 6.** Comparative visualization of various state-of-the-art methods on the ISIC 2017 dataset. (a) Input images. (b) Ground truth. (c) U-Net [17]. (d) Ms RED [43]. (e) FAT-Net [5]. (f) EIU-Net [61]. (g) Ours. The ground truth and segmentation outcomes from various methods are represented by the red and blue lines correspondingly.

**TABLE 8.** Comparative analysis of various state-of-the-art segmentation methods on the ISIC 2018 test set.

| Method | Year | ACC (%) | IoU (%) | Dice (%) | SE (%) | SP (%) |
|---|---|---|---|---|---|---|
| U-Net [17] | 2015 | 95.31 | 79.78 | 87.22 | 89.36 | 97.88 |
| Ms RED [43] | 2022 | 96.19 | 83.45 | 89.99 | - | - |
| FAT-Net [5] | 2022 | 95.78 | 82.02 | 89.03 | 91.00 | 96.99 |
| FTN [67] | 2022 | 96.20 | 82.80 | 89.80 | **96.20** | 97.50 |
| AS-Net [68] | 2022 | 95.68 | 83.09 | 89.55 | 93.06 | 94.69 |
| UNeXt [69] | 2022 | - | 81.70 | 89.70 | - | - |
| ICL-Net [70] | 2023 | **97.10** | 83.27 | 90.03 | 90.82 | **98.49** |
| GFANet [60] | 2023 | 96.29 | 83.66 | 90.13 | 90.75 | 97.79 |
| EIU-Net [61] | 2023 | 96.70 | 83.60 | 90.20 | 90.70 | 96.70 |
| autoSMIM [62] | 2023 | 96.21 | 83.96 | 90.40 | - | - |
| Ours | - | 96.57 | **84.44** | **90.81** | 92.40 | 97.51 |

performance even with complex samples featuring blurred lesion borders, low lesion-to-skin contrast, patient hair artifacts, and varied lesion dimensions. Despite these challenges, DEU-Net generated segmentation results that closely resembled the ground truth masks.

#### 4) EVALUATION ON THE PH2 DATASET

The performance of the proposed DEU-Net on the PH2 dataset was evaluated by comparing it to eight state-of-the-art segmentation methods: U-Net [17], Ms RED [43], FAT-Net [5], Act-AttSegNet [63], AttSwinUNet [71], GFA-Net [60], ULFAC-Net [72], and AMCC-Net [73]. Notably, FAT-Net and AttSwinUNet are transformer-based methods specifically designed for skin lesion segmentation. A statistical comparison for the PH2 dataset can be found in Table 9. As the result indicates, our proposed methodology outshines most competitors across various metrics. Specifically, the DEU-Net achieved scores of 97.58%, 91.81%, 95.65%, 94.60%, and 98.13% in ACC, IoU, Dice, SE, and SP metrics, respectively. Compared to the conventional U-Net,

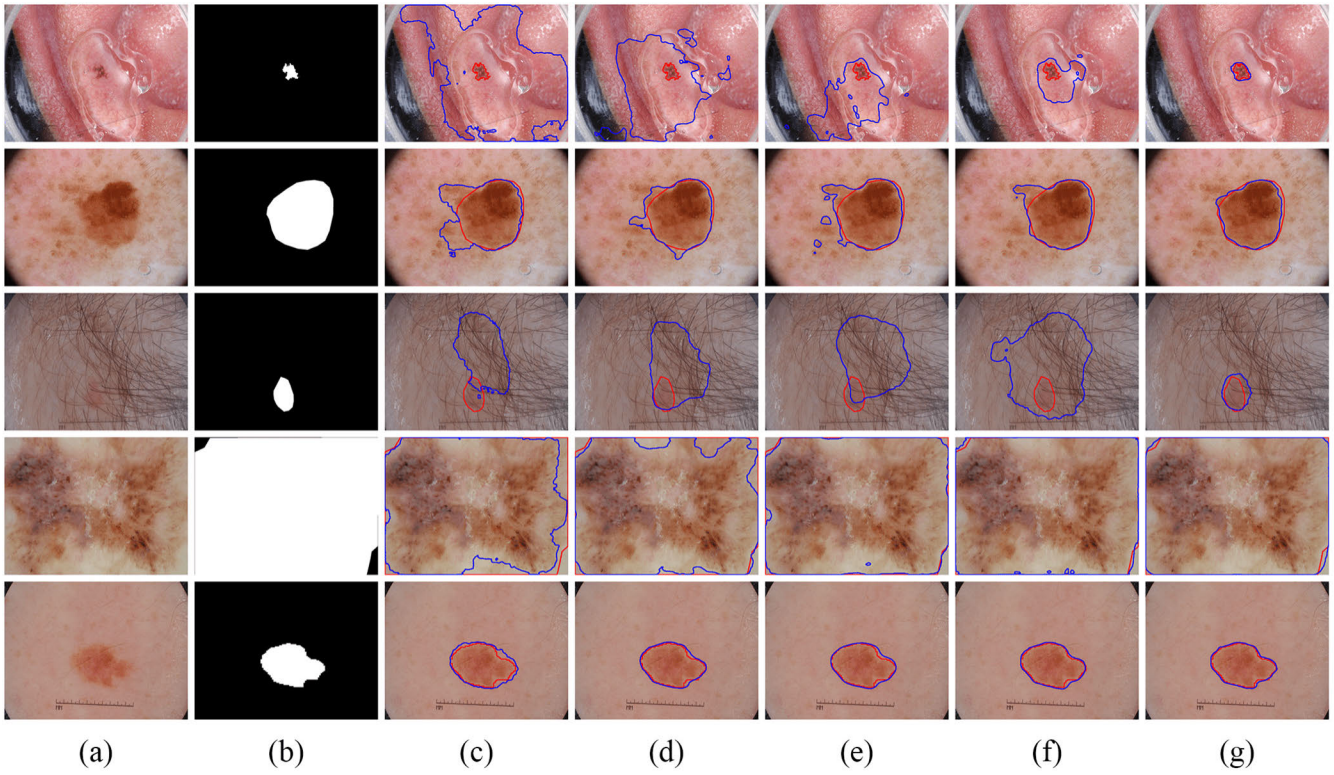



**FIGURE 7.** Comparative visualization of various state-of-the-art methods on the ISIC 2018 dataset. (a) Input images. (b) Ground truth. (c) U-Net [17]. (d) FAT-Net [5]. (e) UNeXt [69] (f) EIU-Net [61]. (g) Ours. The ground truth and segmentation outcomes from various methods are represented by the red and blue lines correspondingly.

**TABLE 9.** Comparative analysis of various state-of-the-art segmentation methods on the PH2 test set.

| Method | Year | ACC (%) | IoU (%) | Dice (%) | SE (%) | SP (%) |
|---|---|---|---|---|---|---|
| U-Net [17] | 2015 | 93.16 | 81.70 | 89.00 | 90.66 | 95.07 |
| Ms RED [43] | 2022 | 96.80 | 90.14 | 94.65 | - | - |
| FAT-Net [5] | 2022 | 97.03 | 89.62 | 94.40 | 94.41 | 97.41 |
| Act-AttSegNet [63] | 2022 | 95.10 | 90.10 | 94.70 | **96.30** | 96.40 |
| AttSwinUNet [71] | 2022 | 96.85 | - | 95.04 | 94.39 | 95.76 |
| GFANet [60] | 2023 | 97.09 | 90.98 | 95.06 | 96.08 | 97.57 |
| ULFAC-Net [72] | 2023 | 97.01 | 91.28 | 95.29 | 95.63 | 97.42 |
| AMCC-Net [73] | 2023 | 97.00 | 89.00 | 94.00 | - | - |
| Ours | - | **97.58** | **91.81** | **95.65** | 94.60 | **98.13** |

our DEU-Net enhances skin lesion segmentation performance by 4.42%, 10.11%, and 6.65% in ACC, IoU, and Dice metrics, respectively. These significant improvements underscore the importance of leveraging global long-range context when segmenting skin lesions.

Continuing the trend observed in other datasets, Fig. 8 offers a visual comparison of segmentation results for five challenging samples from the PH2 dataset test set. These results were obtained using our proposed DEU-Net method and four other approaches: U-Net, Ms RED, FAT-Net, and AttSwinUNet. The comparative analysis clearly demonstrates that our proposed method performs superiorly to competitors, producing results that closely align with the ground truth mask. DEU-Net is particularly successful in demarcating lesion boundaries within samples characterized by ambiguous boundaries, which is a testament to its efficiency and robustness.

## V. DISCUSSION AND LIMITATIONS

In medical image segmentation, encoder-decoder structures have gained significant popularity. Nevertheless, many of these models have overlooked the significance of capturing long-range dependencies, resulting in subpar network performance. Incorporating ample global context information

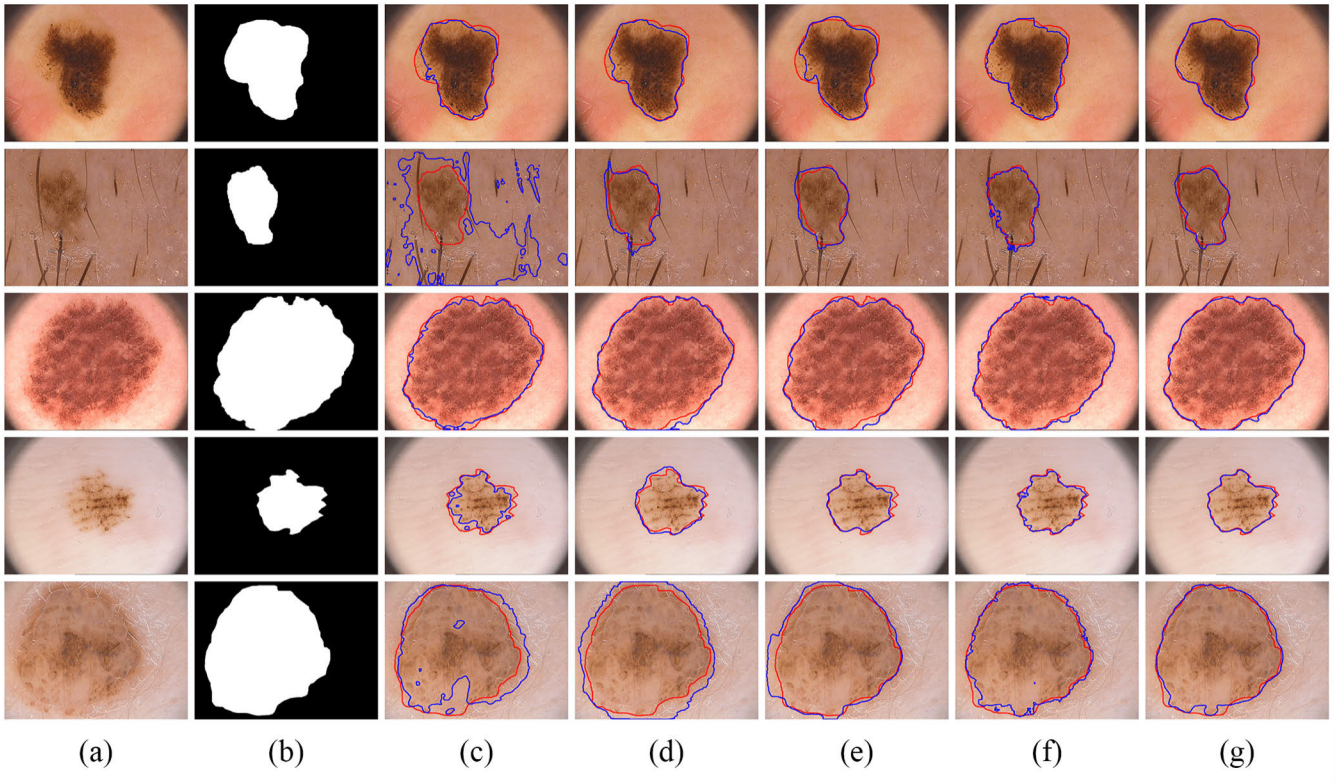



**FIGURE 8.** Comparative visualization of various state-of-the-art methods on the PH2 dataset. (a) Input images. (b) Ground truth. (c) U-Net [17]. (d) Ms RED [43]. (e) FAT-Net [5]. (f) AttSwinUNet [71]. (g) Ours. The ground truth and segmentation outcomes from various methods are represented by the red and blue lines correspondingly.

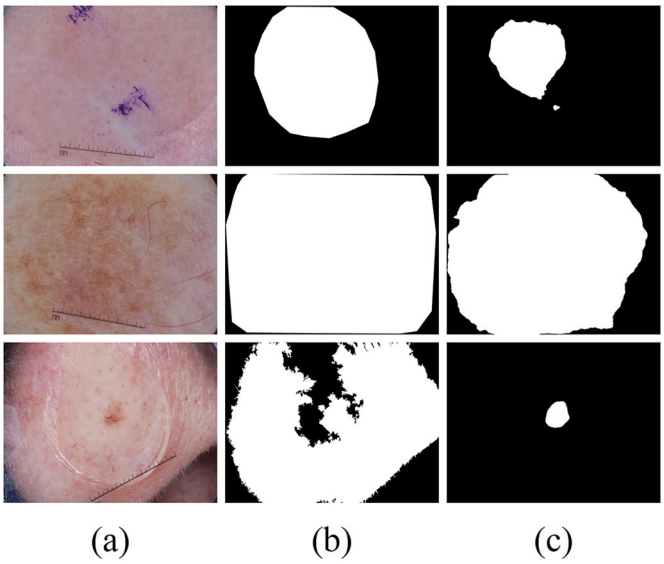



**FIGURE 9.** Visualization of challenging and unsuccessful cases of skin lesion segmentation. (a) Input images. (b) Ground truth. (c) DEU-Net.

proves invaluable in enabling the network to accurately delineate lesion boundaries, particularly in complex cases characterized by indistinct lesion-background boundaries, irregular shapes, and large dimensions. We incorporated a dual encoder within the U-Net framework to enhance feature extraction in the skin lesion segmentation task. This dual encoder combines two branches, one with convolutional layers and the other with transformer components, allowing us to extract local features and global contextual information simultaneously. In certain earlier methods,

such as FAT-Net [5], there was an attempt to extract local features and global contextual information concurrently by employing a fusion of convolutional and transformer networks. However, they did not prioritize choosing the most compatible convolutional and transformer backbones for seamless cooperation. In our research, we have thoroughly assessed several backbones for this specific task, and the pairing of EfficientNet-B6 and MaxViT-T networks has proven to deliver the most outstanding performance. Comprehensive ablation studies and comparative experiments on diverse datasets, including ISIC 2016, ISIC 2017, ISIC 2018, and PH2, showcase the advantages of our dual encoder. Furthermore, we employed the integrated test-time augmentation technique, which improved results by combining predictions from multiple models and various transformations and increased the model's robustness.

However, despite our segmentation network outperforming previous approaches in most evaluation metrics, it still has limitations when confronted with samples characterized by high irregularities and low contrasts. Fig. 9 illustrates instances of these complex cases where our method failed to segment the lesion accurately.

## VI. CONCLUSION

In this study, we presented the Dual Encoder U-Net (DEU-Net), a novel U-shaped network architecture crafted specifically for skin lesion segmentation. The DEU-Net incorporates a convolutional branch in its encoder responsible for extracting local features and an auxiliary transformer branch to amplify the extraction of more complex long-range dependencies. Such an approach is pivotal in handling the challenging task of skin lesion segmentation. To boost the network's effectiveness and robustness, we further incorporated the Integrate test-time augmentation strategy during the testing phase of the network. Moreover, we implemented a thorough evaluation of the suggested approach by conducting a series of rigorous experiments across four publicly available skin lesion datasets: ISIC 2016, ISIC 2017, ISIC 2018, and PH2. The empirical results indicated that our approach outperformed other state-of-the-art skin lesion segmentation methodologies and yielded outputs more aligned with the ground truth mask.

In our proposed approach, we used a simple method in the central part of the DEU-Net to fuse the feature maps obtained from the two convolutional and transformer branches. In future work, we aim to improve the model's performance by enhancing the fusion of feature maps obtained from the two encoder branches.

## REFERENCES


[1] R. L. Siegel, K. D. Miller, N. S. Wagle, and A. Jemal, "Cancer statistics, 2023," *CA, Cancer J. Clinicians*, vol. 73, no. 1, pp. 17–48, Jan. 2023, doi: 10.3322/caac.21763.

[2] American Cancer Society. *Cancer Facts & Figures 2023*. Accessed: Jun. 17, 2023. [Online]. Available: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf

[3] M. Binder, "Epiluminescence microscopy. A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists," *Arch. Dermatol.*, vol. 131, no. 3, pp. 286–291, Mar. 1995, doi: 10.1001/archderm.131.3.286.

[4] J. Mayer, "Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma," *Med. J. Aust.*, vol. 167, no. 4, pp. 206–210, Aug. 1997, doi: 10.5694/j.1326-5377.1997.tb138847.x.

[5] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102327, doi: 10.1016/j.media.2021.102327.

[6] X. Tong, J. Wei, B. Sun, S. Su, Z. Zuo, and P. Wu, "ASCU-Net: Attention gate, spatial and channel attention U-Net for skin lesion segmentation," *Diagnostics*, vol. 11, no. 3, Mar. 2021, Art. no. 3, doi: 10.3390/diagnostics11030501.

[7] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2482–2493, Jul. 2020, doi: 10.1109/TMI.2020.2972964.

[8] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001, doi: 10.1109/42.918473.

[9] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, Sep. 2007, doi: 10.1016/j.compmedimag.2007.01.003.

[10] P. Tang, Q. Liang, X. Yan, S. Xiang, and D. Zhang, "GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 10, pp. 2870–2882, Oct. 2020, doi: 10.1109/JBHI.2020.2977013.

[11] M. A. Khan, M. I. Sharif, M. Raza, A. Anjum, T. Saba, and S. A. Shad, "Skin lesion segmentation and classification: A unified framework of deep neural network features fusion and selection," *Expert Syst.*, vol. 39, no. 7, p. e12497, Aug. 2022, doi: 10.1111/exsy.12497.

[12] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.

[13] H. Iyatomi, H. Oka, M. Saito, A. Miyake, M. Kimoto, J. Yamagami, S. Kobayashi, A. Tanikawa, M. Hagiwara, K. Ogawa, G. Argenziano, H. P. Soyer, and M. Tanaka, "Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system," *Melanoma Res.*, vol. 16, no. 2, pp. 183–190, Apr. 2006, doi: 10.1097/01.cmr.0000215041.76553.58.

[14] A. Wong, J. Scharcanski, and P. Fieguth, "Automatic skin lesion segmentation via iterative stochastic region merging," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 929–936, Nov. 2011, doi: 10.1109/TITB.2011.2157829.

[15] D. D. Gomez, C. Butakoff, B. K. Ersboll, and W. Stoecker, "Independent histogram pursuit for segmentation of skin lesions," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 157–161, Jan. 2008, doi: 10.1109/tbme.2007.910651.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention* (Lecture Notes in Computer Science), N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[18] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, in Lecture Notes in Computer Science, D. Stoyanov, Z. Taylor, G. Carneiro, T. Syeda-Mahmood, A. Martel, L. Maier-Hein, J. M. R. S. Tavares, A. Bradley, J. P. Papa, V. Belagiannis, J. C. Nascimento, Z. Lu, S. Conjeti, M. Moradi, H. Greenspan, and A. Madabhushi, Eds. Cham, Switzerland: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[19] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention*, in Lecture Notes in Computer Science, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds. Cham, Switzerland: Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[20] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571, doi: 10.1109/3DV.2016.79.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," Apr. 2015, *arXiv:1511.07122*, doi: 10.48550/arXiv.1511.07122.

[22] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 636–644.

[23] H. J. Lee, J. U. Kim, S. Lee, H. G. Kim, and Y. M. Ro, "Structure boundary preserving segmentation for medical image with ambiguous boundary," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4816–4825.

[24] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2017.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[28] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in *Computer Vision—ECCV 2022* (Lecture Notes in Computer Science), S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham, Switzerland: Springer, 2022, pp. 459–479, doi: 10.1007/978-3-031-20053-3_27.

[29] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[30] Y. Zhang, Y. Chen, and K. Zhang, "PCANet: Pyramid context-aware network for retinal vessel segmentation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2073–2080, doi: 10.1109/ICPR48806.2021.9412773.

[31] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4119–4128.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. NIPS Deep Learn. Represent. Learn. Workshop*, 2015, pp. 1–9.

[33] D. Gutman, N. C. F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)," May 2016, *arXiv:1605.01397*, doi: 10.48550/arXiv.1605.01397.

[34] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, and A. Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," Mar. 2019, *arXiv:1902.03368*, doi: 10.48550/arXiv.1902.03368.

[35] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Aug. 2018, Art. no. 1, doi: 10.1038/sdata.2018.161.

[36] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "$PH^2$—A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440, doi: 10.1109/EMBC.2013.6610779.

[37] M. E. Celebi, H. A. Kingravi, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, R. H. Moss, J. M. Malters, J. M. Grichnik, A. A. Marghoob, H. S. Rabinovitz, and S. W. Menzies, "Border detection in dermoscopy images using statistical region merging," *Skin Res. Technol.*, vol. 14, no. 3, pp. 347–353, Aug. 2008, doi: 10.1111/j.1600-0846.2008.00301.x.

[38] F. Peruch, F. Bogo, M. Bonazza, V.-M. Cappelleri, and E. Peserico, "Simpler, faster, more accurate melanocytic lesion segmentation through MEDS," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 557–565, Feb. 2014, doi: 10.1109/TBME.2013.2283803.

[39] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017, doi: 10.1109/TMI.2017.2695227.

[40] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564, doi: 10.1109/CBMS49503.2020.00111.

[41] M. K. Hasan, L. Dahal, P. N. Samarakoon, F. I. Tushar, and R. Martí, "DSNet: Automatic dermoscopic skin lesion segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103738, doi: 10.1016/j.compbiomed.2020.103738.

[42] Y. Dong, L. Wang, S. Cheng, and Y. Li, "FAC-Net: Feedback attention network based on context encoder network for skin lesion segmentation," *Sensors*, vol. 21, no. 15, p. 5172, Jul. 2021, doi: 10.3390/s21155172.

[43] D. Dai, C. Dong, S. Xu, Q. Yan, Z. Li, C. Zhang, and N. Luo, "Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102293, doi: 10.1016/j.media.2021.102293.

[44] L. Bi, M. Fulham, and J. Kim, "Hyper-fusion network for semi-automatic segmentation of skin lesions," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102334, doi: 10.1016/j.media.2021.102334.

[45] B. Lei, Z. Xia, F. Jiang, X. Jiang, Z. Ge, Y. Xu, J. Qin, S. Chen, T. Wang, and S. Wang, "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, Aug. 2020, Art. no. 101716, doi: 10.1016/j.media.2020.101716.

[46] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Computer Vision—2022 Workshops* (Lecture Notes in Computer Science), L. Karlinsky, T. Michaeli, and K. Nishino, Eds. Cham, Switzerland: Springer, 2023, pp. 205–218, doi: 10.1007/978-3-031-25066-8_9.

[47] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin transformer U-Net for medical image segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–15, 2022, doi: 10.1109/TIM.2022.3178991.

[48] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," Feb. 2021, *arXiv:2102.04306*, doi: 10.48550/arXiv.2102.04306.

[49] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*, in Lecture Notes in Computer Science, M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham, Switzerland: Springer, 2021, pp. 14–24, doi: 10.1007/978-3-030-87193-2_2.

[50] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in *Medical Image Computing and Computer Assisted Intervention* (Lecture Notes in Computer Science), M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, Eds. Cham, Switzerland: Springer, 2021, pp. 206–216, doi: 10.1007/978-3-030-87193-2_20.

[51] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[53] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, 2021, pp. 3965–3977.

[54] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associate, 2021, pp. 14745–14758.

[55] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 464–468, doi: 10.18653/v1/N18-2074.

[56] X. Chu, Z. Tian, B. Zhang, X. Wang, and C. Shen, "Conditional positional encodings for vision transformers," presented at the 11th Int. Conf. Learn. Represent., Feb. 2023.

[57] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[58] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin transformer v2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11999–12009.

[59] Y. Dong, L. Wang, and Y. Li, "TC-Net: Dual coding network of transformer and CNN for skin lesion segmentation," *PLoS ONE*, vol. 17, no. 11, Nov. 2022, Art. no. e0277578, doi: 10.1371/journal.pone.0277578.

[60] S. Qiu, C. Li, Y. Feng, S. Zuo, H. Liang, and A. Xu, "GFANet: Gated fusion attention network for skin lesion segmentation," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106462, doi: 10.1016/j.compbiomed.2022.106462.

[61] Z. Yu, L. Yu, W. Zheng, and S. Wang, "EIU-Net: Enhanced feature extraction and improved skip connections in U-Net for skin lesion segmentation," *Comput. Biol. Med.*, vol. 162, Aug. 2023, Art. no. 107081, doi: 10.1016/j.compbiomed.2023.107081.

[62] Z. Wang, J. Lyu, and X. Tang, "AutoSMIM: Automatic superpixel-based masked image modeling for skin lesion segmentation," *IEEE Trans. Med. Imag.*, early access, Jun. 28, 2023, doi: 10.1109/TMI.2023.3290700.

[63] T.-T. Tran and V.-T. Pham, "Fully convolutional neural network with attention gate and fuzzy active contour model for skin lesion segmentation," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 13979–13999, Apr. 2022, doi: 10.1007/s11042-022-12413-1.

[64] Q. Liu, J. Wang, M. Zuo, W. Cao, J. Zheng, H. Zhao, and J. Xie, "NCRNet: Neighborhood context refinement network for skin lesion segmentation," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105545, doi: 10.1016/j.compbiomed.2022.105545.

[65] Y. Wang, J. Su, Q. Xu, and Y. Zhong, "A collaborative learning model for skin lesion segmentation and classification," *Diagnostics*, vol. 13, no. 5, Feb. 2023, Art. no. 5, doi: 10.3390/diagnostics13050912.

[66] C. Ji, Z. Deng, Y. Ding, F. Zhou, and Z. Xiao, "RMMLP: Rolling MLP and matrix decomposition for skin lesion segmentation," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104825, doi: 10.1016/j.bspc.2023.104825.

[67] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, Apr. 2022, Art. no. 102357, doi: 10.1016/j.media.2022.102357.

[68] K. Hu, J. Lu, D. Lee, D. Xiong, and Z. Chen, "AS-Net: Attention synergy network for skin lesion segmentation," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117112, doi: 10.1016/j.eswa.2022.117112.

[69] J. M. J. Valanarasu and V. M. Patel, "UNeXt: MLP-based rapid medical image segmentation network," in *Medical Image Computing and Computer Assisted Intervention*, (Lecture Notes in Computer Science), L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham, Switzerland: Springer, 2022, pp. 23–33, doi: 10.1007/978-3-031-16443-9_3.

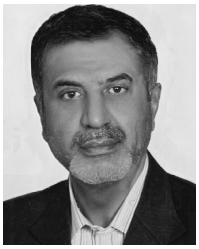
[70] W. Cao, G. Yuan, Q. Liu, C. Peng, J. Xie, X. Yang, X. Ni, and J. Zheng, "ICL-Net: Global and local inter-pixel correlations learning network for skin lesion segmentation," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 145–156, Jan. 2023, doi: 10.1109/JBHI.2022.3162342.

[71] E. K. Aghdam, R. Azad, M. Zarvani, and D. Merhof, "Attention Swin U-Net: Cross-contextual attention mechanism for skin lesion segmentation," Oct. 2022, *arXiv:2210.16898*, doi: 10.48550/arXiv.2210.16898.

[72] Y. Ma, L. Wu, Y. Gao, F. Gao, J. Zhang, and Z. Luo, "ULFAC-Net: Ultra-lightweight fully asymmetric convolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 6, pp. 2886–2897, Jun. 2023, doi: 10.1109/JBHI.2023.3259802.

[73] C. Dayananda, N. Yamanakkanavar, T. Nguyen, and B. Lee, "AMCC-Net: An asymmetric multi-cross convolution for skin lesion segmentation on dermoscopic images," *Eng. Appl. Artif. Intell.*, vol. 122, Jun. 2023, Art. no. 106154, doi: 10.1016/j.engappai.2023.106154.

**ALI KARIMI** received the B.S. degree in electrical engineering from the Shahrood University of Technology, in 2020. He is currently pursuing the M.S. degree in electronic engineering with the Electrical Engineering Department, Amirkabir University of Technology, Iran. His current research interests include computer vision, artificial intelligence, and deep learning.



**KARIM FAEZ** (Life Member, IEEE) received the B.S. degree (Hons.) in electrical engineering from Tehran Polytechnic University, in June 1973, and the M.S. and Ph.D. degrees in computer science from the University of California at Los Angeles (UCLA), in 1977 and 1980, respectively. Before joining the Amirkabir University of Technology, Iran, he was with the Iran Telecommunication Research Center, from 1981 to 1983. He was the Founder of the Computer Engineering Department, Amirkabir University, in 1989. His current research interests include pattern recognition, biometric identification and recognition, image processing, steganography, neural networks, signal processing, farsi handwritten recognition, earthquake signal processing, fault-tolerant system design, and computer networks. He is a member of IEICE and ACM. He has served as the first chairperson, from April 1989 to September 1992. He was the Chairperson of the Planning Committee for Computer Engineering and Computer Science of the Ministry of Science, Research and Technology, from 1988 to 1996.

**SOHEILA NAZARI** received the M.Sc. and Ph.D. degrees in electronic engineering from the Amirkabir University of Technology, Tehran, Iran, in 2014 and 2018, respectively. Her current research interests include digital circuit design, signal processing, image processing, neuromorphic engineering, artificial intelligence, and bioinspired pattern recognition.