2024 13th International Conference on Electrical and Computer Engineering (ICECE) | 979-8-3315-3178-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECE64886.2024.11024876

ADTNet: Attention-Guided U-Net with Dynamic CNN and Transformers for Skin Cancer Detection

Ashfak Yeafi
Department of Electrical and Electronic Engineering,
Khulna University of Engineering & Technology.
E-mail: yeafiashfak@gmail.com

Liton Sarker
Cash Transfer Modernization (CTM) Project,
Department of Social Services, Ministry of Social Welfare.
E-mail: sarker.liton@gmail.com

***Abstract*—Skin cancer is one of the most common malignancies worldwide, with rising incidence rates emphasizing the need for effective diagnostic and treatment strategies. This study presents the Attention-Guided U-Net, which incorporates Dynamic Convolution and Transformers (ADTNet), a novel deep-learning architecture specifically designed for skin cancer segmentation. ADTNet leverages the strengths of dynamic convolution and transformers within a U-Net framework, enhancing the model's ability to accurately detect cancerous regions in skin images. A key innovation of ADTNet is its integration of attention-guided mechanisms in the skip connections and transformer modules located at the bottleneck of the architecture. Dynamic convolution improves feature extraction by adjusting to the spatial properties of input images, facilitating more detailed and accurate segmentation of skin lesions. Through comprehensive experimentation, we developed a four-stage version of ADTNet, which achieves an optimal balance between segmentation accuracy and computational efficiency. This version yielded a Dice score of 92.4% and an IoU of 87.2%, demonstrating robust performance in distinguishing cancerous regions from healthy tissue. These metrics indicate the model's potential for practical implementation in real-world clinical scenarios, where accurate and efficient segmentation is necessary for effective diagnosis.**

***Index Terms*—Skin Cancer Segmentation, Dynamic Convolution, Attention Mechanism, Transformer**

I. Introduction

Skin cancer, one of the most prevalent malignancies globally, arises from uncontrolled growth of malignant cells in the epidermis. Recent statistics from the American Cancer Society reveal that over 96,480 new melanoma cases are diagnosed annually in the U.S., with approximately 7,230 fatalities. This data indicates an increase in the incidence of melanoma compared to previous years. [1]. Skin cancer is broadly categorized into melanoma and non-melanoma types, with melanoma being the deadliest, originating from abnormal melanocyte proliferation. Early detection of melanoma markedly enhances mortality rates, with a five-year overall survival rate of 98% for localized cases, dropping to 14% in advanced stages. Dermoscopy, a non-invasive method, enhances visualization of pigmented skin lesions, aiding in better diagnosis than traditional visual inspection [2]. However, dermoscopic image interpretation is subjective, relying heavily on the expertise of dermatologists, and prone to errors [3]. Automated dermoscopy analysis has emerged as a promising approach to reduce human error and improve diagnostic accuracy. However, automatic detection of skin lesions continues to be difficult due to the wide variation in skin tones, lesion textures, and the interference caused by artifacts like hair and shadows. [4]. Early segmentation methods, such as threshold-based algorithms, often fail to provide accurate results. Recent breakthroughs in deep learning have transformed medical image analysis, with designs such as U-Net emerging as the standard for medical imagery segmentation due to their skip connections that maintain spatial information [5]. However, traditional CNNs are limited by their static filters. Dynamic Convolutional Neural Networks were introduced to address this, dynamically adjusting filters based on input features, leading to improved feature extraction [6]. Transformers have further advanced vision tasks, with Vision Transformers (ViTs) leveraging attention processes to preserve long-range dependencies by segmenting images into patches [6]. This study proposes an innovative framework that combines dynamic convolution, transformers in the U-Net bottleneck, and attention mechanisms in skip links to improve skin cancer segmentation accuracy. Dynamic convolution enhances adaptability, ViTs capture long-range relationships, and the attention mechanism focuses on critical features during segmentation. This paper emphasizes the following key points:

1) We introduce dynamic convolution into the segmentation framework, allowing convolutional filters to adapt based on input features.
2) We embed transformers in the bottleneck of the U-Net architecture, enabling the model to maintain global context and capture long-range dependencies, enhancing segmentation accuracy.

The structure of the material is as follows: Section Section II examines recent progress in skin cancer segmentation. Section Section III delineates the materials and methodologies employed. Section Section IV delineates the experiments and analyzes the results. Ultimately, Section V conclude the final observations.

II. Related work

Advancements in deep learning, especially in computer vision, have led to substantial progress in skin lesion segmentation. One of the foundational works in this domain is UNet [7], which introduced skip connections, allowing feature information from the encoder to be directly utilized by the decoder. These skip connections mitigate information loss

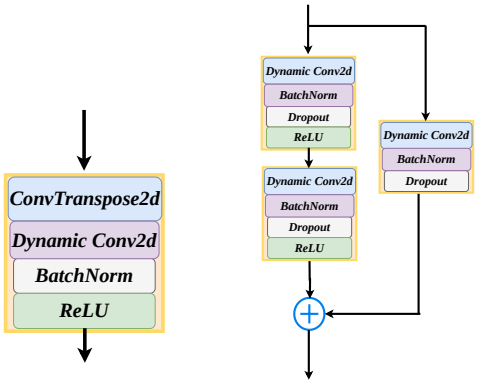
during downsampling, leading to improved segmentation outcomes. However, different medical imaging tasks often require specialized networks to enhance feature learning for specific datasets. Consequently, several extensions and variations of the UNet architecture have been proposed, including UNet++ [8], UNet3+ [9], R2U-Net [10], and CE-Net [11]. While these models show promising results in various tasks, they share a common limitation—the reliance on convolutional operations. Convolution, by nature, struggles to capture global contextual information, which limits its generalization capability in more complex tasks. Researchers have adopted transformer-based topologies to overcome these intrinsic restrictions. Initially developed for natural language processing purposes, transformers have also demonstrated notable efficiency in visual tasks. The ViT [6] was an innovative study that utilized the transformer architecture for image classification by segmenting pictures into patches and processing them in sequence. This innovation inspired further developments in image segmentation. For instance, Valanarasu et al. [12] proposed a medical image detection method leveraging gated axial attention combined with transformers. Chen et al. [13] introduced TransUNet, a hybrid model that integrates the advantages of U-Net and transformer designs, yielding significant outcomes in multi-organ and cardiac segmentation tasks. In summary, while traditional convolution-based models have demonstrated remarkable performance in skin lesion segmentation, their limitations in global feature representation have prompted the adoption of transformer-based models. These advancements not only address the shortcomings of convolutional networks but also provide a robust framework for tackling complex medical image segmentation challenges.

III. Material and Methods

This section introduces our proposed model, ADTNet, specifically developed for skin cancer segmentation. The architecture of ADTNet is built upon a U-Net framework, featuring a U-shaped backbone that comprises both an encoder and a decoder. Within this architecture, we utilize the Residual Dynamic Convolution (RDC) module as a core component for feature extraction. To enhance the model's performance, an attention-guided layer is integrated into the skip connections of ADTNet, while a transformer module is incorporated within the bottleneck of the U-Net structure. The comprehensive design of ADTNet is depicted in Fig. 5.

A. RDC Module

The Residual Dynamic Convolution (RDC) module is a critical component of our proposed ADTNet architecture, aimed at improving the feature extraction abilities essential for skin cancer segmentation. This module employs a residual link to ensure the seamless transmission of information, thereby mitigating the vanishing gradient problem frequently observed in deep neural networks [14]. At the core of the RDC module is the dynamic convolution, which allows the convolutional filters to adapt based on the input features [15]. The construction of the dynamic convolution is depicted in Fig. 2. First,



(a) UpConv Module (b) RDC module

Fig. 1. Architectures of UpConv and RDC Modules. (a) Structure of the UpConv module utilized in the ADTNet architecture, showcasing the upsampling process to restore spatial resolution. (b) Detailed architecture of the RDC module.

a 2D Adaptive Average Pooling layer compresses the spatial information, making the subsequent processing more robust to variations in input size. Next, two 2D Convolution layers are applied, with the first convolutional layer paired with a ReLU activation function used to integrate non-linearity within the model. The second convolutional layer is succeeded by Softmax activation methods, producing normalized attention weights for the convolution kernels. This approach enables the model to focus on the most relevant aspects of the data, thereby enhancing the effectiveness of the dynamic convolution.

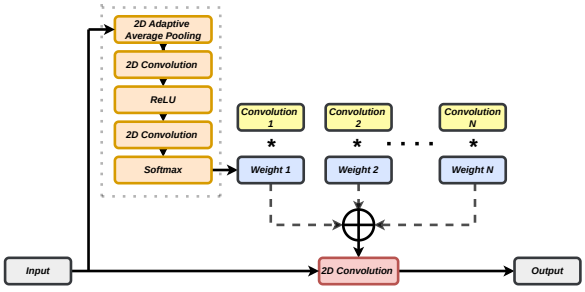


Fig. 2. Illustration of the Dynamic Convolution structure, featuring adaptive average pooling, convolutional layers, and attention weight generation.

The structure of the RDC module is illustrated in Fig. 1(b). The RDC module handles input data via two consecutive blocks. Each block contains several essential components: the initial block includes dynamic convolution, batch normalization, and dropout layers. After the dynamic convolution, Batch Normalization is implemented to stabilize the learning process. This technique normalizes activations, hence accelerating convergence and enhancing the model's overall robustness. A Dropout layer is implemented after the batch normalization process to mitigate overfitting. We incorporate a ReLU activation function following the dropout layer to bring further non-linearity into the model. The skip connection layer of the RDC module includes dynamic convolution,

batch normalization, and dropout layers, in addition to the sequential blocks. The outputs from the skip connection layer and the sequential blocks are concatenated, enhancing the representation of features.

B. Attention Module

In our proposed ADTNet architecture, we incorporate an attention mechanism to enhance the functionality of the skip connections. This process efficiently generates a query accompanied by a collection of key-value pairs, yielding an output calculated as a weighted sum of the values. The weights are determined through the attention mechanism, which establishes a compatibility metric between the query and its associated keys. The implementation of the attention-based skip connection is

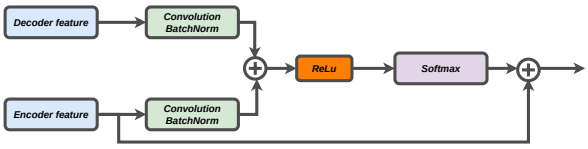


Fig. 3. Visualization of the attention-based skip connection mechanism, demonstrating the integration of encoder and decoder features.

illustrated in Fig. 3. In our network architecture, inputs from both the encoder and decoder sides are directly fed into the attention module. The operational principle of the attention layers can be described mathematically. The encoder generates H hidden state vectors, each with a dimension of α. The input shape for the feedforward layer then becomes $(H, 2\alpha)$. Upon adding a bias term σ, this input P is multiplied by a weight matrix W of shape $(2\alpha, 1)$. The resulting score K is generated, yielding an output with a dimension of $(H, 1)$ as represented in Eq. 1.

$$K = P[H * 2\alpha] * W[2\alpha * 1] + \sigma[H * 1] \tag{1}$$

$$Q = softmax(tanh(K)) \tag{2}$$

$$Output = Input\ term * Q \tag{3}$$

Following this, the score S is processed through a hyperbolic tangent function, followed by a $softmax$ activation function. The output of this process, denoted as Q, is subsequently multiplied by one of the input terms. In this approach, skip links are used to establish connections between the encoder and decoder. The attention mechanism assigns weights to feature maps at each level based on their importance, enabling the model to focus on the most relevant features during the segmentation process.

C. Transformer Module

Our proposed ADTNet architecture incorporates a transformer module at the bottleneck to understand long-distance dependencies and gather global contextual information. The input feature maps are split into patches and processed through attention mechanisms based on the ViT framework. The main block of this transformer module is the Multi-Head Self-Attention (MHSA) mechanism. First, the input feature map

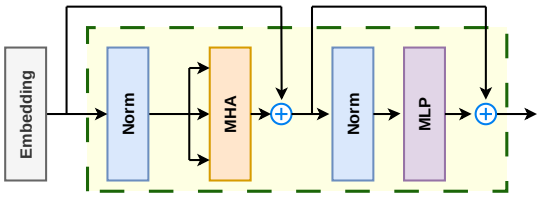


Fig. 4. Illustration of the transformer architecture integrated within the ADTNet, highlighting the self-attention mechanism and positional encoding.

X is projected into query Q, key K, and value V matrices and the attention score is calculated as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where d_k denotes the dimension of the key vectors. The MHSA output is generated by concatenating the outputs from each attention head and applying a linear transformation:

$$\text{MHSA}(Q,K,V) = \text{Concat}(\text{H}_1,\ldots,\text{H}_n)P_O$$

where P_O is the output projection matrix, and n is the number of attention heads. This output is passed through a feedforward neural network (MLP) comprising two fully connected layers with ReLU activation:

$$\text{MLP}(X) = \text{ReLU}(XW_1 + b_1)W_2 + b_2$$

where W_1, W_2 represent the weight matrices, and b_1, b_2 are the bias terms. Both MHSA and MLP blocks are followed by layer normalization:

$$\text{Norm}(X) = \frac{X - mean}{standard \quad deviation}$$

Additionally, residual connections are applied to the outputs of both the MHSA and MLP blocks, leading to the final transformer module output:

$$\text{Output} = X + \text{MLP}(\text{MHSA}(X))$$

D. Proposed ADTNet

This paper introduces ADTNet for segmenting skin lesions. ADTNet extends the U-Net framework by incorporating dynamic convolutions, transformers, and attention mechanisms, enhancing segmentation accuracy and feature extraction capabilities. ADTNet's design adheres to a conventional U-shaped encoder-decoder framework. The encoder path consists of four stages, each stage comprising an RDC module, which integrates two dynamic convolutional layers, batch normalization, ReLU activation, and dropout layers. After each block, a max-pooling layer is applied to reduce feature map dimensions while preserving essential information. In total, the encoder features 8 convolutional layers across its four stages. We incorporate a transformer module with an MHSA mechanism at the architecture bottleneck, therefore augmenting the model's capacity to capture long-range dependencies and global context. This module comprises multi-head attention and feedforward layers, succeeded by normalization layers. The decoder

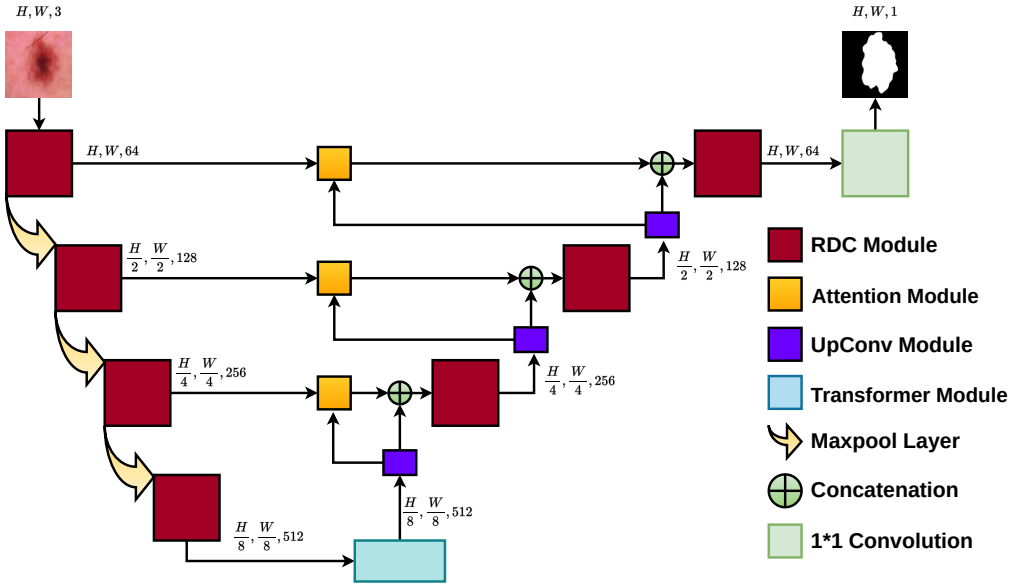


Fig. 5. Overview of the ADTNet architecture, highlighting the integration of RDC modules for adaptive feature extraction, attention-guided skip connections to focus on critical features, and transformers for capturing long-range dependencies. The variables H and W represent the height and width of the feature maps, respectively.

path mirrors the encoder and contains four UpConvolution (UpConv) modules, which consist of transpose convolutional layers to upsample the feature maps, restoring them to their original resolution. The configuration of the UpConv module is depicted in Fig. 1(a). Skip links interconnect the relevant encoder and decoder layers, augmented by an attention module to emphasize the most significant elements of the encoder. The attention mechanism highlights important areas in the feature maps, allowing ADTNet to focus on the most critical areas for segmentation. Finally, a 1×1 convolution layer is applied to generate the output segmentation map, where each pixel is classified as either cancerous or non-cancerous. The design of ADTNet is illustrated in Fig. 5, where the RDC module, attention module, and transformer module are organized to maximize feature extraction and segmentation performance.

IV. RESULTS AND DISCUSSION

A. Dataset description

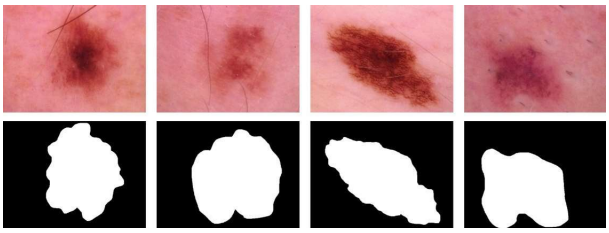


Fig. 6. Visualization of Data Samples from the ISIC 2018 Dataset top row presents skin cancer images, while the bottom row shows the corresponding ground truth masks.

In this study, we utilized the International Skin Imaging Collaboration (ISIC) 2018 dataset, a widely used benchmark for skin cancer segmentation tasks [16]. The dataset comprises 2,594 dermoscopic RGB images, each accompanied by its corresponding ground truth segmentation mask. The average resolution of the images is 2166×3188 pixels, providing high-quality data for accurate segmentation. To ensure a comprehensive assessment, we adhered to the established approach and partitioned the dataset into training, validation, and test subsets. A total of 1,815 images were utilized for training, 259 for validation, and 520 for testing. To enhance computational performance, all photos were downsized to 128×128 pixels. No data augmentation techniques were applied in this experiment, apart from standard data normalization. Both the images and their corresponding masks were cropped uniformly to maintain consistency. Fig. 6 presents a visual representation of various samples alongside their corresponding ground truth masks from the ISIC 2018 dataset.

B. Evaluation Criteria

In this study, we compute the Dice Coefficient Score (Dice), Intersection over Union (IoU), and Sensitivity (SEN) to assess model performance comprehensively. These metrics are calculated as following equations:

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

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

$$SEN = \frac{TP}{TP + FN}$$

The symbols TP, FN, FP, and TN denote the counts of true positives, false negatives, false positives, and true negatives, respectively. A cancerous pixel region is classified as a true

TABLE I
COMPARISON OF DICE, IOU, AND SEN FOR THE 3-STAGE, 4-STAGE, AND 5-STAGE ADTNET ARCHITECTURES.

Model Stage	Dice (%)	IoU (%)	SEN(%)
3-stage	89.4	82.3	90.2
4-stage	92.4	87.2	93.4
5-stage	92.9	87.9	92.4

TABLE II
PARAMETER COUNT FOR THE 3-STAGE, 4-STAGE, AND 5-STAGE ADTNET ARCHITECTURES.

Model Stage	Parameters (M)
3-stage	15.9
4-stage	60.4
5-stage	234.6

positive when the model accurately identifies it as cancerous; conversely, it is labeled as a false positive if the model incorrectly identifies it as cancerous. Similarly, a normal skin pixel region is labeled as a true negative when correctly detected as normal, but if misclassified, it is considered a false negative.

C. Experimental setup

The experiments utilized an NVIDIA Tesla P100 GPU with 16GB of RAM. The ADTNet model was implemented using the PyTorch framework in Python, and training was performed over 30 epochs with a batch size of 8. The Adam optimizer was utilized with a learning rate of 0.001, while all other hyper-parameters were maintained at their default values to ensure uniformity. This configuration enabled efficient convergence and resulted in improved accuracy for segmenting skin cancer regions. For the segmentation tasks, the primary loss function employed was Dice loss [17].

D. Result analysis

In this work, we thoroughly investigated different configurations of our proposed ADTNet architecture for skin cancer segmentation. Specifically, the model was evaluated with 3, 4, and 5 stages, examining its performance using standard metrics such as the Dice coefficient, IoU, and SEN. The outcomes of these tests are encapsulated in Table I. The 3-stage ADTNet architecture achieved a Dice score of 89.4%, IoU of 82.3%, and a sensitivity of 90.2%. When the architecture was expanded to 4 stages, performance improved, yielding a Dice score of 92.4% and IoU of 87.2%, with sensitivity better at 93.4%. The 5-stage ADTNet exhibited marginal improvements over the 4-stage version, achieving Dice, IoU, and SEN scores of 92.9%, 87.9%, and 92.4%, respectively. The results indicate that adding stages from 3 to 4 substantially enhanced the segmentation accuracy, as the 4-stage model was able to capture more complex patterns and features from the dermoscopic images. However, the performance improvement from 4 to 5 stages was minimal, suggesting that increasing the model depth beyond four stages led to diminishing returns. In terms of computational complexity, the number of parameters for each model architecture is provided in Table II. The 3-stage model consisted of 15.9 million(M) parameters, while the 4-stage and 5-stage models had 60.4M and 234.6M parameters, respectively. Although the 5-stage model achieved a slightly higher score, the increase in computational cost was significant, making the 4-stage model a more optimal solution. Therefore, we selected the four-stage ADTNet as the best configuration for our proposed approach, striking a balance between segmentation accuracy and computational efficiency. The performance of the ADTNet model was monitored through the training and validation loss curves, as shown in Fig. 7. The steady decrease in both losses throughout 30 epochs demonstrates that the model successfully learned the intricate patterns within the dataset without overfitting. After 30 epochs, the training loss converged to 0.074, while the validation loss stabilized at 0.068. This consistency between training and validation losses serves as a robust sign of the model's ability to generalize. Fig. 8 presents qualitative results from the test set, showcasing several examples of

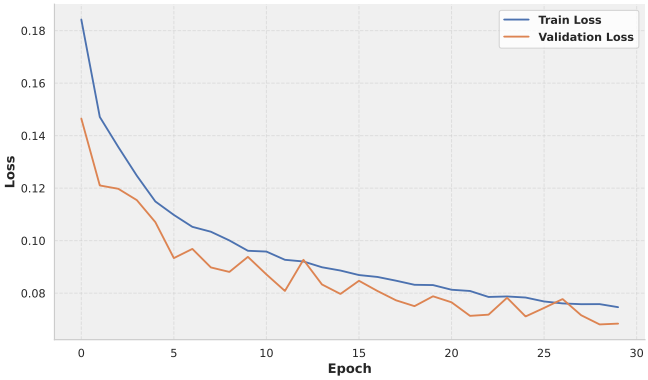


Fig. 7. Loss curves depicting the training and validation performance of the ADTNet model over 30 epochs.

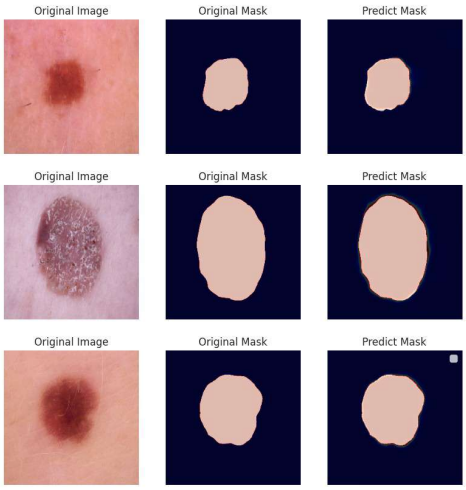


Fig. 8. Examples of model predictions compared with ground truth annotations from the test dataset.

TABLE III
COMPARISON OF ADTNET PERFORMANCE WITH STATE-OF-THE-ART MODELS ON THE ISIC 2018 DATASET.

Methods	Dice (%)	IoU (%)	SEN (%)
UNet [18]	88.5	81.6	88.4
AttUNet [19]	88.6	81.7	88.9
CPFNet [20]	89.7	83.0	90.0
GFANNet [21]	90.1	83.5	90.0
Proposed ADTNet	**92.4**	**87.2**	**93.4**

predicted segmentation masks alongside their corresponding ground truth annotations. These visualizations highlight the accuracy of our model in detecting and segmenting skin cancer regions, even in challenging cases where the lesions may vary in size, shape, and texture. Finally, we compared our ADTNet model with several state-of-the-art architectures on the ISIC 2018 dataset. The comparative results, detailed in Table III, indicate that our proposed ADTNet outperformed existing models in terms of key metrics such as Dice, IoU, and SEN. The superior performance of our model underscores the effectiveness of incorporating dynamic convolutions and transformer-based attention mechanisms in the segmentation pipeline.

V. CONCLUSION

This study presents ADTNet, a deep learning architecture for skin cancer segmentation, utilizing dynamic CNNs and transformers within a U-Net framework. The key innovation lies in integrating attention-guided mechanisms in the skip connections and transformer modules in the U-Net bottleneck, alongside dynamic CNNs for feature extraction. Extensive experiments were conducted using the ISIC 2018 dataset, where 3, 4, and 5-stage variations of ADTNet were evaluated. The 4-stage ADTNet achieved the best balance between segmentation accuracy and computational efficiency, yielding a Dice score of 92.4% and an IoU of 87.2%. It outperformed both the 3-stage and 5-stage models, with significantly reduced complexity compared to the 5-stage variant. Moreover, our approach demonstrated competitive results compared to state-of-the-art models. These findings underscore the potential of ADTNet for accurate and efficient skin cancer segmentation, making it a promising tool for clinical applications in dermatology. Future work will focus on expanding the model's generalization to other datasets, incorporating additional image pre-processing techniques, and exploring lightweight architectures for real-time applications in clinical settings. Despite the promising results, further evaluation on more diverse and challenging datasets is required to ensure its adaptability across different skin cancer types and imaging conditions.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," *CA: a cancer journal for clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[2] N. S. Zghal and N. Derbel, "Melanoma skin cancer detection based on image processing," *Current Medical Imaging*, vol. 16, no. 1, pp. 50–58, 2020.

[3] K. Polat and K. O. Koc, "Detection of skin diseases from dermoscopy image using the combination of convolutional neural network and one-versus-all," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 80–97, 2020.

[4] N. K. Mishra and M. E. Celebi, "An overview of melanoma detection in dermoscopy images using image processing and machine learning," *arXiv preprint arXiv:1601.07843*, 2016.

[5] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[6] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[9] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[10] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *arXiv preprint arXiv:1802.06955*, 2018.

[11] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, no. 10, pp. 2281–2292, 2019.

[12] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in *Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, proceedings, part I 24*. Springer, 2021, pp. 36–46.

[13] 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," *arXiv preprint arXiv:2102.04306*, 2021.

[14] A. Yeafi, M. Islam, S. K. Mondal, K. I. H. Nashad, and M. S. U. Yusuf, "A semi-supervised approach for brain tumor classification using wasserstein generative adversarial network with gradient penalty," in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE, 2023, pp. 1–6.

[15] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[16] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[17] M. T. Jawad, A. Yeafi, and K. K. Halder, "Gsnet: a multi-class 3d attention-based hybrid glioma segmentation network," *Optics Express*, vol. 31, no. 24, pp. 40 881–40 906, 2023.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[19] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197–207, 2019.

[20] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, and X. Chen, "Cpfnet: Context pyramid fusion network for medical image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 10, pp. 3008–3018, 2020.

[21] S. Qiu, C. Li, Y. Feng, S. Zuo, H. Liang, and A. Xu, "Gfanet: Gated fusion attention network for skin lesion segmentation," *Computers in Biology and Medicine*, vol. 155, p. 106462, 2023.