Swin-DSA-Net: A Dual-Stage Attention Guided Swin Transformer Network with Multi-scale Refinement for Skin Lesion Segmentation

Nitin Yadav
Dept. of EEE
IIT Guwahati
nitin.yadav@iitg.ac.in

Bibek Goswami
Dept. of EEE
IIT Guwahati
b.goswami@iitg.ac.in

Marjana Bhuyan
Dept. of CSE
VIT Vellore
marjana.bhuyan2023@vitstudent.ac.in

M.K. Bhuyan
Dept. of EEE
IIT Guwahati
mkb@iitg.ac.in

***Abstract*—Skin cancer is one of the most prevalent cancers in the world, and thus early and precise detection is critical for effective treatment. Skin lesion segmentation is an important aspect of computer-aided diagnostic tools for the precise segmentation of lesion regions from surrounding healthy tissue. However, due to large variations in lesion shapes, colors, and textures, segmentation is a complex task. Recent developments in deep learning models have shown promising results in medical image segmentation tasks. In this work, we introduce a new hybrid segmentation network, Swin-DSA-Net, which incorporates a Swin Transformer encoder with attention-driven and multi-scale refinement mechanisms. This network utilizes a Dual-Stage Attention (DSA) module that refines the features of skip connections with sequential channel and spatial attention and an Adaptive Dilated Multi-scale Refinement (ADM) module that collects contextual information from multiple branches of dilated convolution. The network has shown promising results with a Jaccard Index value of 84.05% and Dice coefficient of 90.10% on the ISIC 2018 datasets.**

***Index Terms*—Swin, Transformer, Segmentation, Skin Cancer.**

I. Introduction

Skin cancer is one of the most common cancers in the world, and it is caused by the uncontrolled growth of abnormal cells in the skin. According to global health reports, it is a major public health problem, with over 330,000 cases of melanoma and almost 60,000 deaths related to it in 2022 [1]. In addition to melanoma, there are other non-melanoma cancers of the skin, including squamous cell carcinoma (SCC) and basal cell carcinoma (BCC), with millions of cases reported worldwide [2]. The traditional method of clinical diagnosis is based on visual observation and dermoscopic examination. However, it is a tedious and observer-dependent task. To overcome the limitations of traditional diagnosis, computer-aided diagnosis (CAD) techniques have been developed.

For effective computer-aided diagnosis, accurate lesion segmentation is a critical factor. Although deep learning methods have achieved great success in image processing, accurate segmentation of lesions is a challenging task. To address these limitations, we introduced Swin-DSA-Net, a transformer-based segmentation framework. Swin-DSA-Net utilizes a Swin Transformer encoder to hierarchically extract features, which can effectively address local and global dependencies. Furthermore, a Dual-Stage Attention (DSA) module is utilized in skip connections to further enhance feature representation using sequential channel and spatial attention, enabling better discrimination of lesion regions. Finally, an Adaptive Dilated Multi-scale Refinement (ADM) module is incorporated in the decoder to effectively fuse multi-scale contextual information using adaptively weighted dilated convolution.

II. Related Work

Skin lesion segmentation has been widely studied in medical image analysis. Initial approaches relied on image processing techniques such as region-based methods, clustering, thresholding, and color-based approaches. However, these methods are sensitive to illumination variations, irregular boundaries, and complex textures, limiting their robustness. With the emergence of deep learning, convolutional neural networks (CNNs) have significantly improved segmentation performance. U-Net [3] introduced an encoder–decoder architecture with skip connections to capture both spatial and contextual features. Subsequent variants like Attention U-Net [4] employs attention gates to focus on relevant regions, while ResUNet-a [5] improves feature propagation through residual connections, and other extensions [6] incorporate atrous convolutions and task interaction strategies to enhance performance.

More recently, transformer-based models have gained attention for their ability to capture long-range dependencies. Hybrid architectures such as CSAP-UNet [7], TransUNet [8], and HTC-Net [9], combine CNNs with transformers to model both local and global context. Fully transformer-based approaches have also been explored. Methods such as Pyramid Transformer Network [10], Spatial Pyramid Transformer [11], and DA-TransUNet [12] enhance contextual modeling and feature representation. Other frameworks, including I^2U-Net [13], MSCA-Net [14], and DSU-Net [15], DSNET [16], focus on multi-scale aggregation and boundary refinement for improved segmentation accuracy. Similarly, ASP-VMUNet [17] leverages Vision Mamba with atrous scanning to improve receptive fields while maintaining efficiency.

Despite these developments, challenges persist in effectively combining multi-scale context information and spatial details. To mitigate these challenges, this paper proposes Swin-DSA-Net, a segmentation framework that utilises a Swin Transformer encoder for hierarchical feature learning, a Dual-Stage Attention (DSA) module for feature refinement, and the Adaptive Dilated Multi-scale Refinement (ADM) module for combining multi-scale context information using dilated convolutions.

III. Material & Methods

A. Datasets

In this work, the ISIC 2018 [18] dataset is used, which is a popular benchmark dataset for skin lesion segmentation provided by the International Skin Imaging Collaboration (ISIC). The ISIC 2018 consists of a larger dataset with 2,594 training images, 100 validation images, and 1,000 test images with corresponding segmentation masks.

B. Methodology

The proposed segmentation framework follows an encoder-decoder architecture like the U-Net [3] model, introducing two key modules to enhance feature selection and multi-scale contextual representation. Despite the success of transformer-based segmentation models, two key limitations remain: (i) skip connections directly transfer encoder features without assessing their relevance, introducing redundant background information, and (ii) fixed receptive fields in convolution layers limit the ability to capture lesions with varying scales and complex boundaries. To address these challenges, the following modules are introduced:

Dual-Stage Attention (DSA) Module: It refines skip connection features using sequential channel and spatial attention.

Adaptive Dilated Multi-scale Refinement (ADM) Module: It captures multi-scale contextual information using adaptively weighted dilated convolutions.

As shown in Fig. 1, a Swin Transformer encoder extracts hierarchical features, which are refined via DSA modules before fusion with decoder features. The decoder progressively reconstructs the segmentation mask using ADM modules and transposed convolutions.

Swin Transformer Encoder

The encoder is built upon the Swin Transformer [19], which enables the extraction of rich hierarchical feature representations. Unlike conventional CNNs, it effectively captures long-range dependencies through a window-based self-attention mechanism, leading to improved contextual understanding. The Swin Transformer is organized into four hierarchical stages, where each stage comprises a predefined number of Swin Transformer blocks. In the base configuration, the number of blocks in each stage follows a $(2, 2, 8, 2)$ arrangement. The detailed architecture of an individual Swin Transformer block is illustrated in Fig. 2.

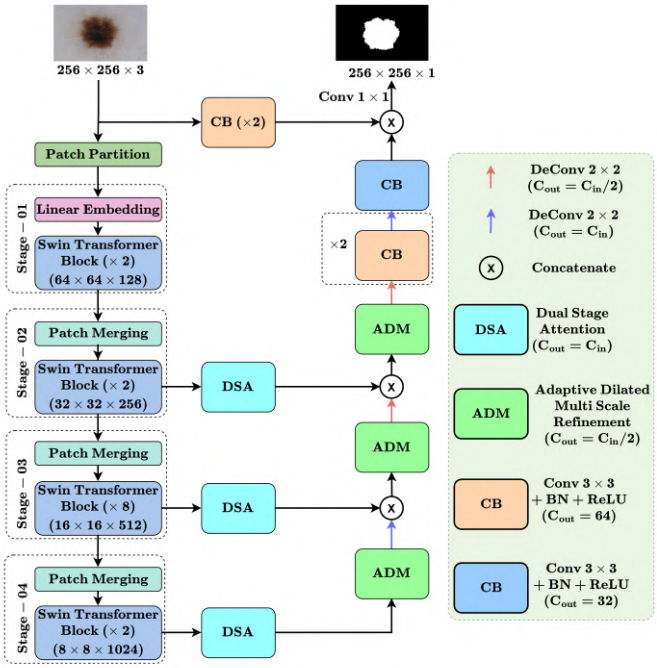

Fig. 1. Overview of the proposed Swin-DSA-Net architecture. It utilises a Swin Transformer encoder for hierarchical feature learning and a decoder with DSA and ADM for feature refinement.

Given an input image $X \in \mathbb{R}^{H\times W\times 3}$, it is partitioned into non-overlapping patches and embedded as:

$$Z_0 = \text{LinearEmbed}(\text{PatchPartition}(X)) \tag{1}$$

The tokens are processed through hierarchical Swin Transformer stages with patch merging, reducing spatial resolution while increasing channel depth. The encoder outputs multi-scale features:

$$F_1 \in \mathbb{R}^{256\times 32\times 32},\ F_2 \in \mathbb{R}^{512\times 16\times 16},\ F_3 \in \mathbb{R}^{1024\times 8\times 8} \tag{2}$$

These features capture both local and global lesion characteristics and are utilized via skip connections in the decoder.

Decoder Architecture

The decoder progressively reconstructs the segmentation mask using transposed convolutions. The deepest feature F_3 is first refined using DSA and ADM modules.

At each decoding stage, upsampled features are fused with corresponding encoder features:

$$F^{(i)}_{\text{fusion}} = \text{Up}(F^{(i+1)}) \parallel F_i \tag{3}$$

The fused features are refined using ADM modules in early stages and lightweight convolution blocks in later stages to recover spatial details.

To enhance fine-grained representation, the final decoder output is concatenated with shallow features extracted from the input:

$$F_{\text{final}} = F_{\text{decoder}} \parallel F_{\text{local}} \tag{4}$$

A 1×1 convolution generates the final segmentation map.

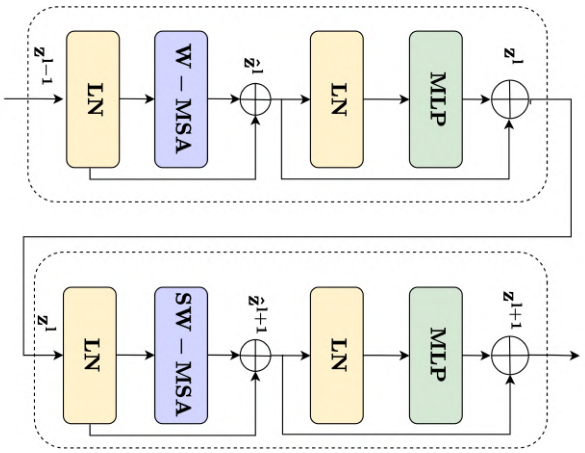

Fig. 2. Structure of the Swin Transformer block, which includes window-based multi-head self-attention (W-MSA), shifted window mechanism for self-attention (SW-MSA), and feed-forward network (MLP) with residual connections and layer normalization.

Dual-Stage Attention (DSA) Module

Skip connections often introduce irrelevant background information. The proposed DSA module refines encoder features using sequential channel and spatial attention, followed by a novel fusion strategy.

Let $F \in \mathbb{R}^{C\times H\times W}$ denote the input feature map, where C, H, and W are the number of channels, height, and width, respectively. Let $\odot$ denote element-wise multiplication (with broadcasting), $\sigma(\cdot)$ the sigmoid function, $\delta(\cdot)$ the ReLU activation, and $\mathrm{GAP}(\cdot)$ global average pooling (producing a $C\times 1\times 1$ descriptor).

Channel Attention: A channel descriptor $z \in \mathbb{R}^{C\times 1\times 1}$ is generated as

$$z = \mathrm{GAP}(F). \tag{5}$$

Channel weights $s_c \in \mathbb{R}^{C\times 1\times 1}$ (reduction ratio $r = 16$) are computed via two successive 1×1 convolutions:

$$s_c = \sigma\big(W_2\big(\delta(W_1(z))\big)\big), \tag{6}$$

where W_1 reduces channels from C to C/r and W_2 expands back to C. The channel-refined feature map is

$$F_c = F \odot s_c. \tag{7}$$

Spatial Attention: From F_c, the average- and max-pooled maps across channels are

$$M_{\mathrm{avg}} = \tfrac{1}{C}\sum_{c=1}^{C} F_c[c,:,:], \quad M_{\max} = \max_c F_c[c,:,:]$$

(both $1\times H\times W$). These are concatenated along the channel dimension to form

$$M = [M_{\mathrm{avg}}; M_{\max}] \in \mathbb{R}^{2\times H\times W}.$$

Spatial weights $s_s \in \mathbb{R}^{1\times H\times W}$ are obtained as

$$s_s = \sigma\big(\mathrm{Conv}_{7\times 7}(M)\big) \tag{8}$$

(7×7 convolution with padding 3). The spatially-refined feature map is

$$F_s = F_c \odot s_s. \tag{9}$$

Feature Fusion:

$$F_{\mathrm{DSA}} = \mathrm{Conv}_{1\times 1}\big([F_c; F_s]\big) + F, \tag{10}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation and the 1×1 convolution reduces the channel count from $2C$ back to C. The residual connection $+F$ preserves the original information. Unlike CBAM [20], DSA fuses the intermediate channel-refined map F_c with the final spatially-refined map F_s before the residual addition, preserving complementary information and improving skip-connection quality.

Adaptive Dilated Multi-scale Refinement (ADM) Module

To capture multi-scale context, the ADM module employs three parallel dilated convolutions with adaptive weighting.

Let $F \in \mathbb{R}^{C\times H\times W}$ be the input feature map. Three parallel branches produce

$$F_i = \mathrm{Conv}_{3\times 3}^{d_i}(F), \quad d_i \in \{1,2,4\}, \quad i = 1,2,3, \tag{11}$$

where each 3×3 convolution (with appropriate padding) preserves the spatial dimensions $H\times W$.

Adaptive Weighting: Global descriptors are first extracted from each branch:

$$g = [\mathrm{GAP}(F_1);\ \mathrm{GAP}(F_2);\ \mathrm{GAP}(F_3)] \in \mathbb{R}^{3C}, \tag{12}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling. A learnable matrix $W \in \mathbb{R}^{3\times 3C}$ then projects g to three logits, and the adaptive weights are obtained via softmax:

$$w = [w_1,\ w_2,\ w_3]^\top = \mathrm{Softmax}(Wg), \quad \sum_{i=1}^{3} w_i = 1. \tag{13}$$

The final output is the weighted sum

$$F_{\mathrm{ADM}} = \sum_{i=1}^{3} w_i F_i, \tag{14}$$

where the scalar weights w_i (computed per input sample) are broadcast across all spatial locations. In addition, batch normalization, ReLU, and dropout (with a dropout rate of 0.1) are adopted for stability. Unlike traditional fixed multi-scale methods such as ASPP in [21], ADM adaptively determines the importance of multi-scale features using global contextual information, enabling the network to emphasize the most relevant receptive field for lesions of varying sizes.

C. Implementation Details

All experiments were conducted using Python 3.12 and PyTorch 2.8 with CUDA 12.6 on an NVIDIA Tesla P100 GPU (16 GB VRAM). Images were resized to 256×256 pixels. To improve generalization, data augmentation techniques, including horizontal/vertical flipping, rotation, brightness/contrast adjustment, elastic transformation, and random cropping, were applied. Training was performed for a maximum of 500 epochs

using the Adam optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, and a batch size of 8. The learning rate was dynamically adjusted using the ReduceLROnPlateau strategy, while early stopping with a patience of 15 epochs was employed to avoid overfitting. To optimize segmentation quality, the network was trained with Dice loss, which directly maximizes the overlap between the predicted masks and the corresponding ground-truth annotations.

IV. Result & Discussion

The performance of the proposed model was extensively evaluated on the publicly available ISIC 2018 dataset using five widely adopted segmentation metrics: Accuracy (AC), Dice Coefficient (DC), Intersection over Union (IoU), Specificity (SP), and Sensitivity (SE). The proposed model has resulted in an IoU of 84.05% and Dice of 90.10%. A quantitative comparison with state-of-the-art (SOTA) techniques is also shown in Table I. It can be noted that the proposed model has outperformed existing techniques in terms of all evaluation parameters. This demonstrates the effectiveness of the proposed architecture in learning both global and local spatial details. The proposed model contains **123.44 M** trainable parameters and requires **40.58 GFLOPs**. The average inference time is **3.48 ms** per image on an NVIDIA Tesla P100 GPU.

TABLE I
Performance Comparison on the ISIC 2018 Dataset

Model	Yr	AC	IoU	Dice	SE	SP
U-Net* [3]	2015	92.25	78.49	86.48	90.42	95.63
Attention U-Net* [4]	2018	92.59	79.49	87.20	91.08	95.55
ResU-Net-a* [5]	2020	92.90	80.37	87.90	92.99	93.46
TransUNet* [8]	2021	92.01	80.44	87.66	90.17	94.74
DA-TransUNet* [12]	2024	93.25	81.39	88.95	91.37	94.76
CSAP-UNet* [7]	2024	93.76	82.11	89.05	93.01	94.78
I^2U-Net* [13]	2024	93.76	80.89	88.31	90.27	94.84
ASP-VMUNet [17]	2025	93.83	80.32	89.09	89.97	95.33
DSU-Net* [15]	2025	92.80	81.97	89.26	91.02	94.73
DSNET [16]	2025	–	81.30	89.81	–	–
Proposed	–	**94.51**	**84.05**	**90.10**	**93.10**	**94.27**

* indicates retrained model.

In addition, qualitative segmentation results are presented in Fig. 3, where the results are compared to ground truth and input images for each case. The proposed model is capable of effectively segmenting lesion boundaries and structural features for different types of skin lesions.

A. Failure Case Analysis

Although PGAF-Net achieves strong overall segmentation performance, some challenging cases remain difficult to segment accurately. Figure 4 presents representative failure examples from the ISIC 2018 dataset. Most failures occur when the lesion and healthy skin share similar color and texture characteristics, making precise delineation difficult. Consequently, the model may underestimate lesion extent, over-segment surrounding regions, or fail to accurately capture irregular boundaries.

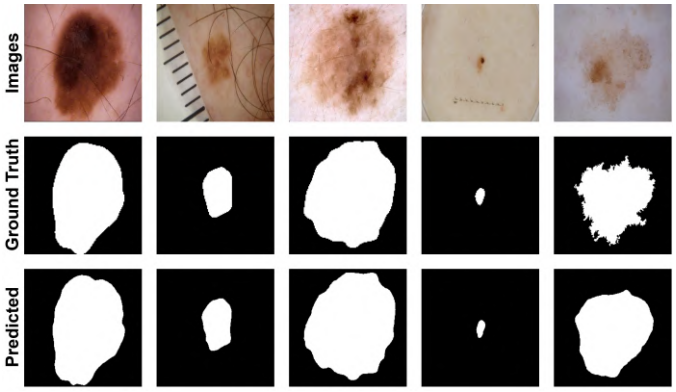

Fig. 3. Comparison of segmentation results on some samples of test images.

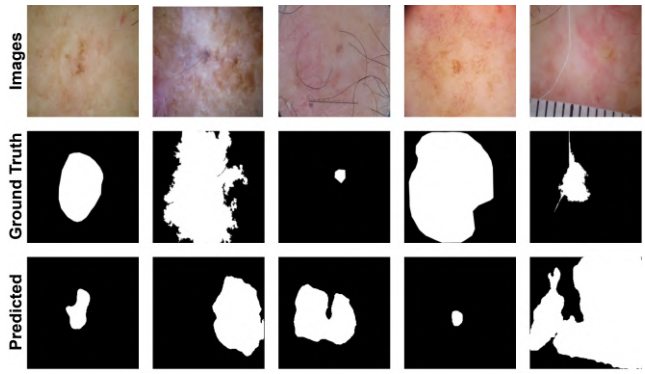

Fig. 4. Representative failure cases of PGAF-Net on the ISIC 2018 test dataset

B. Ablation Study

To comprehensively evaluate the effectiveness of the proposed framework, a series of ablation experiments were conducted on the ISIC 2018 dataset. The results are summarized in Table II.

The first section of Table II evaluates the contribution of DSA and ADM. Compared with the baseline, both modules improve segmentation performance individually, while their combination achieves the best results with a Dice score of 90.10

The second and third sections analyze the design choices of DSA and ADM. For DSA, both CBAM and the variant without feature fusion perform worse than the proposed design, validating the effectiveness of the dual-scale attention fusion strategy. Similarly, the proposed ADM outperforms single-scale convolution, ASPP, and fixed fusion, highlighting the benefit of adaptive multi-scale feature aggregation.

The fourth section compares different loss functions. Dice loss achieves the highest Accuracy, Dice, and IoU scores, indicating its suitability for handling class imbalance and directly optimizing segmentation overlap.

Finally, the backbone comparison shows that larger Swin Transformer variants generally provide better feature representations and improved segmentation accuracy. The selected backbone offers the best balance between performance and

TABLE II
COMPREHENSIVE ABLATION STUDY ON THE ISIC 2018 DATASET.

Category	Method	Dice (%)	IoU (%)
Overall	Baseline (w/o DSA, w/o ADM)	89.70	82.46
	Baseline + DSA	90.01	83.74
	Baseline + ADM	89.77	83.09
	Baseline + DSA + ADM	**90.10**	**84.05**
DSA	Identity	90.01	83.74
	CBAM [20]	88.73	81.50
	DSA w/o Fusion	89.67	82.61
	DSA (Proposed)	**90.10**	**84.05**
ADM	Single Conv	88.67	81.54
	ASPP [21]	89.69	83.35
	ADM w/o Weighting	90.03	83.23
	ADM (Proposed)	**90.10**	**84.05**
Loss	BCE	86.14	83.00
	BCE + Dice	89.66	83.24
	Dice Loss	**90.10**	**84.05**
Backbone	Swin-Tiny	89.06	82.18
	Swin-Small	89.46	82.27
	Swin-Base	**84.05**	**90.10**
	Swin-Large	89.70	82.70

computational complexity and is therefore adopted in the final model.

Overall, the ablation results confirm that the Swin Transformer encoder, together with the proposed DSA and ADM modules, collectively enhances feature representation, multi-scale aggregation, and lesion boundary delineation, leading to improved robustness and segmentation performance.

V. CONCLUSION

In this paper, a novel deep learning framework for skin lesion segmentation is proposed. The proposed framework uses a Swin Transformer-based encoder to extract hierarchical feature representations and long-range dependencies. To further enhance the feature learning process, two important modules have been introduced: the DSA module for the attention-based feature refinement process and the ADM module for the adaptive multi-scale feature aggregation process. The proposed model has been trained and tested on the ISIC 2018 dataset, and the results show that the model has performed better than the existing state-of-the-art techniques. The results show that the proposed architecture can produce a precise and robust segmentation mask, even for complex lesion structures. In conclusion, the proposed framework is a viable solution for the problem of automatic skin lesion segmentation, which can help dermatologists in the diagnosis of the disease in the future. The proposed model can be extended and tested on a large dataset to check its generalization capacity in the future.

REFERENCES

[1] "Skin cancer statistics — World Cancer Research Fund — wcrf.org," https://www.wcrf.org/preventing-cancer/cancer-statistics/skin-cancer-statistics/, [Accessed 09-06-2026].

[2] "Radiation: Ultraviolet (UV) radiation and skin cancer — who.int," https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer, [Accessed 09-06-2026].

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[6] R. Arora, B. Raman, K. Nayyar, and R. Awasthi, "Automated skin lesion segmentation using attention-based deep convolutional neural network," *Biomedical Signal Processing and Control*, vol. 65, p. 102358, 2021.

[7] X. Fan, J. Zhou, X. Jiang, M. Xin, and L. Hou, "Csap-unet: Convolution and self-attention paralleling network for medical image segmentation with edge enhancement," *Computers in Biology and Medicine*, vol. 172, p. 108265, 2024.

[8] 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.

[9] H. Tang, Y. Chen, T. Wang, Y. Zhou, L. Zhao, Q. Gao, M. Du, T. Tan, X. Zhang, and T. Tong, "Htc-net: A hybrid cnn-transformer framework for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 88, p. 105605, 2024.

[10] 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 Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 145–156, 2022.

[11] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Medical Image Analysis*, vol. 77, p. 102357, 2022.

[12] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation," *Frontiers in Bioengineering and Biotechnology*, vol. 12, p. 1398237, 2024.

[13] D. Dai, C. Dong, Q. Yan, Y. Sun, C. Zhang, Z. Li, and S. Xu, "I2u-net: A dual-path u-net with rich information interaction for medical image segmentation," *Medical Image Analysis*, vol. 97, p. 103241, 2024.

[14] Y. Sun, D. Dai, Q. Zhang, Y. Wang, S. Xu, and C. Lian, "Msca-net: Multi-scale contextual attention network for skin lesion segmentation," *Pattern Recognition*, vol. 139, p. 109524, 2023.

[15] L. Zhong, T. Li, M. Cui, S. Cui, H. Wang, and L. Yu, "Dsu-net: Dual-stage u-net based on cnn and transformer for skin lesion segmentation," *Biomedical Signal Processing and Control*, vol. 100, p. 107090, 2025.

[16] Y. Chen, G. Yang, X. Dong, J. Zeng, and C. Qin, "Dsnet: A lightweight segmentation model for segmentation of skin cancer lesion regions," *IEEE Access*, vol. 13, pp. 31 095–31 104, 2025.

[17] M. Bao, S. Lyu, Z. Xu, Q. Zhao, C. Zeng, W. Bai, and G. Cheng, "Asp-vmunet: Atrous shifted parallel vision mamba u-net for skin lesion segmentation," *arXiv preprint arXiv:2503.19427*, 2025.

[18] 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.

[19] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.