Applying and Adapting DUCK-Net for Skin Lesion Segmentation

Thoi-Toan Nguyen Dang
University of Science, VNU-HCM
Vietnam National University Ho Chi Minh City
21120149@student.hcmus.edu.vn

***Abstract*—This study explores the adaptation of DUCK-Net, a compact convolutional network originally designed for polyp segmentation, for skin lesion segmentation on the ISIC-2016 dataset. DUCK-Net employs DUCK blocks that combine residual, dilated, and separable convolutions to enable efficient multiscale representation while maintaining a lightweight structure. We investigate three architectural variants with different depths: DUCK-Net with five encoder–decoder layers, DUCK-Net-Lite with one fewer layer, and DUCK-Net-Tiny with two fewer layers. Experimental results show that DUCK-Net achieves 92.52% Dice and 86.08% IoU, while DUCK-Net-Lite retains 92.00% Dice and 85.19% IoU despite using fewer layers and less computation. In contrast, DUCK-Net-Tiny exhibits notable performance degradation, highlighting the importance of deeper semantic features. Additionally, ablation studies confirm that replacing DUCK blocks with standard convolutional blocks leads to consistent drops in performance, demonstrating the effectiveness of the DUCK block design. These findings suggest that DUCK-Net-Lite offers an optimal balance between accuracy, efficiency, and model size, making it well suited for deployment in real-time or resource-constrained environments.**

***Index Terms*—DUCK-Net, skin lesion segmentation, medical image analysis, lightweight convolutional networks, multiscale feature extraction, ISIC-2016**

I. Introduction

Skin cancer, particularly malignant melanoma, ranks among the most aggressive and fatal cancers when not detected at an early stage [1]. The increasing exposure to ultraviolet (UV) radiation has been identified as a major contributing factor to its rising incidence worldwide. Dermoscopic imaging has emerged as a valuable non-invasive technique for aiding dermatologists in the evaluation of suspicious skin lesions [2]. A crucial component of automated diagnostic systems is accurate lesion segmentation, which facilitates the delineation of lesion boundaries and supports objective assessments of lesion size, shape, and temporal progression [3].

Although recent advances in medical imaging and deep learning have led to promising results, accurate skin lesion segmentation remains a challenging task. Lesions exhibit significant variations in morphology, such as differences in size, shape, color, and texture, and are often affected by imaging artifacts like hair, glare, and shadows [4]. Additionally, low contrast between lesions and healthy skin, along with the limited availability of high-quality annotated datasets, hampers model generalizability [5].

Convolutional neural networks (CNNs) [6] have demonstrated strong performance in medical image segmentation due to their ability to learn spatially hierarchical representations. Architectures such as U-Net [7], Attention U-Net [8], SwinUNet [9], and TransFuse [10] have shown competitive results on various skin lesion datasets. However, these models often rely on pretrained weights, involve complex architectural components, and require substantial computational resources, factors that limit their practical adoption in clinical environments [11].

To mitigate these limitations, we explore the use of DUCK-Net [12], a lightweight and efficient encoder-decoder CNN architecture originally proposed for colorectal polyp segmentation. DUCK-Net is designed to extract multiscale spatial features while maintaining computational simplicity. Its core components include the DUCK block, a parallel module combining residual, dilated, and separable convolutions, and a residual downsampling path that preserves fine-grained spatial details. Notably, DUCK-Net can be trained from scratch on domain-specific datasets without requiring external pretrained weights, making it well-suited for low-resource medical applications.

In this study, we not only adapt the original DUCK-Net architecture [12] to the ISIC-2016 skin lesion segmentation dataset [5], but also introduce a compressed variant (termed *DUCK-Net-Lite*) that removes or reduces specific parallel branches and filter widths to further decrease model complexity. Our hypothesis is twofold: (i) DUCK-Net's architectural characteristics, originally tailored for gastrointestinal image segmentation, can effectively transfer to dermoscopic images; and (ii) a carefully simplified version can retain most of the segmentation accuracy while significantly lowering the number of parameters and computational cost.

Our main contributions are summarized as follows:

- We apply the original DUCK-Net to skin lesion segmentation and introduce a compact variant, *DUCK-Net-Lite*, tailored for resource-constrained settings.
- DUCK-Net achieves strong performance on ISIC-2016 without pretraining, reaching a Dice score of 92.52% and IoU of 86.08% using only 17 filters per layer.
- DUCK-Net-Lite reduces parameters by 44% with minimal accuracy loss, achieving a Dice score of 92.00%.
- We conduct two ablation studies: (1) replacing DUCK blocks with standard convolutional blocks to evaluate the effectiveness of the block design, and (2) reducing the

number of encoder-decoder layers to assess the contribution of architectural depth.
- We provide a comprehensive efficiency evaluation, including parameter count, model size, and inference time, highlighting the models' practicality for clinical deployment.

The remainder of this paper is organized as follows. Section II reviews related work. Section III introduces the DUCK-Net and DUCK-Net-Lite architectures. Section IV describes the experimental setup and evaluation results. Section V presents an ablation study that verifies the effectiveness of the DUCK block compared to standard convolutional blocks. Section VI provides a discussion on model complexity and inference efficiency. Finally, Section VII concludes the paper.

II. Related Work

A. Convolutional Neural Networks

Convolutional neural networks (CNNs) have long been central to medical image segmentation tasks due to their strong capacity for extracting local features. The introduction of U-Net [7], an encoder-decoder architecture with skip connections, significantly improved spatial resolution recovery in segmentation problems. Building upon this foundation, many variants have been proposed to improve performance and close the semantic gap between the encoder and decoder stages. For example, Attention U-Net [8] integrates attention gates to emphasize relevant lesion regions while suppressing background noise. nnU-Net [13] eliminates manual architecture design by automatically adapting to the input dataset's properties. SwinUNet [9] incorporates Swin Transformer blocks into the U-Net backbone to extend the receptive field through hierarchical attention. Other architectures like DAEFormer [14] and SU-Net [15] further enhance the U-shaped design by using dense feature fusion, multi-path encoders, or attention-based refinements. Despite their effectiveness, CNNs are inherently constrained in modeling global context due to the locality of convolution operations [16].

B. Transformer-Based Models

Transformer architectures [17], originally introduced in natural language processing, have been increasingly adopted in vision applications for their ability to model long-range dependencies. Vision Transformer (ViT) [18] applies this concept to image classification by treating patches as sequential tokens. Building on this, TransUNet [19] combines ViT with a U-Net backbone, integrating CNN-based spatial features with transformer-driven global attention. Subsequent models such as TransFuse [10], MISSFormer [16], HiFormer [20], and D-LKA [21] explore different hybridization strategies. TransFuse employs parallel CNN and Transformer branches merged through a prediction head. MISSFormer introduces a lightweight self-attention mechanism for efficient feature refinement, while HiFormer applies multi-scale attention to better handle morphologically diverse lesions. D-LKA enhances the receptive field using large-kernel attention while maintaining low computational complexity. Despite their strong performance, transformer-based architectures typically require pretrained weights and high memory consumption, which can be impractical in clinical environments where computational resources are limited [10].

C. Hybrid and Multi-scale Architectures

To reconcile local feature precision with global context modeling, hybrid models combining CNNs and transformers have emerged. ScaleFusionNet [22] is one such architecture specifically tailored for skin lesion segmentation. It integrates Adaptive Fusion Blocks based on deformable convolutions to extract flexible multi-scale features, while the Cross-Attention Transformer Module (CATM) improves information flow between encoder and decoder. This design enables strong performance on datasets such as ISIC-2016 and ISIC-2018 [23], particularly in terms of Dice coefficient and Intersection over Union (IoU) scores.

D. Lightweight and Compressed Segmentation Networks

Several works have pursued parameter-efficient or real-time medical segmentation models via model pruning, depth/width reduction, or lightweight backbones. Examples include Mobile-UNet [24], ESPNet [25], ENet-like [26] adaptations for segmentation, and pruning/quantization strategies specialized for U-shaped architectures. These approaches highlight the growing demand for deployable models under strict memory and latency constraints. However, most existing efforts either target generic natural images or still rely on pretrained backbones.

E. Research Gap and Motivation

Most state-of-the-art methods rely on complex backbones, pretrained transformers, or heavy compute, which limits deployment in resource-constrained clinics. DUCK-Net [12], though lightweight and trainable from scratch, has not been examined on skin lesion datasets nor explored for further compression. Motivated by this gap, we adapt DUCK-Net to ISIC-2016 and propose a depth-truncated variant (DUCK-Net-Lite). We further assess the impact of block choice and depth reduction on both accuracy and efficiency.

III. Method

A. Overview

Our approach builds on DUCK-Net [12] and targets two goals: (i) adapt it to dermoscopic skin-lesion images, and (ii) compress it for resource-constrained deployment. The baseline architecture retains DUCK-Net's core modules for multi-scale feature extraction and detail preservation. On top of that, we introduce a depth-truncated ("Lite") variant to further reduce parameters and MACs.

B. Key Components

1) DUCK Block (Deep Understanding Convolutional Kernel): A parallel module that aggregates six convolutional variants (e.g., residual, dilated, separable). This allows the network to adaptively choose receptive fields and

capture salient lesion regions. Repeated use, however, may attenuate very fine details in deeper layers.

2) Auxiliary U-Net style Downsampling Path: A residual downsampling branch that performs only strided convolutions (no feature transformation). By bypassing DUCK blocks, it preserves low-level spatial information that might otherwise be compressed in early stages.
3) Depth Truncation (DUCK-Net-Lite): To reduce architectural complexity, we remove the deepest encoder-decoder level and its corresponding decoder stage (early exit).

C. DUCK-Net-Lite Architecture

DUCK-Net-Lite is a compact variant of the original DUCK-Net, specifically designed to reduce computational overhead while preserving segmentation performance. The original DUCK-Net, although lightweight, still contains deep layers and complex blocks that may be excessive for deployment on resource-limited devices. DUCK-Net-Lite addresses this by introducing several architectural simplifications.

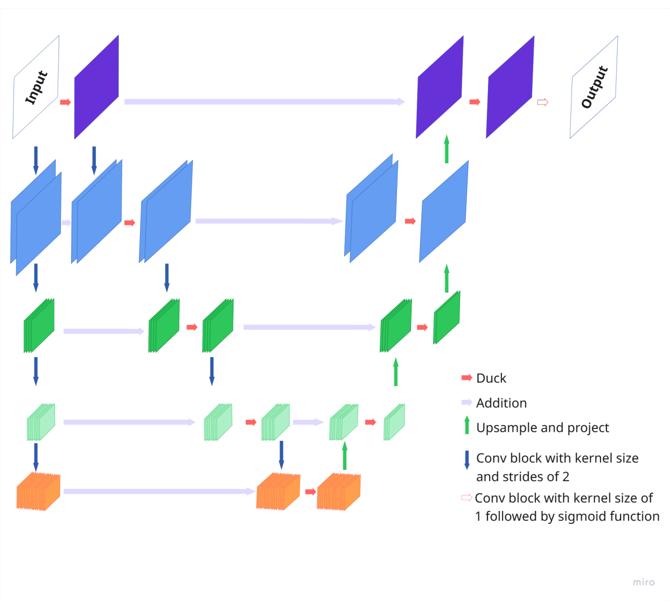
Input
Output
Duck
Addition
Upsample and project
Conv block with kernel size and strides of 2
Conv block with kernel size of 1 followed by sigmoid function
miro

Fig. 1. Overview of DUCK-Net-Lite architecture. The network follows a typical encoder-decoder structure with four downsampling and upsampling stages. Compared to the original DUCK-Net [12], DUCK-Net-Lite removes the bottom-most encoder-decoder stage to reduce computational cost. Complex DUCK and residual blocks are replaced with standard Conv-BN-ReLU layers. After each upsampling step, a 1×1 convolution aligns the feature dimensions to enable skip connections via addition.

The proposed architecture applies the following key modifications:

1) Depth truncation (early exit): The deepest encode-decoder stage (with feature width $F \times 32$) is removed. The model retains four downsampling/upsampling levels, with maximum feature width reduced to $F \times 16$.
2) Block simplification: Heavier DUCK blocks and ResNet-style modules are selectively replaced with plain convolutional blocks (Conv-BN-ReLU), reducing the number of branches and the per-layer cost.
3) Channel projection after upsampling: After each upsampling step, a 1×1 convolution projects the feature map to match the corresponding encoder feature dimensions. This enables efficient skip connections via element-wise addition.

By combining these strategies, DUCK-Net-Lite achieves a favorable balance between accuracy and efficiency, making it suitable for real-time or edge-device applications.

D. Block Components

1) Residual Block:

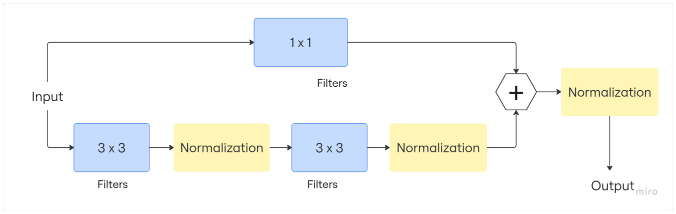

Fig. 2. Residual Block [27]. A standard ResNet-style unit used to maintain gradient flow and improve feature representation for detailed lesion boundaries.

Residual Block (Fig. 2) was first introduced in the ResUNet++ architecture [28], and it serves as a core component of the proposed DUCK-Net model for skin lesion segmentation on the ISIC 2016 dataset. The primary goal of this block is to help the network effectively learn fine-grained and detailed features, such as lesion boundaries and textures, crucial elements for accurately identifying the location and shape of skin lesions.

Stacking multiple small convolutional layers (e.g., 3×3) is a common strategy to increase the receptive field. However, increasing network depth may lead to vanishing gradients and hinder convergence. To address this issue, residual connections are employed, allowing gradients to “shortcut” through intermediate layers, which facilitates stable training even in deeper architectures.

Furthermore, combining one, two, or three residual blocks in parallel can emulate the effect of using larger convolutional kernels such as 5×5, 9×9, and 13×13. This multi-scale approach enables the network to learn both local features and broader contextual information, which is essential when dealing with lesions that vary in size, shape, and contrast. In addition to improving feature reuse, the use of residual blocks enhances the model’s generalization capability across diverse lesion appearances.

2) Midescope Block and Widescope Block:

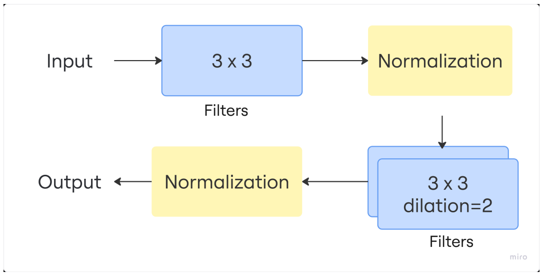

Fig. 3. Midescope Block [12]. A dilated convolutional structure that approximates a 7×7 receptive field to capture mid-level contextual features.

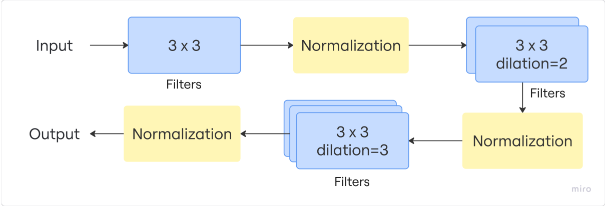

Fig. 4. Widescope Block [12]. Utilizes large dilation rates to simulate a 15×15 kernel for learning global context in large lesion regions.

Two new blocks introduced in the architecture are the Midescope (Fig. 3) and Widescope (Fig. 4) blocks. Both utilize dilated convolution techniques to simulate the effect of larger kernels without significantly increasing the number of parameters, while still preserving the ability to learn high-level abstract features.

Specifically, dilated convolutions distribute the sampling points of a standard 3×3 kernel over a larger region of the input image, thereby expanding the receptive field without resorting to pooling operations or additional layers. However, this method may sacrifice some local detail, which is why these blocks are primarily designed to capture prominent global features rather than fine boundaries.

In the proposed model:

- The Midescope block approximates a 7×7 kernel.
- The Widescope block approximates a 15×15 kernel.

This approach enables the model to incorporate broader contextual awareness, which is particularly helpful when segmenting lesions with unclear or diffuse boundaries.

3) *Separated Block:*

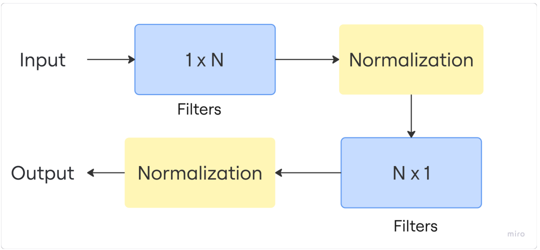

Fig. 5. Separated Block [12]. A separable convolutional design that approximates large kernels using $1 \times N$ and $N \times 1$ layers, trading efficiency for reduced diagonal sensitivity.

The Separated block (Fig. 5) represents a third approach to simulating large convolutional kernels. The core idea is to approximate a traditional $N \times N$ kernel by combining a $1 \times N$ kernel with an $N \times 1$ kernel.

However, this method introduces a limitation related to what is referred to as "diagonality", the ability of a convolutional layer to capture and retain spatial details related to diagonal patterns in images, a property inherent in two-dimensional $N \times N$ kernels. These full kernels naturally capture spatial dependencies in both horizontal and vertical directions while also covering diagonal structures.

In contrast, the Separated block applies separable convolution, where the operations are performed sequentially, first along one dimension (e.g., $1 \times N$) and then the other (e.g., $N \times 1$). This sequential processing restricts the filters to operate on one axis at a time, potentially reducing their capacity to encode diagonal features, leading to what is known as "loss of diagonality."

In many cases, diagonal relationships are crucial for identifying complex patterns and shapes. Therefore, other blocks in the architecture, such as Residual, Midescope, and Widescope are designed to compensate for this limitation.

4) *DUCK Block:*

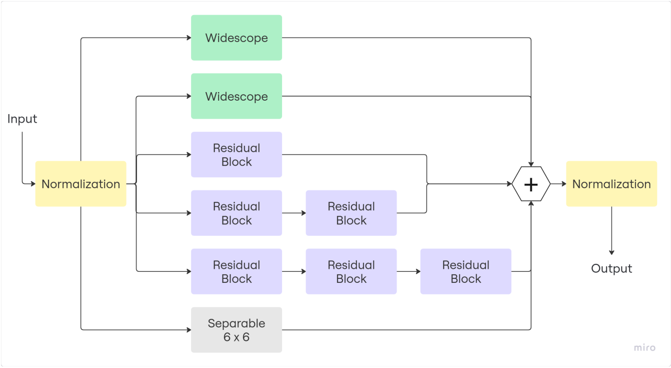

Fig. 6. DUCK Block [12]. A composite convolutional module that integrates Residual, Midescope, Widescope, and Separated blocks in parallel to simulate large kernel behavior. This design enables simultaneous learning of both fine details and global context. Empirically, the Residual path is implemented using a stack of three residual blocks for optimal performance-efficiency tradeoff.

The DUCK block (Fig. 6) is a newly proposed convolutional block that integrates the previously discussed components (Residual, Midescope, Widescope, and Separated) into a parallel structure. This design allows the network to flexibly choose the most suitable processing behavior at each learning step.

The core idea of the DUCK block is to leverage diverse approaches to simulate large kernel effects, enabling the model to automatically balance the strengths and weaknesses of each technique. By using a variety of kernel types, the model can both localize the general lesion region and capture fine details such as boundaries and shapes.

In its implementation, the DUCK block employs a stack of three Residual blocks. This choice is empirically driven by the observation that duplicating Midescope, Widescope, or

Separated blocks does not yield significant performance gains, while substantially increasing computational costs.

In summary, the DUCK block is a versatile convolutional module capable of learning both low-level and high-level features simultaneously, contributing to the model's ability to adapt to various forms of skin lesions in the segmentation task.

E. Model Evaluation

In the task of skin lesion segmentation, each input image is paired with a corresponding binary mask of the same size. Each pixel in the mask is assigned a label of 0 or 1, depending on whether it belongs to the lesion region. A pixel is labeled as 1 if it falls within the lesion area marked by experts, and 0 if it corresponds to normal skin.

This binary representation simplifies the segmentation task into a pixel-level binary classification problem, in which the model learns to distinguish lesion areas from the background on a per-pixel basis.

Accurate evaluation is essential for assessing the effectiveness of different neural network architectures. Various metrics have been proposed for this purpose. In this study, we focus on five of the most commonly used evaluation metrics: Dice Coefficient, Jaccard Index, Precision, Recall, and Accuracy.

Dice Loss is a widely used loss function in medical image segmentation. It is derived from the Dice Coefficient, which quantifies the overlap between the predicted and ground truth segmentation masks.

IV. Experiments

A. Implementation Details

We implemented the DUCK-Net and DUCK-Net-Lite architecture using TensorFlow 2.11 integrated with Keras. All experiments were conducted on a dual NVIDIA T4 GPU system (15 GB VRAM each), provided via the Kaggle platform.

Training was performed entirely from scratch, without any pretrained weights. The complete training configuration is summarized in Table I. A fixed random seed was used for reproducibility across all experiments.

TABLE I
Training Configuration for DUCK-Net and DUCK-Net-Lite

Parameter	Value
Input image size	352×352
Batch size	4
Epochs	200
Optimizer	RMSprop
Learning rate	0.0001
Loss function	Dice Loss
Output activation	Sigmoid

B. Data Augmentation

To improve generalization and reduce overfitting, we applied online data augmentation using the Albumentations library (v1.3.0). Each training image was randomly augmented at every epoch. This strategy, adapted from [29], includes:

- Horizontal and vertical flipping;
- Color jitter (brightness: $[0.6, 1.6]$, contrast: ± 0.2, saturation: ± 0.1, hue: ± 0.01);
- Affine transforms: rotation ($[-180°, 180°]$), translation ($\pm 12.5\%$), scaling ($[0.5, 1.5]$), and shearing ($[-22.5°, 22.5°]$).

All geometric transformations were applied simultaneously to both input images and masks to maintain alignment. Color transformations were applied to images only. This augmentation proved sufficient for regularization, and we did not use additional techniques such as dropout.

C. Dataset

We evaluated DUCK-Net on the ISIC-2016 dataset [23], a public dermoscopic image dataset with binary lesion masks annotated by dermatologists. It contains 900 images for training and 379 for testing.

Ground truth masks are provided in PNG format with pixel values of 1 (lesion) or 0 (background). We split the training set into 80% training and 20% validation, maintaining class balance and ensuring no overlap. The official test set was used as-is to enable comparison with prior works.

D. Quantitative Results

Tables II present the segmentation performance of DUCK-Net and DUCK-Net-Lite using 17/34 filters, respectively. We report Dice Similarity Coefficient (DSC), Intersection over Union (IoU), precision, recall, and accuracy on training, validation, and test sets.

TABLE II
Segmentation performance (%) of DUCK-Net and DUCK-Net-Lite on ISIC-2016 with 17/34 filters.

Model	Filter	Set	DSC	IoU	Prec.	Rec.	Acc.
DUCK-Net	17	Train	94.39	89.38	95.73	93.09	97.07
		Validation	93.38	87.58	94.61	92.18	96.16
		Test	92.52	86.08	92.59	92.45	95.79
DUCK-Net	34	Train	94.25	89.13	94.41	94.09	96.95
		Validation	93.57	87.92	93.03	94.11	96.20
		Test	92.51	86.07	91.34	93.72	95.72
DUCK-Net-Lite	17	Train	94.27	89.16	94.78	93.77	96.91
		Validation	92.41	85.89	93.63	91.22	95.94
		Test	92.00	85.19	91.74	92.26	95.48
DUCK-Net-Lite	34	Train	93.72	88.18	92.34	95.13	96.55
		Validation	92.06	85.28	91.41	92.72	95.66
		Test	91.90	85.02	89.56	94.37	95.31

Although the 34-filter model has more parameters, the performance difference is marginal. The 17-filter version achieves comparable metrics with fewer parameters, suggesting a better trade-off for resource-constrained settings.

E. Computational Complexity and Efficiency

While DUCK-Net achieves competitive performance across multiple metrics, its architectural complexity can pose limitations for deployment in resource-constrained environments. To address this, we introduce DUCK-Net-Lite, a streamlined variant that reduces computational overhead while maintaining comparable segmentation performance.

Table III compares the number of model parameters between DUCK-Net and DUCK-Net-Lite at two filter scales ($F = 17$ and $F = 34$). The Lite versions exhibit a reduction of approximately 44-45% in total parameters, significantly decreasing the model's memory footprint.

TABLE III
PARAMETER COUNTS OF DUCK-NET VS. DUCK-NET-LITE (EXCLUDING OPTIMIZER PARAMETERS).

F	Model	Total Params (#)	Trainable (#)	Non-trainable (#)
17	DUCK-Net	38,921,088	38,867,668	53,420
	DUCK-Net-Lite	21,585,151	21,542,951	42,200
34	DUCK-Net	155,410,341	155,303,507	106,834
	DUCK-Net-Lite	86,123,781	86,039,387	84,394

Similarly, Table IV reports the corresponding file sizes of the trained models. DUCK-Net-Lite achieves up to 44% reduction in model size, with the 17-filter version occupying only 175MB compared to the original DUCK-Net's 314MB. This reduction is achieved through architectural simplification, without significant performance degradation, as demonstrated in Table II.

TABLE IV
MODEL FILE SIZES (.KERAS) AND RELATIVE REDUCTION.

F	Model	Size	Δ (%)
17	DUCK-Net	314.11 MB	
	DUCK-Net-Lite	175.15 MB	-44.2
34	DUCK-Net	1,216.61 MB	
	DUCK-Net-Lite	691.27 MB	-43.2

F. Qualitative Results

To further evaluate the segmentation capability of the model, we provide qualitative results on representative test samples (Fig. 7). These include an easy case with high overlap, a moderately challenging case, and a difficult case with poor segmentation performance.

In the easy case (top row), the lesion is clearly visible with high contrast against the surrounding skin. The predicted mask closely matches the ground truth with minimal error, demonstrating the model's ability to accurately detect well-defined lesion boundaries.

In the medium case (middle row), the lesion is partially occluded by hair, making it difficult to localize precisely. As a result, the model oversegments the lesion area, highlighting the challenge of dealing with visual noise in dermoscopic images.

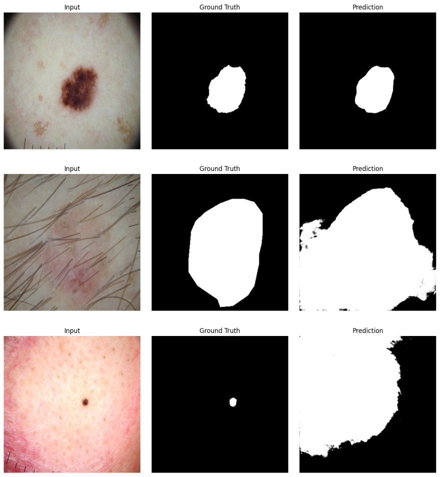

Fig. 7. Qualitative results on three representative test cases. Each row shows the input image, the ground truth mask, and the predicted mask. From top to bottom: easy case (high Dice), medium case (moderate Dice), and difficult case (low Dice).

In the difficult case (bottom row), the lesion is small and exhibits low contrast with the surrounding skin. Although the model still localizes the lesion region correctly, the predicted mask slightly underestimates its boundaries, indicating reduced performance under more challenging conditions.

These examples highlight the strengths and limitations of the model under varying lesion characteristics, offering insight into potential areas for further improvement.

G. Comparison with Prior Work

TABLE V
PERFORMANCE (%) COMPARISON ON ISIC-2016 TEST SET

Model	**DSC**	**IoU**
U-Net	87.81	80.25
Att-Unet	87.43	79.70
nnU-Net	90.45	84.52
SwinUNet	90.12	83.21
MISSFormer	90.46	83.92
DAEFormer	91.19	85.40
HiFormer	91.48	85.15
TransFuse	92.03	**86.19**
DUCK-Net (17 filters)	**92.52**	86.08
DUCK-Net (34 filters)	92.51	86.07
DUCK-Net-Lite (17 filters)	92.00	85.19
DUCK-Net-Lite (34 filters)	91.90	85.02

Baseline results (excluding DUCK-Net and DUCK-Net-Lite variants) are adopted from Table 1 in [22].

We benchmarked DUCK-Net and DUCK-Net-Lite against eight state-of-the-art models on the ISIC-2016 test set using DSC and IoU (Table V. DUCK-Net with 17 filters attains the highest DSC (92.52%) and the second-best IoU (86.08%), only 0.11 below TransFuse, despite being trained from scratch without pretrained weights. The compressed DUCK-Net-Lite

(17 filters) still delivers 92.00% DSC and 85.19% IoU, an acceptable drop given its 45% reduction in parameters. These results indicate that both variants remain highly competitive while offering a far more favorable accuracy-efficiency trade-off for resource-constrained clinical deployment.

V. Ablation Studies

To assess the contribution of both the DUCK block design and the architectural depth, we conduct two sets of ablation studies: (1) replacing DUCK blocks with standard convolutional blocks, and (2) reducing the number of encoder-decoder layers to analyze the role of each resolution scale.

A. *DUCK Block vs. Convolution*

We first evaluate the impact of DUCK blocks by replacing them with standard convolutional blocks, while keeping all other components and training settings unchanged (e.g., encoder–decoder structure, skip connections, and residual paths). This replacement is applied to both the full DUCK-Net and the Lite version with four layers.

Each configuration is evaluated under two filter widths (F=17 and F=34) on the ISIC-2016 test set. Table VI summarizes the results.

TABLE VI
Ablation results comparing DUCK and Conv blocks on the ISIC-2016 test set (F: number of filters).

Depth	Block	F	DSC	IoU	Prec.	Rec.	Acc.
Full	Conv	17	92.08	85.32	91.91	92.25	95.53
	DUCK	17	**92.52**	**86.08**	92.59	92.45	95.79
			+0.44	+0.76	+0.68	+0.20	+0.26
	Conv	34	92.27	85.65	91.01	93.57	95.58
	DUCK	34	**92.51**	**86.07**	91.34	93.72	95.72
			+0.24	+0.42	+0.33	+0.15	+0.14
Lite	Conv	17	88.91	80.04	86.18	91.83	93.55
	DUCK	17	**92.00**	**85.19**	91.74	92.26	95.48
			+3.09	+5.15	+5.56	+0.43	+1.93
	Conv	34	89.21	80.53	87.86	90.61	93.83
	DUCK	34	**91.90**	**85.02**	89.56	94.37	95.31
			+2.69	+4.49	+1.70	+3.76	+1.48

B. *Reducing Network Depth*

To evaluate the importance of each encoder-decoder layer, we gradually reduce the depth of DUCK-Net by removing the deepest layers in the architecture.

- DUCK-Net (Full): 5-layer model.
- DUCK-Net-Lite: 4-layer model (removes the deepest layer).
- DUCK-Net-Tiny: 3-layer model (removes the two deepest layers).

Table VII shows the performance of each version using F=17 and DUCK blocks.

TABLE VII
Impact of depth reduction on segmentation performance (DUCK Block)

Architecture	DSC	IoU	Prec.	Rec.	Acc.
17 Filters					
DUCK-Net (17 filters)	**92.52**	**86.08**	**92.59**	92.45	**95.79**
DUCK-Net-Lite (17 filters)	92.00	85.19	91.74	92.26	95.48
DUCK-Net-Tiny (17 filters)	90.59	82.80	91.21	89.98	94.73
34 Filters					
DUCK-Net (34 filters)	92.51	86.07	91.34	**93.72**	95.72
DUCK-Net-Lite (34 filters)	91.90	85.02	89.56	94.37	95.31
DUCK-Net-Tiny (34 filters)	90.62	82.85	88.31	93.06	94.57

These results indicate that removing the final bottleneck layer (from Full to Lite) results in only a minor drop in performance, suggesting that the deepest layer may not be essential for effective segmentation on ISIC-2016. However, further removing the next encoder layer (Lite to Tiny) leads to a substantial performance drop, highlighting the necessity of preserving sufficient depth for multi-scale feature extraction. Thus, DUCK-Net-Lite represents a balanced trade-off between performance and efficiency.

VI. Discussion

To assess how each encoder–decoder layer contributes to segmentation quality, we experiment with progressively shallower versions of DUCK-Net. DUCK-Net-Lite eliminates the deepest layer, while DUCK-Net-Tiny removes the two innermost levels. Results show that DUCK-Net-Lite retains most of the performance of the original architecture, indicating that the deepest layer contributes relatively little to accuracy while incurring considerable computational cost.

However, DUCK-Net-Tiny experiences a measurable drop in all metrics, suggesting that the second-deepest layer plays a more substantial role in encoding semantic context. These outcomes confirm that selectively removing only the bottommost layer yields an efficient model, whereas further pruning harms performance more significantly. Hence, DUCK-Net-Lite offers a practical balance between simplicity and effectiveness.

Interestingly, although DUCK-Net-Tiny has fewer layers, it requires longer training time than both DUCK-Net and DUCK-Net-Lite. This is due to the higher-resolution feature maps processed throughout the shallower network, which increase computational load per iteration. The absence of deep downsampling reduces the efficiency gains typically associated with smaller models. This observation emphasizes that model compactness does not always correlate with faster training and highligh

VII. Conclusion

This study adapts DUCK-Net for skin lesion segmentation on the ISIC-2016 dataset and demonstrates that accurate results can be achieved with a compact design using just 17 filters per layer and no external pretraining.

To optimize the model for deployment, we proposed DUCK-Net-Lite by removing the deepest layer of the network. Our experiments show that this reduced version maintains strong performance while lowering the computational footprint.

We also evaluated DUCK-Net-Tiny, a more aggressive variant that removes two deep layers. While this version is even smaller, it shows a noticeable decline in segmentation quality. These findings highlight that only the deepest layer can be safely pruned, and further reduction compromises semantic understanding.

Future directions may involve testing on more complex datasets, designing smarter pruning techniques, or adding lightweight attention to further enhance boundary sensitivity.

References

[1] W. H. Organization, "Skin cancers." https://www.who.int/news-room/fact-sheets/detail/ultraviolet-(uv)-radiation, 2023. Accessed: 2024-07-01.

[2] A. A. Marghoob and H. S. Rabinovitz, "Dermoscopy for melanoma detection in everyday practice: Current status and future directions," *Dermatologic Clinics*, vol. 21, no. 4, pp. 665–678, 2003.

[3] M. E. Celebi, T. Mendonça, and J. S. Marques, "Lesion border detection in dermoscopy images," *Dermatologic Image Analysis*, pp. 107–129, 2015.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, p. 180161, 2018.

[5] 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 2016 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pp. 168–172, IEEE, 2018.

[6] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[7] 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*, pp. 234–241, Springer, 2015.

[8] 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.

[9] 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 *European conference on computer vision*, pp. 205–218, Springer, 2022.

[10] M. De Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*, vol. 12903. Springer Nature, 2021.

[11] M. M. Islam, S. Anwar, and N. Barnes, "A survey on deep learning techniques for medical image segmentation," *arXiv preprint arXiv:2009.13120*, 2020.

[12] R.-G. Dumitru, D. Peteleaza, and C. Craciun, "Using duck-net for polyp image segmentation," *Scientific reports*, vol. 13, no. 1, p. 9803, 2023.

[13] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[14] M. Zhang, Y. Zhang, S. Liu, Y. Han, H. Cao, and B. Qiao, "Dual-attention transformer-based hybrid network for multi-modal medical image segmentation," *Scientific Reports*, vol. 14, no. 1, p. 25704, 2024.

[15] X. Li, X. Qin, C. Huang, Y. Lu, J. Cheng, L. Wang, O. Liu, J. Shuai, and C.-a. Yuan, "Sunet: A multi-organ segmentation network based on multiple attention," *Computers in Biology and Medicine*, vol. 167, p. 107596, 2023.

[16] X. Huang, Z. Deng, D. Li, and X. Yuan, "Missformer: An effective medical image segmentation transformer," *arXiv preprint arXiv:2109.07162*, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6824–6835, 2021.

[19] 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.

[20] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E. K. Aghdam, J. Cohen-Adad, and D. Merhof, "Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 6202–6212, 2023.

[21] R. Azad, L. Niggemeier, M. Hüttemann, A. Kazerouni, E. K. Aghdam, Y. Velichko, U. Bagci, and D. Merhof, "Beyond self-attention: Deformable large kernel attention for medical image segmentation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1287–1297, 2024.

[22] S. Qamar, S. F. Qadri, R. Alroobaea, G. M. M. Alshmrani, and R. Jiang, "Scalefusionnet: Transformer-guided multi-scale feature fusion for skin lesion segmentation," *arXiv preprint arXiv:2503.03327*, 2025.

[23] D. Gutman, N. C. 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)," *arXiv preprint arXiv:1605.01397*, 2016.

[24] J. Jing *et al.*, "Mobile-unet: An efficient convolutional neural network for fabric defect detection," *Textile Research Journal*, vol. 92, no. 1-2, pp. 30–42, 2022.

[25] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[26] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[27] D. e. a. Jha, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE International Symposium on Multimedia (ISM)*.

[28] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen, "Resunet++: An advanced architecture for medical image segmentation," in *2019 IEEE international symposium on multimedia (ISM)*, pp. 225–2255, IEEE, 2019.

[29] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020.