bioengineering

MDPI

Article

DermViT: Diagnosis-Guided Vision Transformer for Robust and Efficient Skin Lesion Classification

Xuejun Zhang [1], Yehui Liu [1], Ganxin Ouyang [2], Wenkang Chen [1], Aobo Xu [1], Takeshi Hara [2], Xiangrong Zhou [2] and Dongbo Wu [3,4,*]

[1] School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China; xjzhang@gxu.edu.cn (X.Z.)
[2] Department of Electrical, Electronic and Computer Engineering, Gifu University, Gifu 501-1193, Japan
[3] Department of General Surgery, The Fourth Affiliated Hospital of Guangxi Medical University, Liuzhou 545000, China
[4] Department of Gastrointestinal, Metabolic and Bariatric Surgery, Ruikang Hospital Affiliated to Guangxi University of Chinese Medicine, Nanning 530004, China
* Correspondence: wudongbobo@126.com; Tel.: +86-139-7886-5862

Abstract: Early diagnosis of skin cancer can significantly improve patient survival. Currently, skin lesion classification faces challenges such as lesion–background semantic entanglement, high intra-class variability, artifactual interference, and more, while existing classification models lack modeling of physicians' diagnostic paradigms. To this end, we propose DermViT, a medically driven deep learning architecture that addresses the above issues through a medically-inspired modular design. DermViT consists of three main modules: (1) Dermoscopic Context Pyramid (DCP), which mimics the multi-scale observation process of pathological diagnosis to adapt to the high intraclass variability of lesions such as melanoma, then extract stable and consistent data at different scales; (2) Dermoscopic Hierarchical Attention (DHA), which can reduce computational complexity while realizing intelligent focusing on lesion areas through a coarse screening–fine inspection mechanism; (3). Dermoscopic Feature Gate (DFG), which simulates the observation–verification operation of doctors through a convolutional gating mechanism and effectively suppresses semantic leakage of artifact regions. Our experimental results show that DermViT significantly outperforms existing methods in terms of classification accuracy (86.12%, a 7.8% improvement over ViT-Base) and number of parameters (40% less than ViT-Base) on the ISIC2018 and ISIC2019 datasets. Our visualization results further validate DermViT's ability to locate lesions under interference conditions. By introducing a modular design that mimics a physician's observation mode, DermViT achieves more logical feature extraction and decision-making processes for medical diagnosis, providing an efficient and reliable solution for dermoscopic image analysis.

Keywords: skin cancer; skin lesion classification; transformer; dermoscopic image analysis

check for updates

Academic Editors: Kwang Woo Ahn and Elpiniki I. Papageorgiou

Received: 9 March 2025
Revised: 27 March 2025
Accepted: 10 April 2025
Published: 16 April 2025

Citation: Zhang, X.; Liu, Y.; Ouyang, G.; Chen, W.; Xu, A.; Hara, T.; Zhou, X.; Wu, D. DermViT: Diagnosis-Guided Vision Transformer for Robust and Efficient Skin Lesion Classification. *Bioengineering* **2025**, *12*, 421. https://doi.org/10.3390/bioengineering12040421

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Early diagnosis of skin cancer is crucial for improving patient prognosis. The five-year survival rate for malignant melanoma can reach 99% with early detection, but declines sharply to 23% when diagnosed at advanced stages [1]. Dermoscopy [2], which provides structural information on the epidermis, has become an important non-invasive diagnostic tool [3]. However, this technique is highly dependent on physicians' expertise and subjective assessment, resulting in a considerable risk of misdiagnosis [4]. Therefore, there is an urgent need to develop efficient and reliable computer-aided diagnostic systems [5].

As illustrated in Figure 1, automatic classification of skin diseases faces significant challenges, including high inter-class similarity, high intra-class variability, and artifacts such as hair and air bubbles. Early research was primarily based on traditional methods that combined handcrafted feature extraction and machine learning [6]. For instance, Celebi et al. [7] proposed an image segmentation strategy based on Euclidean distance transformation, while Abbas et al. [8] constructed color–texture feature models in color space. Although these methods achieved some success on limited datasets, they were constrained by inherent limitations such as complex feature engineering and weak generalization capabilities.

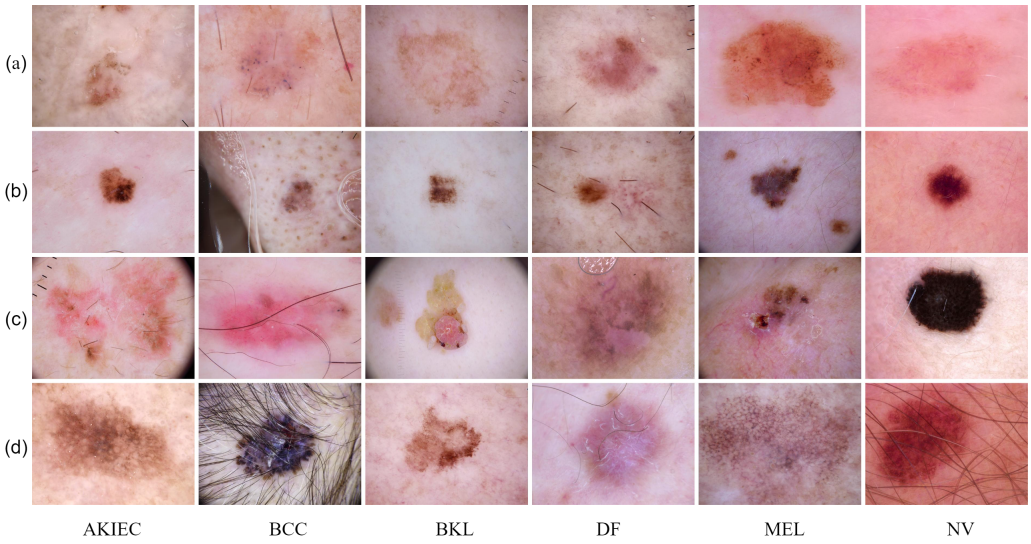


Figure 1. Examples of skin lesions from ISIC2018 dataset; each column represents the same type of disease. One challenge to the classification of skin lesions is high intra-class variation. Columns (**a**,**b**) show examples of high inter-class similarity, while columns (**c**,**d**) show examples of hair and other artifacts.

The advent of convolutional neural networks (CNNs) revolutionized skin lesion classification [9]. Esteva et al. [10] showed that CNNs could match dermatologists in diagnostic accuracy. Harangi et al. [11] achieved enhanced classification robustness through model ensemble strategies, while Aldhyani et al. [12] balanced efficiency and accuracy through a dynamic kernel design. However, their local receptive field makes CNNs insufficient for modeling the global contextual relationships of lesions, which limits their ability to handle diffuse lesions [13]. To address this issue, researchers have incorporated attention mechanisms to enhance classification performance. For instance, Gessert et al. [14] proposed a patch-based attention mechanism combined with diagnosis-guided loss weighting to improve the focus on hard-to-classify cases and strengthen global feature modeling. More recently, Yu et al. [15] have demonstrated the effectiveness of cross-attention mechanisms for multimodal feature integration, achieving 91% accuracy in melanoma classification.

The success of the transformer architecture [16] in natural language processing has provided a new paradigm for visual tasks. The vision transformer (ViT) architecture [17] were the first to validate the visual representation capability of pure transformer models on ImageNet classification tasks by partitioning images into sequential patches and modeling global dependencies through self-attention mechanisms. This breakthrough has facilitated its application in medical imaging. For instance, Soumyya et al. [18] introduced a soft attention mechanism to improve ViT's focus on lesion regions, resulting in reduced dependency on global features and enhanced classification accuracy. Zhao et al. [13] introduced a fully transformer-based FTN network that achieved efficient feature learning, while

Li et al. [19] designed a CNN–transformer dual-branch model that eliminated semantic discrepancies between the two branches using the BC-FCU module, raising the multi-class classification accuracy to 87.6%. These and similar advancements [20–22] demonstrate that transformer-based models offer significant advantages in addressing the locality limitations of CNNs and capturing long-range dependencies.

However, the original ViT design faces significant challenges in medical imaging. First, the computational complexity of the global self-attention mechanism scales quadratically with image resolution [23]. This leads to substantial computational overhead for high-resolution dermoscopic images, which typically exceed 1024 × 1024 pixels [24]. Second, ViT lacks incorporation of domain-specific medical knowledge; dermatologists rely on a multi-scale observation paradigm (i.e., first identifying regions of interest at low magnification and then refining the diagnosis at high magnification) and an attention-focusing mechanism (i.e., initially filtering out irrelevant regions before analyzing lesion features in detail), whereas ViT's uniform allocation of attention does not align with this diagnostic logic [25]. More critically, ViT treats all visual tokens equally during feature propagation [26], which allows irrelevant artifacts such as hair or bubbles to introduce noise into the deep semantic representations. The small sample size typical of medical datasets (e.g., ISIC2018 contains only 10,015 images) further increases the likelihood of overfitting to spurious features. To address these issues, we propose DermViT, which mitigates the inherent limitations of ViT through a structural redesign inspired by medical diagnostic principles. DermViT consists of the following key modules:

1. Dermoscopic Hierarchical Attention (DHA): DHA mimics the attention-shifting mechanism employed by dermatologists, enabling efficient allocation of computational resources through a two-stage routing mechanism. In the first stage, DHA computes region-level attention weights in order to rapidly discard diagnostically irrelevant background regions. In the second stage, fine-grained token-level attention is applied exclusively within the retained regions, concentrating on dermatologically relevant local features. This hierarchical coarse-to-fine strategy reduces computational overhead while enhancing the sensitivity of detection for subtle lesions.
2. Dermoscopic Context Pyramid (DCP): Inspired by the multi-magnification collaborative mechanism in pathological diagnosis, DCP adapts to the high intra-class variability of lesions such as melanoma through cross-scale feature fusion.
3. Dermoscopic Feature Gate (DFG): Inspired by the observation–verification cycle in clinical diagnosis, DFG suppresses semantic noise through channel decoupling and dynamic gating. It decouples input features into a primary pathway for morphological analysis and an auxiliary pathway for detail verification. The primary pathway captures local context (e.g., pigment texture), while the auxiliary pathway calibrates response intensity through channel-wise multiplication. This mechanism emulates the cognitive process of dermatologists, who first observe globally before zooming in to verify local details, effectively attenuating high-frequency artifacts (e.g., hair noise) that may interfere with diagnosis.

The experimental results demonstrate that DermViT achieves strong classification capability without pretraining. In the presence of challenging background artifacts, its attention heatmaps strongly align with ground = truth lesion annotations, surpassing counterparts based on both CNN [27–30] and transformer [17,19,31] architectures. The proposed model outperforms existing SOTA methods in classification accuracy while substantially enhancing efficiency and resource utilization, offering an optimized and dependable framework for automated skin cancer classification.

2. Materials and Methods

This section introduces the proposed methodology. We first describe the key modules of the model, then elaborate on the overall framework in the final section to present the overall design concept of the approach.

2.1. Dermoscopic Hierarchical Attention

The traditional global self-attention mechanism treats all regions uniformly, resulting in inefficient resource utilization and feature contamination [32]. To address this issue, DHA draws inspiration from sparse attention [23,24,26] and employs a hierarchical self-attention mechanism to mimic the hierarchical diagnostic process of dermatologists through a two-stage process consisting of coarse region filtering followed by fine-grained pixel inspection. The specific process is shown in Figure 2.

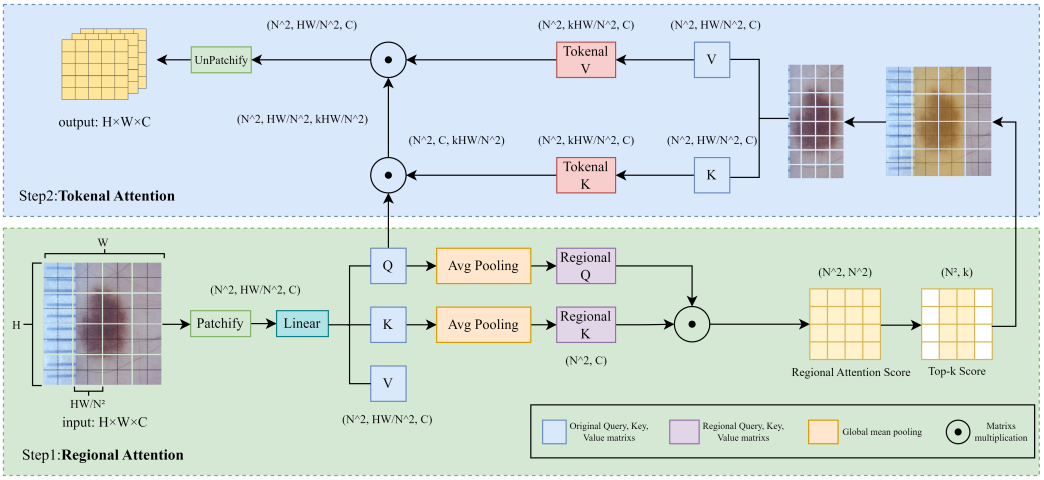


Figure 2. Dermoscopic Hierarchical Attention: The DHA module's hierarchical attention involves (1) region-level coarse screening for global relevance followed by (2) fine-grained attention focusing on the top-K regions. This design filters obvious background noise (e.g., rulers) while improving both efficiency and precision.

Regional Coarse Screening for Filtering Out Irrelevant Background. The input feature map is first divided into **N** regions through a reshape operation. After reshaping, the image undergoes linear mapping to generate the corresponding original matrices for the query, key, and value, i.e., $Q, K, V \in \mathbb{R}^{N \times \frac{HW}{N} \times d}$, where N is the number of regions, HW is the size of the feature map, and d is the feature dimension. To quickly filter key regions, DHA applies global average pooling (GAP) to Q and K to obtain region-level queries and keys:

$$Q^{\mathrm{r}} = \mathrm{GAP}(\mathrm{Q}) \in \mathbb{R}^{N \times d}, \quad \mathrm{K}^{\mathrm{r}} = \mathrm{GAP}(\mathrm{K}) \in \mathbb{R}^{N \times d}. \tag{1}$$

Based on $\mathbf{Q}^r$ and $\mathbf{K}^r$, the regional similarity score matrix is computed as follows:

$$S = Q^r \cdot (K^r)^T \in \mathbb{R}^{N \times N}. \tag{2}$$

DHA employs a selective screening mechanism that retains only the top-k most relevant regions:

$$\mathcal{R}\mathrm{top} - \mathrm{k} = \mathrm{Topk}(S) \subset 1, 2, \ldots, N. \tag{3}$$

By discarding the $N - k$ least relevant regions, this strategy significantly reduces computational complexity while effectively suppressing background artifacts such as hair and measurement scales.

Token-Level Refinement Focusing on Lesion Characteristics. Among the k candidate regions identified in the previous step, DHA extracts the token-level key ($\mathbf{K}^t$) and value ($\mathbf{V}^t$) matrices from $\mathbf{K}, \mathbf{V}$ based on the similarity score matrix $\mathbf{S}$, then applies fine-grained self-attention as follows:

$$\text{Attention}(Q, K^t, V^t) = \text{Softmax}\left(\frac{Q \cdot (K^t)^T}{\sqrt{d_k}}\right) \cdot V^t. \tag{4}$$

In the above equation, $\mathbf{Q}$ represents the original query matrix, while $\mathbf{K}^t$ and $\mathbf{V}^t$ respectively correspond to the key and value matrices of the selected candidate regions. This stage simulates the detailed examination process of a dermatologist using magnification, which helps to ensure precise focus on subtle lesion features. After computing the self-attention, the feature map is restored to its original size of $Y \in \mathbb{R}^{N \times W \times C}$ before being passed to the next stage.

2.2. Dermoscopic Feature Gate

In dermoscopic images, lesion regions are often highly coupled with artifacts such as hairs and blisters, making it difficult for traditional feature propagation mechanisms to effectively suppress the semantic leakage of artifact-related features. To address this issue, the proposed Dermoscopic Feature Gate (DFG) is inspired by the "observation–verification" diagnostic loop used by dermatologists. Specifically, clinicians first perform a global observation to locate suspicious areas, followed by a detailed examination under magnification to verify pathological features, then finally apply their clinical expertise to filter out irrelevant information. As shown in Figure 3, DFG formalizes this process into a learnable feature operation through a dual-path gated convolutional mechanism, which simulates the diagnostic logic of dermatologists in feature space to significantly increase the model's robustness against interference from artifacts. Given an input feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, the DFG module operates as shown below.

Global Observation (Feature Decoupling). The input feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ is first processed through a hidden layer to increase feature complexity. This step simulates a dermatologist's initial global observation, where the clinician scans the entire image to identify potential lesion regions. The resulting feature map $\mathbf{Z} \in \mathbb{R}^{2C \times H \times W}$ is then split into two parts, $\mathbf{X}_1$ and $\mathbf{V}$, respectively representing the global features and gating features:

$$\mathbf{X}_1, \mathbf{V} = \text{split}(\text{Conv}_{2\times C}(\mathbf{X})) \in \mathbb{R}^{2C \times H \times W}. \tag{5}$$

Detailed Inspection (Spatial Feature Extraction). Next, $\mathbf{X}_1$ is processed through a 3×3 depthwise separable convolution, which enhances spatial feature extraction without significantly increasing the parameter count. This step mimics a dermatologist's detailed inspection under magnification, where fine-grained features such as irregular borders and color heterogeneity are analyzed. The output $\mathbf{X}_2$ captures the refined spatial features:

$$\mathbf{X}_2 = \text{ReLU}(\text{Conv}_{\text{depthwise}}(\mathbf{X}_1)) \in \mathbb{R}^{2C \times H \times W}. \tag{6}$$

Clinical Decision (Feature Filtering). Finally, $\mathbf{X}_2$ is multiplied by the gating signal $\mathbf{V}$ to dynamically calibrate the response strength of the observing branch. This helps to ensure that only information relevant to the lesion is enhanced while effectively suppressing background noise, resulting in a calibrated feature. This step emulates the clinician's decision-making process, where irrelevant information (e.g., hairs, rulers) is filtered out based on clinical experience and only diagnostically relevant features are retained:

$$\mathbf{Y} = \text{Dropout}(\text{Conv}_{1\times 1}((\mathbf{X_2} \odot \mathbf{V}))) \in \mathbb{R}^{C \times H \times W}. \tag{7}$$

By integrating these steps, the DFG module effectively mimics the dermatologist's diagnostic workflow, enabling the model to dynamically focus on lesion-specific features while suppressing interference from artifacts. Furthermore, this mechanism significantly reduces computational complexity, achieving more efficient lesion feature extraction compared to traditional feature modeling approaches.

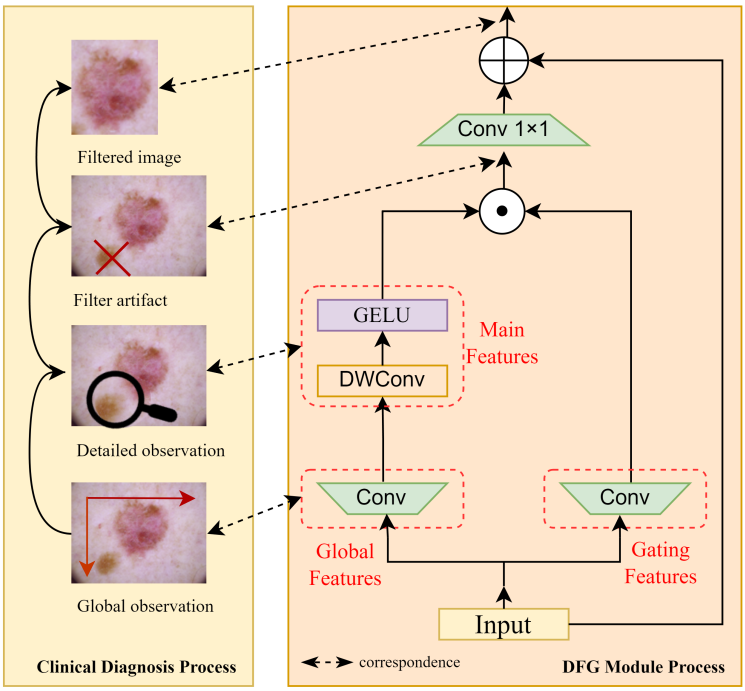


Figure 3. Dermoscopic Feature Gate: The left figure shows the diagnostic logic of a dermatologist; the right figure shows how the design of the DFG module is based on this logic, simulating the process of filtering out artifacts during a doctor's diagnostic process.

2.3. Dermoscopic Context Pyramid

ViT exhibits an inherent contradiction between its global unified attention mechanism and the multi-scale observational paradigm of dermatological diagnosis. Clinicians typically first localize lesion regions at low magnification, then switch to high magnification to analyze the detailed features. In contrast, ViT's fixed-resolution processing mode fails to accommodate this dynamic cognitive process. Additionally, ViT's neglect of cross-scale contextual information results in blurred lesion boundaries, severely limiting diagnostic accuracy. The Dermoscopic Contextual Perception (DCP) module addresses this limitation by structurally modeling a multi-magnification collaborative mechanism that simulates a clinician's cognitive process of cross-scale feature comparison. This provides multi-granularity semantic support for the subsequent DHA mechanism.

As shown in Figure 4, the DCP consists of a spatial compressor with a multiscale pyramid transformer (MPT). For simplicity, we only draw the key attention schema in the MPT shown in Figure 4. The computational flow is closely aligned with the dermoscopic diagnostic criteria:

Low-Magnification Semantic Compression: Spatial downsampling of input feature maps by convolutional operations to generate coarse-grained semantic anchors:

$$X_{\text{down}} = \text{Conv2D}(X) \in \mathbb{R}^{\frac{H}{k} \times \frac{W}{k} \times C}. \tag{8}$$

This procedure mimics rapid lesion localization at low magnification, compressing extraneous background areas while preserving the core morphological features of the lesion.

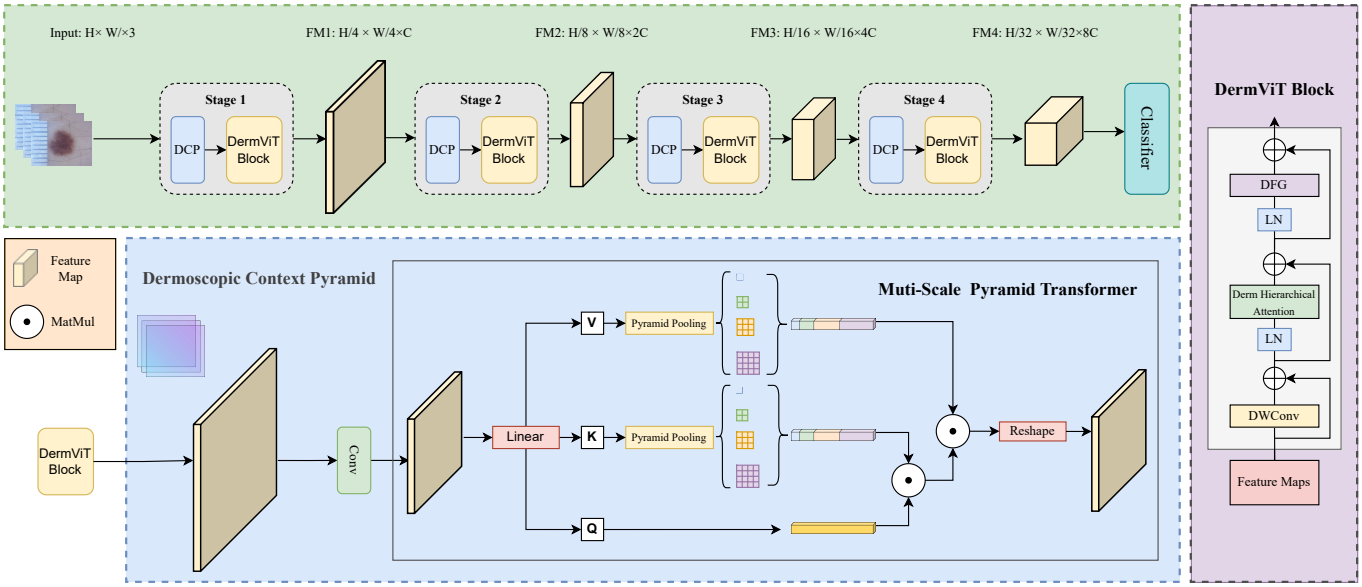


Figure 4. DermViT: Our proposed multi-stage framework. Each stage contains the following modules: (1) DCP for multi-scale feature-preserving downsampling; (2) DHA with two-stage attention (region filtering and token refinement) to suppress irrelevant background; and (3) DFG for imaging artifact removal, enabling robust lesion classification.

High-magnification multi-scale comparisons: Cross-scale contextual features are generated via pyramid vooling (PSP) in a compressed feature space. First, PSP is performed on the key and value to extract the multi-resolution semantics:

$$K_{\text{psp}} = \text{Concat}(\text{AvgPools}_i(K))s_i \in 1,2,4,16, \quad V_{\text{psp}} = \text{Concat}(\text{AvgPools}_i(V))s_i \in 1,2,4,16. \tag{9}$$

Multi-scale pyramid attention is then applied. The original spatial coordinates of the query vector (Q) are maintained, and fine-grained feature focusing is achieved by interacting with multi-scale key values:

$$\text{Attn}(Q, K_{\text{psp}}, V_{\text{psp}}) = \text{Softmax}\left(\frac{QK_{\text{psp}}^T}{\sqrt{d}}\right)V_{\text{psp}}. \tag{10}$$

This mechanism allows the model to simultaneously capture the overall distribution pattern of the lesion (s = 16) with microstructural details (s = 1), simulating the visual switching of a physician between different magnifications. In addition, this design transforms the multi-scale cognitive process of a doctor into a differentiable computational structure, providing a globally consistent and locally discriminative feature representation for subsequent hierarchical attention and forming a complete diagnostic simulation chain.

2.4. The Entire Network Structure

The whole network structure is shown in Figure 4. The input of the network is the original dermatoscopy image $I \in \mathbb{R}^{H\times W\times 3}$. Subsequently, the multi-scale semantics are decoupled step-by-step through a four-level cascade of DCP-DHA-DFG processing units. At each level of processing, the DCP module performs spatial compression and cross-scale context fusion of the input features to generate low-resolution multi-granularity features. The DHA module then performs a two-phase route based on the multi-scale a priori output of the DCP. The background area is filtered through coarse region-level screening to filter the background region, after which token-level fine-grained attention is executed within the reserved region. The DFG module then performs channel decoupling

and dynamic calibration on the attention output to suppress semantic contamination due to high-frequency noise such as hairs and bubbles. The final output is the lesion classification probability.

2.5. Experimental Configuration

To validate the performance of our dermatological classification model, we conducted extensive experiments on the publicly available ISIC 2018 and ISIC 2019 datasets [33].

The ISIC 2018 dataset is widely used for skin lesion analysis. It contains 10,015 images of skin lesions across seven different types: Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanocytic Nevi (NV), Melanoma (MEL), and Vascular Skin Lesions (VASC). The ISIC 2019 dataset is larger and more diverse, comprising 25,331 images spanning nine lesion categories. In this study we divided the datasets into training, validation, and test sets with a ratio of 0.8:0.05:0.15.

All experiments were conducted using the PyTorch (version 2.1.0) framework on a single NVIDIA RTX 4090 GPU with 24 GB of memory and running on a Linux system. To enhance model generalization [34,35], the data augmentation techniques of random flipping and random rotation were applied to the training set. Additionally, input images were normalized using a mean of [0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225] to maintain consistency in training.

The models were trained from scratch using a class-weighted cross-entropy loss function to address dataset imbalance [14]. Optimization was carried out with the AdamW optimizer. The initial learning rate was set to 0.0002, and a cosine annealing strategy was employed to gradually reduce the learning rate throughout the training process. Specifically, the learning rate decayed to 0.01 times its initial value within 100 epochs, ensuring smooth convergence towards an optimal solution. To mitigate overfitting, a weight decay parameter of 0.0002 was applied to constrain model complexity and enhance generalization performance. In order to balance efficient training with computational constraints, the batch size was set to 32 [36,37].

3. Results

3.1. Comparison Experiments

In this study, we systematically conducted comparative experiments to evaluate the performance of our proposed model against several classical CNNs (e.g., ResNet34, ConvNeXt) as well as the latest transformer networks (e.g.,ViT and SwinT). In addition, we compared DermViT with models that have performed well in recent years in the field of dermatology classification, such as RCSABC and Biformer. Tables 1 and 2 present the experimental results of the different models on the ISIC2018 and ISIC2019 datasets. From the results of our comparative experiments, it can be observed that the proposed method outperforms other baseline models in terms of multiple key metrics while maintaining lower computational complexity and parameter count.

Table 1. Experimental results on the ISIC2018 dataset.

Model	FLOPs (G)	Parameters (M)	ISIC2018				
			Precision (%)	Recall (%)	F1 (%)	Accuracy (%)	MAUC (%)
Resnet34	3.68	21.8	69.02 ± 2.38	66.80 ± 0.90	67.24 ± 1.42	81.67 ± 0.52	95.56 ± 0.20
Resnet50	4.13	23.52	71.33 ± 1.04	69.58 ± 1.60	70.14 ± 0.94	82.35 ± 0.46	95.58 ± 0.17
Resnet101	7.87	44.55	68.92 ± 1.77	69.20 ± 1.36	68.67 ± 0.92	81.93 ± 0.41	95.68 ± 0.11
ConvNeXt-tiny	4.45	27.8	61.50 ± 1.16	59.05 ± 1.76	59.18 ± 1.33	78.09 ± 0.33	94.27 ± 0.07
Swin-tiny	8.74	27.5	62.28 ± 0.95	62.41 ± 1.14	61.88 ± 1.11	80.04 ± 0.25	95.42 ± 0.04
Swin-small	17.09	48.8	52.80 ± 0.64	47.63 ± 0.73	49.70 ± 0.51	74.11 ± 0.27	92.80 ± 0.04
ViT-Base	16.86	85.65	60.86 ± 2.22	55.03 ± 1.38	56.88 ± 1.50	77.75 ± 0.42	93.00 ± 0.21
RCSABC [19]	22.89	81.19	80.38	76.66	**78.3**	**87.39**	\
Biformer	4.42	27.65	71.17 ± 1.10	71.44 ± 1.49	69.86 ± 0.97	83.00 ± 0.45	95.18 ± 0.13
DermViT	**2.75**	**16.82**	**80.75** ± 1.07	**76.82** ± 0.50	77.90 ± 0.43	85.34 ± 0.23	**96.29** ± 0.05

Note: Bolded data is the best under this indicator, and underlined data indicates second place.

Table 2. Experimental results on the ISIC2019 dataset.

Model	FLOPs (G)	Parameters (M)	ISIC2019				
			Precision (%)	Recall (%)	F1 (%)	Accuracy (%)	MAUC (%)
Resnet34	3.68	21.8	68.42 ± 0.93	66.59 ± 0.91	67.11 ± 0.87	78.92 ± 0.46	95.75 ± 0.06
Resnet50	4.13	23.52	70.31 ± 0.86	68.00 ± 0.43	68.83 ± 0.51	79.38 ± 0.21	96.07 ± 0.06
Resnet101	7.87	44.55	66.47 ± 0.73	65.49 ± 0.88	65.57 ± 0.61	77.44 ± 0.34	95.41 ± 0.07
ConvNeXt-tiny	4.45	27.8	58.90 ± 1.21	51.71 ± 1.50	53.94 ± 0.93	71.55 ± 0.27	93.21 ± 0.05
swin-tiny	8.74	27.5	56.96 ± 1.40	42.23 ± 1.39	45.17 ± 1.36	67.48 ± 1.19	89.79 ± 0.23
swin-small	17.09	48.8	63.40 ± 0.38	58.53 ± 0.37	60.43 ± 0.15	74.80 ± 0.12	93.89 ± 0.02
Vit-base	16.86	85.65	58.66 ± 0.52	52.61 ± 0.90	54.54 ± 0.78	70.40 ± 0.33	92.39 ± 0.07
RCSABC [19]	22.89	81.19	\	\	\	\	\
Biformer	4.42	27.65	72.35 ± 1.04	69.52 ± 1.17	70.68 ± 0.75	78.77 ± 0.31	95.71 ± 0.12
Ours	**2.75**	**16.82**	**72.39** ± 0.41	**71.24** ± 0.31	**71.57** ± 0.30	**80.82** ± 0.20	**96.42** ± 0.03

Note: Bolded data is the best under this indicator, and underlined data indicates second place.

Performance Analysis: On the ISIC2018 dataset, DermViT outperforms benchmark models such as ResNet, ViT, SwinT, and ConvNeXt in terms of precision (80.75% ± 1.07%), recall (76.82% ± 0.50%), and MAUC (96.29% ± 0.05%). Although its F1-score (77.90% ± 0.43%) and accuracy (85.34% ± 0.23%) are slightly lower than those of RCSABC (78.3% and 87.39%, respectively), the differences are minimal, with RCSABC having a much larger parameter size (81.19 M vs. 16.82 M for DermViT). On the ISIC2019 dataset, DermViT maintains its lead in all evaluation metrics. The performance comparison is shown in Figure 5, where DermViT's balanced results highlight its superior capability. As seen in Figure 6, DermViT also outperforms ConvNext-tiny, ResNet-50, and ViT-base in accuracy across all seven skin lesion categories, with only a slight performance gap in the BCC category.

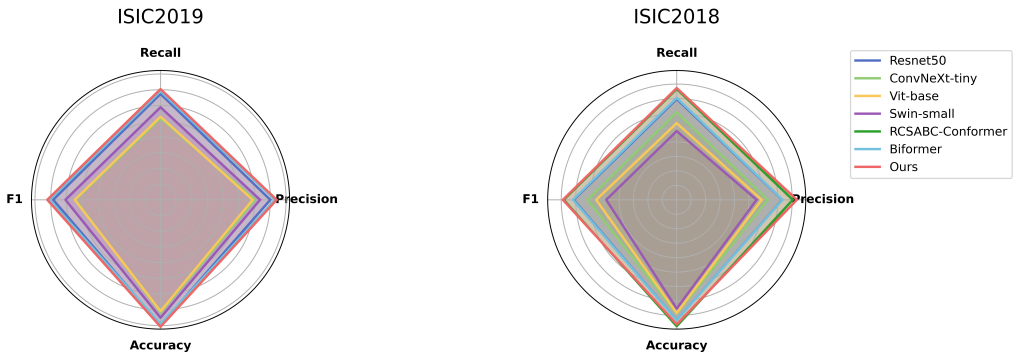


Figure 5. Radar charts for comparison experiments.

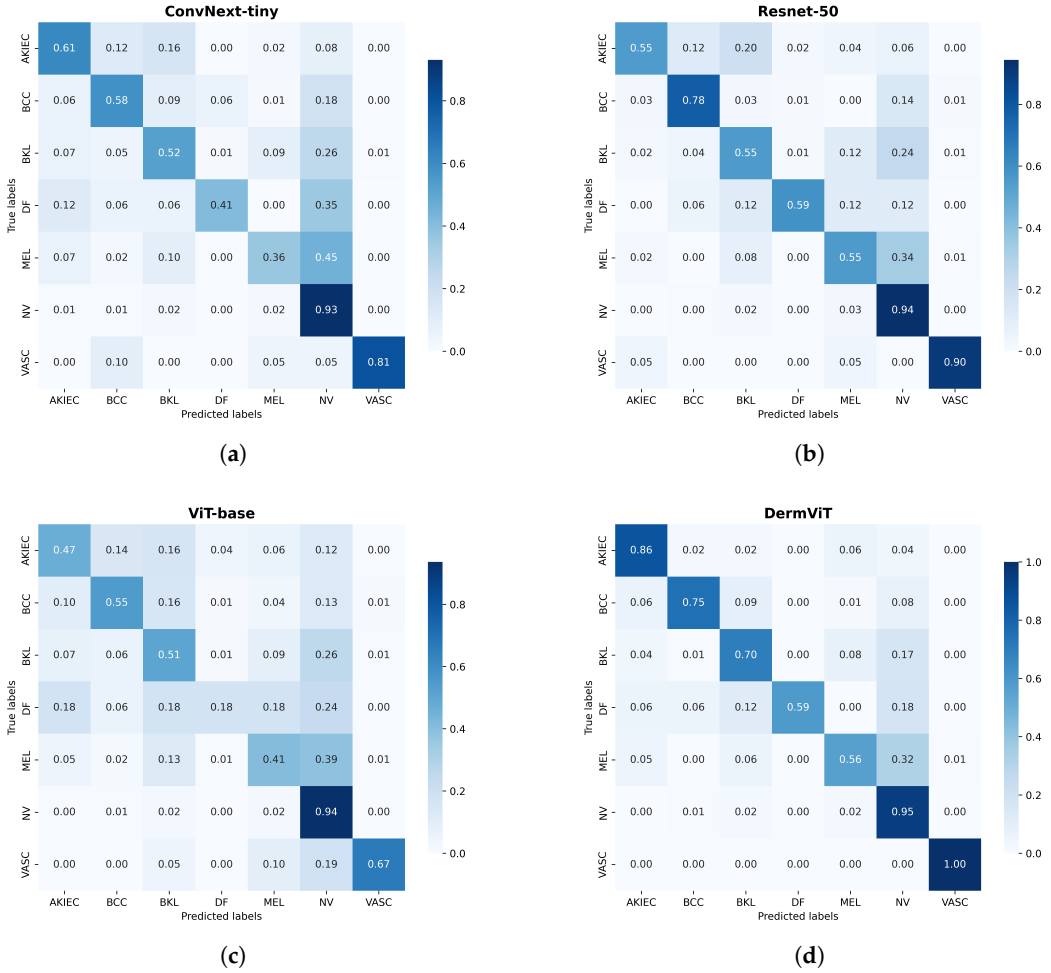


Figure 6. Multiclassification confusion matrices on the ISIC2018 dataset: (**a**) ConvNext-tiny model, (**b**) Resnet-50 model, (**c**) ViT-base model, and (**d**) our proposed DermViT model.

Computational Efficiency Analysis: Our method significantly enhances computational efficiency, reducing FLOPs (2.75 G) and parameters (16.82 M) compared to ResNet50 (4.13 G FLOPs, 23.52 M parameters) and RCSABC (22.89 G FLOPs, 81.19 M parameters). This improvement stems from our redesigned DermViT block, which employs a more efficient attention mechanism for global feature modeling. Additionally, instead of traditional FFN layers, we use the convolution-based DFG module for feature transmission, which drastically reduces the number of parameters while preserving performance.

These results show that our model meets or even exceeds advanced models, with higher parameters in most of the key metrics, while still provides excellent classification results in resource-constrained application scenarios, verifying that our model can outperform traditional CNNs and transformers on the task of skin lesion classification while keeping computational and storage costs low. These results demonstrate DermViT's combination of efficient parameter utilization and excellent classification performance.

3.2. Ablation Experiments

Based on the ablation experiment results shown in Table 3, we analyzed the impact of each proposed module on model performance and resource consumption.

Table 3. Ablation experiment results on the ISIC dataset.

Model	FLOPs	Parameters (M)	Precision (%)	Recall (%)	F1 (%)	Accuracy (%)	MAUC
Baseline	16.86	85.65	58.89	53.00	55.79	78.3	94.01
Baseline + DHA	4.42	27.65	73.34	70.50	71.89	83.26	96.02
Baseline + DCP	5.12	32.36	74.76	71.18	72.93	82.99	95.21
Baseline + DFG	2.02	11.52	75.40	71.69	73.37	82.52	95.40
Baseline + DHA + DCP + DFG	**2.75**	**16.82**	**83.24**	**77.25**	**78.88**	**86.12**	**96.54**

Note: Bolded data is the best under this indicator.

The ViT model achieves 58.89% precision, 53.00% recall, 55.79% F1 score, 78.3% accuracy, and 94.1% on the MAUC metric, with FLOPS of 16.86 G and 85.65 M parameters.

Adding the DCP module increases the FLOPS to 5.12 G and the number of parameters to 32.36 M. Although the precision, recall, and F1 score all improve, the overall accuracy decreases slightly to 82.99% and the MAUC drops to 95.21%, indicating that while DCP enhances feature extraction, the increased parameter count impacts model efficiency.

Introducing the DFG module significantly reduces the FLOPS to 2.02 G and number of parameters to 11.52 M while improving precision to 75.4%, recall to 71.69%, and F1 to 73.37%. While overall accuracy drops slightly, MAUC remains at 95.4%, demonstrating that the DFG module maintains performance while reducing resource usage.

Finally, integrating all the modules results in FLOPS of 2.75 G and parameters of 16.82 M. Precision increases to 83.24%, recall reaches 77.25%, F1-score rises to 78.88%, accuracy improves to 86.12%, and MAUC reaches 96.54%. This combination shows a synergistic effect, significantly boosting classification accuracy and robustness while reducing resource consumption.

3.3. Training Process Analysis

In Figure 7, we present the training and validation process of the model on the ISIC2018 and ISIC2019 datasets, including accuracy curves and loss curves.

As shown in Figure 7a,b, the accuracy curves on both the ISIC2018 and ISIC2019 datasets exhibit an overall upward trend. The accuracy improvement on the ISIC2019 dataset is more stable, whereas the accuracy curve on the ISIC2018 dataset exhibits noticeable fluctuations. These fluctuations can be attributed to the smaller size of the ISIC2018 dataset leading to greater variance in training, whereas the larger ISIC2019 dataset contributes to a more stable training process.

In Figure 7c,d, it can be seen that the loss curves decrease rapidly at the beginning of training before gradually converging. The validation loss curve on the ISIC2019 dataset closely follows the training loss curve, indicating a more stable training process. In contrast, the gap between the training and validation loss curves is relatively larger on the ISIC2018 dataset, which may be due to the limited dataset size and lower sample diversity.

Overall, the four subfigures in Figure 7 demonstrate that the proposed model is able to effectively learn the features and converge on both the ISIC2018 and ISIC2019 datasets. The larger size of the ISIC2019 dataset contributes to a more stable training process, with the loss curves of the training and validation sets maintaining more consistent downward trends. Meanwhile, the fluctuations in the ISIC2018 curves reflect the impact of dataset size on the training process. These results further validate the effectiveness of the proposed method, and indicate that increasing the dataset size can enhance both training stability and generalization performance.

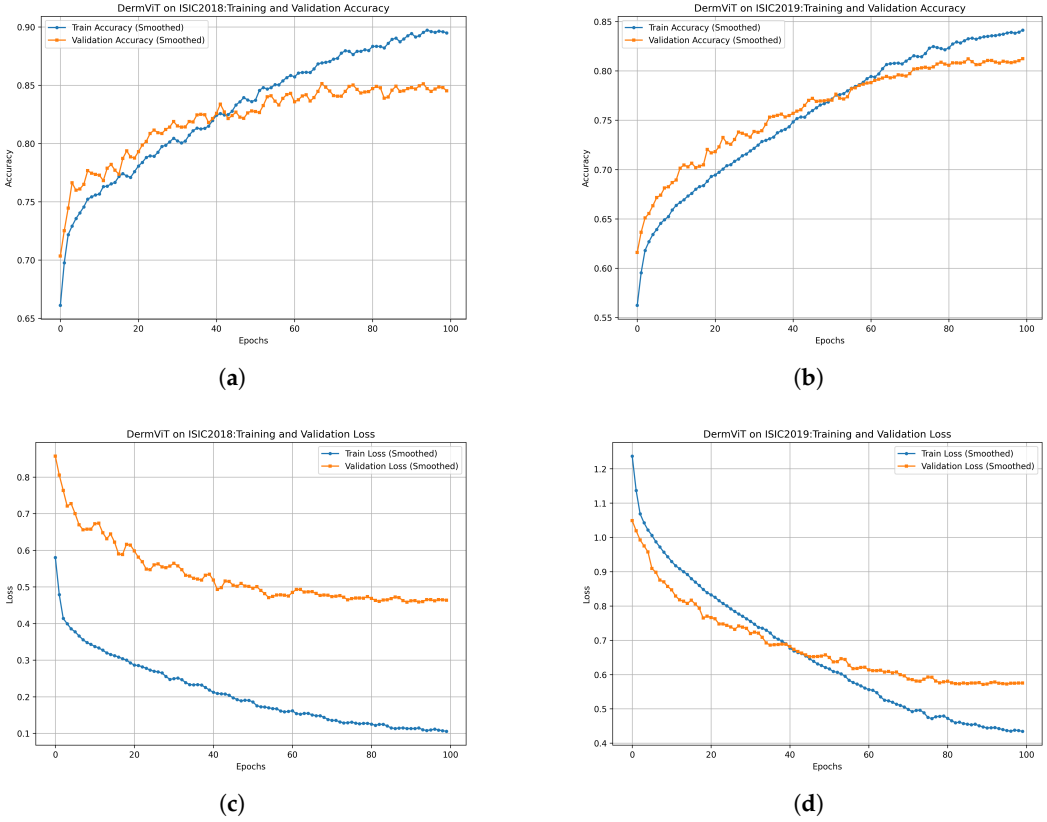


Figure 7. DermViT training and validation curves on the ISIC2018 and ISIC2019 datasets: (**a**) accuracy curves on ISIC2018, (**b**) accuracy curves on ISIC2019, (**c**) loss curves on ISIC2018, and (**d**) loss curves on ISIC2019.

3.4. Visualization

The classification of skin lesion images often faces significant challenges due to artifacts such as hair and measurement tools, which can diminish the model's attention and compromise diagnostic accuracy. To validate the clinical applicability of the DermViT model in scenarios with complex backgrounds, we conducted a visual analysis of lesion images containing typical interfering factors, as illustrated in Figure 8.

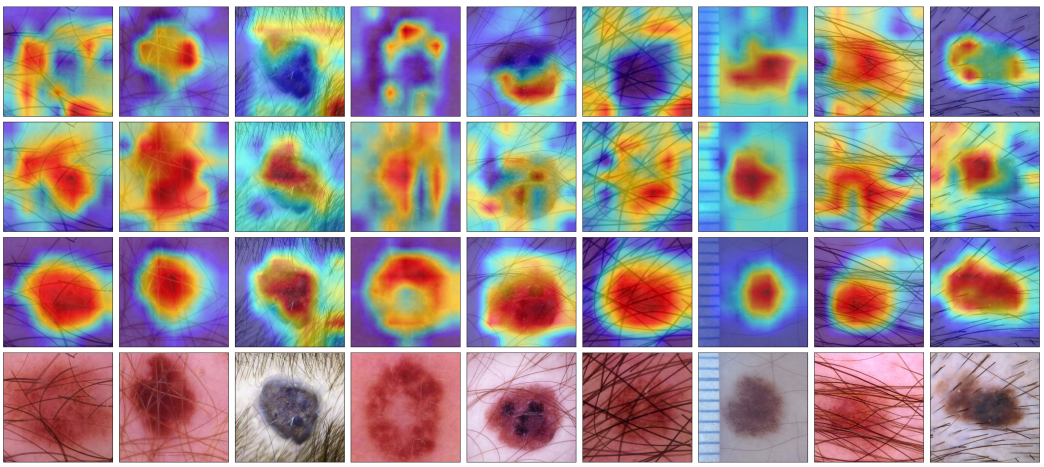


Figure 8. Grad-CAM visualization of images. In order from top to bottom, each line shows the results of ResNet50, Biformer, DermViT, and the original images.

The experimental results demonstrate that DermViT exhibits an attention distribution that is highly consistent with dermatologists compared to the ResNet50 and Biformer models, with its heatmaps precisely covering the core lesion areas. Specifically, in melanoma samples with hair artifacts, the ResNet50 model's attention is dispersed across the surrounding hair regions, whereas DermViT accurately focuses on the lesion area. This alignment with clinical diagnostic logic can be attributed to the synergistic mechanism formed by the DHA module's hierarchical feature guidance, the DCP module's cross-scale semantic association, and the gating mechanism of the DFG module.

The visualization results not only demonstrate the interpretability of the model's decision-making process but also effectively validate the feasibility of modeling dermatologists' diagnostic paradigms based on cognitive pathways.

4. Discussion

The DermViT model proposed in this paper achieves significant performance improvements in skin lesion classification tasks, validating the effectiveness of its medical-driven design. The success of DermViT highlights the importance of healthcare-driven architectural design in bridging the gap between technological advancements and clinical needs. By simulating the diagnostic workflow of dermatologists, DermViT not only achieves exceptional performance but also provides a more interpretable and trustworthy solution for skin lesion classification. This approach can be extended to other medical imaging tasks where domain-specific knowledge is crucial for accurate diagnosis, such as radiology and pathology [38].

Despite its notable advantages, DermViT still faces several challenges, particularly around addressing intra-class variability and inter-class similarity, which are both central issues in skin cancer diagnosis. Through an analysis of misclassified extreme cases from the test set (Figure 9), we identified specific limitations and potential areas for improvement.

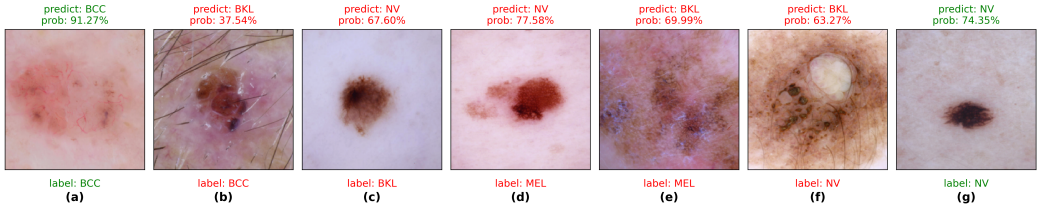


Figure 9. Example predictions of extreme cases from the ISIC2018 test set: (**a**) BCC (early), (**b**) BCC (late), (**c**) BKL (early), (**d**) MEL (early), (**e**) MEL (late), (**f**) NV (late), (**g**) NV (early).

Intra-Class Variability: The morphological characteristics of skin lesions such as melanomas can exhibit significant variations across different stages of progression. For example, early-stage BCC presents as a superficial lesion with indistinct borders, while late-stage BCC manifests as a nodular lesion with central ulceration. Although DermViT utilizes the DCP module to capture local texture information, it remains insufficiently sensitive to temporal morphological changes. The experimental results indicate that the model misclassified early- and late-stage BCC as distinct categories, highlighting its limited ability to model dynamic intra-class variations. This limitation is likely due to the model's inability to account for the temporal evolution of lesions. In clinical practice, dermatologists often incorporate patient history to refine diagnoses. Thus, future research could explore integrating time series data or dynamic feature modeling in order to enhance DermViT's ability to capture and adapt to intra-class variability.

Inter-Class Similarity: Inter-Class Similarity: The morphological overlap between different lesion categories (e.g., MEL and BKL) poses another significant challenge. For example, BKL and MEL both share visual characteristics that are similar to NV, such as

clear border pigmentation, which leads DermViT to misclassify them as NV. The model appears to rely too heavily on superficial visual similarity, failing to distinguish the essential pathological differences between BKL, MEL, and NV. This suggests that DermViT may struggle to differentiate categories with similar visual features but distinct underlying pathological mechanisms, especially when it comes to identifying deeper pathological features such as cellular atypia. Enhancing the model's ability to analyze this pathological substructure remains an ongoing challenge. Future work could explore higher-resolution feature encoding or incorporate pathological priors, which could improve the robustness of the model when handling such challenging cases.

Data Scarcity and Generalizability: While DermViT performs well on the ISIC datasets, its generalizability to other datasets and imaging modalities requires further validation. Future work could explore self-supervised or semi-supervised learning techniques [39,40] to further reduce reliance on large annotated datasets.

Computational Efficiency and Lightweight Design: Although DermViT demonstrates superior computational efficiency compared to other models, its parameter count needs further optimization in order to better adapt to mobile device deployment and enable efficient diagnosis in resource-constrained environments [41]. This is of significant practical value for regions with limited access to healthcare services [42].

Interpretability: While the attention mechanism in DermViT provides a degree of interpretability, more advanced tools could be developed to enhance clinicians' trust in the model [43]. Techniques such as saliency maps or explainable AI frameworks could be integrated to improve transparency [44].

Multimodal Fusion: Integrating additional data sources such as patient metadata or genomic data could further enhance diagnostic accuracy and enable personalized treatment recommendations. Multimodal approaches have shown promise in improving the robustness and applicability of medical AI systems [45,46].

By addressing these challenges, DermViT and similar models can continue to advance the field of medical imaging, offering more reliable, accessible, and interpretable diagnostic solutions for diverse clinical applications.

5. Conclusions

This paper presents DermViT, a medically-driven vision transformer tailored for skin lesion classification. DermViT tackles key challenges such as semantic entanglement, multi-scale feature variations, and computational inefficiency through three novel modules (DCP, DFG, and DHA) that mimic the diagnostic processes of dermatologists. This design enhances performance, interpretability, and robustness, making for significantly reduced misclassification against complex backgrounds. Experiments on the ISIC2018 and ISIC2019 datasets validate the effectiveness of DermViT, reaching 86.12% accuracy on ISIC2018 (a 7.8% improvement over ViT-Base) and strong generalization on ISIC2019. Additionally, DermViT maintains high accuracy while reducing the number of parameters compared to ViT, enhancing its suitability for real-time diagnosis in resource-limited settings. By integrating medical expertise with deep learning, DermViT offers a reliable high-performance solution for skin lesion classification.

Author Contributions: Conceptualization, Y.L. and X.Z. (Xuejun Zhang); methodology, Y.L.; software, Y.L. and G.O.; validation, Y.L., G.O. and W.C.; formal analysis, X.Z. (Xiangrong Zhou) and D.W.; investigation, Y.L. and A.X.; resources, G.O. and T.H.; data curation, D.W. and T.H.; writing—original draft preparation, Y.L.; writing—review and editing, Y.L. and X.Z. (Xuejun Zhang); visualization, Y.L.; supervision, X.Z. (Xuejun Zhang); project administration, X.Z. (Xuejun Zhang); funding acquisition, X.Z. (Xuejun Zhang). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Liuzhou Science and Technology Plan (Grant No. 2022SB019), the Guangxi Zhuang Autonomous Region Health Commission (Grant Nos. Z-B20231432 and Z20210069), and the Central Guidance for Local Science and Technology Development Funds Projects (Grant No. 202201002).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the International Skin Imaging Collaboration (ISIC) challenges at https://challenge.isic-archive.com/landing/2018/ (accessed on 2 May 2024) and https://challenge.isic-archive.com/landing/2019/ (accessed on 5 May 2024). The code will be available in the future at https://github.com/llyafei/DermViT (accessed on 1 April 2025).

Acknowledgments: We would like to thank RongQi Liu from the Guangxi University School of Computers, Electronics and Information for his assistance in setting up the experimental environment.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Balch, C.M.; Gershenwald, J.E.; Soong, S.J.; Thompson, J.F.; Atkins, M.B.; Byrd, D.R.; Buzaid, A.C.; Cochran, A.J.; Coit, D.G.; Ding, S.; et al. Final Version of 2009 AJCC Melanoma Staging and Classification. *J. Clin. Oncol.* **2009**, *27*, 6199–6206. [CrossRef]
2. Braun, R.P.; Rabinovitz, H.S.; Oliviero, M.; Kopf, A.W.; Saurat, J.H. Dermoscopy of pigmented skin lesions. *J. Am. Acad. Dermatol.* **2005**, *52*, 109–121. [CrossRef] [PubMed]
3. Steppan, J.; Hanke, S. Analysis of skin lesion images with deep learning. *arXiv* **2021**, arXiv:2101.03814.
4. Vestergaard, M.; Macaskill, P.; Holt, P.; Menzies, S. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: A meta-analysis of studies performed in a clinical setting. *Br. J. Dermatol.* **2008**, *159*, 669–676. [CrossRef] [PubMed]
5. Gouda, W.; Sama, N.U.; Al-Waakid, G.; Humayun, M.; Jhanjhi, N.Z. Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning. *Healthcare* **2022**, *10*, 1183. [CrossRef]
6. Hasan, M.K.; Ahamad, M.A.; Yap, C.H.; Yang, G. A survey, review, and future trends of skin lesion segmentation and classification. *Comput. Biol. Med.* **2023**, *155*, 106624. [CrossRef]
7. Celebi, M.E.; Kingravi, H.A.; Uddin, B.; Iyatomi, H.; Aslandogan, Y.A.; Stoecker, W.V.; Moss, R.H. A methodological approach to the classification of dermoscopy images. *Comput. Med Imaging Graph.* **2007**, *31*, 362–373. [CrossRef]
8. Abbas, Q.; Celebi, M.; Serrano, C.; Fondón García, I.; Ma, G. Pattern classification of dermoscopy images: A perceptually uniform model. *Pattern Recognit.* **2013**, *46*, 86–97. [CrossRef]
9. Goceri, E. Classification of skin cancer using adjustable and fully convolutional capsule layers. *Biomed. Signal Process. Control* **2023**, *85*, 104949. [CrossRef]
10. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
11. Harangi, B. Skin lesion classification with ensembles of deep convolutional neural networks. *J. Biomed. Inform.* **2018**, *86*, 25–32. [CrossRef] [PubMed]
12. Aldhyani, T.H.H.; Verma, A.; Al-Adhaileh, M.H.; Koundal, D. Multi-Class Skin Lesion Classification Using a Lightweight Dynamic Kernel Deep-Learning-Based Convolutional Neural Network. *Diagnostics* **2022**, *12*, 2048. [CrossRef] [PubMed]
13. He, X.; Tan, E.L.; Bi, H.; Zhang, X.; Zhao, S.; Lei, B. Fully Transformer Network for Skin Lesion Analysis. *Med. Image Anal.* **2022**, *77*, 102357. [CrossRef]
14. Gessert, N.; Sentker, T.; Madesta, F.; Schmitz, R.; Kniep, H.; Baltruschat, I.; Werner, R.; Schlaefer, A. Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. *IEEE Trans. Biomed. Eng.* **2019**, *67*, 495–503. [CrossRef] [PubMed]
15. Yu, Y.; Jia, H.; Zhang, L.; Xu, S.; Zhu, X.; Wang, J.; Wang, F.; Han, L.; Jiang, H.; Zhou, Q.; et al. Deep Multi-Modal Skin-Imaging-Based Information-Switching Network for Skin Lesion Recognition. *Bioengineering* **2025**, *12*, 282. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is All you Need. *arXiv* **2017**, arXiv:1706.03762.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.

18. Datta, S.K.; Shaikh, M.A.; Srihari, S.N.; Gao, M. Soft Attention Improves Skin Cancer Classification Performance. In *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data, Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2021; pp. 13–23. [CrossRef]
19. Li, A.; Zhang, D.; Yu, L.; Kang, X.; Tian, S.; Wu, W.; You, H.; Huo, X. Residual cosine similar attention and bidirectional convolution in dual-branch network for skin lesion image classification. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108386. [CrossRef]
20. Matsoukas, C.; Haslum, J.; Söderberg, M.; Smith, K. Is it Time to Replace CNNs with Transformers for Medical Images. *arXiv* **2021**, arXiv:2108.09038.
21. Wu, H.; Chen, S.; Chen, G.; Wang, W.; Lei, B.; Wen, Z. FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. *Med. Image Anal.* **2022**, *76*, 102327. [CrossRef]
22. Parvaiz, A.; Khalid, M.; Zafar, R.; Ameer, H.; Ali, M.; Fraz, M. Vision Transformers in Medical Computer Vision—A Contemplative Retrospection. *Eng. Appl. Artif. Intell.* **2023**, *122*, 106126. [CrossRef]
23. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R.W. BiFormer: Vision Transformer With Bi-Level Routing Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
24. Wang, S.; Li, B.; Khabsa, M.; Han, F.; Ma, H. Linformer: Self-Attention with Linear Complexity. *arXiv* **2020**, arXiv:2006.04768.
25. Madan, S.; Li, Y.; Zhang, M.; Pfister, H.; Kreiman, G. Improving generalization by mimicking the human visual diet. *arXiv* **2022**, arXiv:2206.07802.
26. Huang, L.; Yuan, Y.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. Interlaced Sparse Self-Attention for Semantic Segmentation. *arXiv* **2019**, arXiv:1907.12273.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
28. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
29. Rezvantalab, A.; Safigholi, H.; Karimijeshni, S. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms. *arXiv* **2018**, arXiv:1810.10348.
30. Foahom Gouabou, A.C.; Damoiseaux, J.L.; Monnier, J.; Iguernaissi, R.; Moudafi, A.; Merad, D. Ensemble Method of Convolutional Neural Networks with Directed Acyclic Graph Using Dermoscopic Images: Melanoma Detection Application. *Sensors* **2021**, *21*, 3999. [CrossRef]
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
32. Fan, Q.; Huang, H.; Chen, M.; He, R. Vision Transformer with Sparse Scan Prior. *arXiv* **2024**, arXiv:2405.13335.
33. Codella, N.; Rotemberg, V.; Tschandl, P.; Çelebi, M.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *arXiv* **2019**, arXiv:1902.03368.
34. Onakpojeruo, E.P.; Mustapha, M.T.; Ozsahin, D.U.; Ozsahin, I. A Comparative Analysis of the Novel Conditional Deep Convolutional Neural Network Model, Using Conditional Deep Convolutional Generative Adversarial Network-Generated Synthetic and Augmented Brain Tumor Datasets for Image Classification. *Brain Sci.* **2024**, *14*, 559. [CrossRef]
35. Onakpojeruo, E.P.; Mustapha, M.T.; Ozsahin, D.U.; Ozsahin, I. Enhanced MRI-based brain tumour classification with a novel Pix2pix generative adversarial network augmentation framework. *Brain Commun.* **2024**, *6*, fcae372. [CrossRef]
36. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary PhD Workshop (IIPhDW), Swinoujscie, Poland, 9–12 May 2018; pp. 117–122. [CrossRef]
37. Shen, S.; Xu, M.; Zhang, F.; Shao, P.; Liu, H.; Xu, L.; Zhang, C.; Liu, P.; Yao, P.; Xu, R. A low-cost high-performance data augmentation for deep learning-based skin lesion classification. *BME Front.* **2022**, *2022*, 9765307. [CrossRef] [PubMed]
38. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
39. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709.
40. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
41. Howard, A.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.

42. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
43. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?". In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [CrossRef]
44. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
45. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [CrossRef]
46. Azam, M.A.; Khan, K.B.; Salahuddin, S.; Rehman, E.; Khan, S.A.; Khan, M.A.; Kadry, S.; Gandomi, A.H. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med.* **2022**, *144*, 105253. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.