

Article

SCGViT: A Pseudo-Multimodal Low-Latency Framework for Real-Time Skin Lesion Diagnosis

Zirui Luo ¹, Chengyu Hou ² and Haishi Wang ^{1,*} 

¹ College of Communication Engineering, Chengdu University of Information Technology, Chengdu 610225, China; 3230404005@stu.cuit.edu.cn

² College of Atmospheric Sciences, Chengdu University of Information Technology, Chengdu 610225, China; 5250101019@stu.cuit.edu.cn

* Correspondence: whs@cuit.edu.cn

Abstract

In order to solve the problems of insufficient medical image feature extraction, high classification accuracy, and computational complexity in automatic diagnosis of skin lesions in the edge computing environment, this paper proposes a real-time pseudo-multimodal low-delay diagnosis framework, SCGViT, based on a vision transformer. The framework is constructed around three functional objectives: mitigating data imbalance through generative modeling, capturing diverse representations via multi-dimensional perception, and optimizing feature fusion through adaptive refinement. Firstly, using Class-Conditioned Generative Adversarial Networks (CGANs) simulates manifolds of minority class samples in latent space, achieving preliminary balance of data distribution. Secondly, a branch feature-extraction path is constructed to simulate inversion (INV) and infrared (IR) modes in the original visual primary color mode (RGB), in order to achieve multi-dimensional perception. Finally, a cross-attention mechanism is combined for cross-branch feature aggregation, and a channel-attention mechanism (squeeze and excitation) is embedded for secondary refinement of the mixed global local features to enhance the representation ability of key pathological regions by integrating complementary structural and contrast information. The experimental results on the HAM10000 dataset showed that the F1 score reached 0.973, the inference speed reached 304.439 FPS, the parameter count was only 0.524 M, and the computational complexity was only 0.866 G FLOPs, achieving a balance between high accuracy and light weight.

Keywords: vision transformer; skin lesion; multi-dimensional perception; channel-attention mechanism



Academic Editor: Jichai Jeong

Received: 2 January 2026

Revised: 5 February 2026

Accepted: 14 February 2026

Published: 16 February 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Skin cancer is one of the most common malignant tumors globally, with over 3 million new cases reported annually. The 5-year survival rate for malignant melanoma patients is over 99% in the early stages, but only 15% in advanced stages, making accurate early diagnosis crucial for improving patient survival rates [1]. Traditional diagnosis relies on visual examination by dermatologists and pathological biopsy, which is time-consuming, labor-intensive, and limited by the doctor's experience level; the shortage of professional dermatologists in primary healthcare institutions and remote areas further exacerbates diagnostic delays [2]. With the development of portable medical devices and edge computing technology, real-time on-site diagnosis of skin lesions has become a promising solution. Edge computing supports local data processing, avoiding the privacy risks of

uploading sensitive medical data to the cloud, while also reducing reliance on network bandwidth [3]. However, edge devices have inherent constraints, including limited computing power, memory, and battery capacity, requiring diagnostic models to be lightweight, low-latency, and energy-efficient. The most commonly used input for existing models is single-modal RGB images, but RGB images are susceptible to variations in lighting, similarities in lesion morphology, and color distortion, leading to insufficient feature representation and poor generalization ability. Multimodal data [4], such as RGB, IR, and INV, can provide complementary lesion information and improve diagnostic robustness. However, existing multimodal fusion methods, such as early concatenation and late fusion, either ignore inter-modal correlations or have high computational complexity. Lightweight visual transformers combine the local feature-extraction capabilities of convolutional neural networks (CNNs) with the global feature modeling advantages of transformers, achieving a balance between performance and efficiency. CGANs have shown significant potential in medical image enhancement, effectively mitigating data scarcity problems [5]. However, how to integrate these technologies into a unified framework to achieve edge-friendly multimodal skin lesion diagnosis remains an unresolved challenge. This article proposes the SCGViT model, a pseudo-multimodal low-latency framework for real-time skin lesion diagnosis, demonstrating high computational efficiency, making it highly compatible with the resource constraints of edge-based hardware. The main contributions are as follows:

1. This article extends the MobileViT architecture to pseudo-multimodal inputs and designs a branch-specific feature extractor with learnable weights. This design achieves collaborative modeling of local texture and global long-range dependencies of skin lesions while maintaining low computational complexity (with only 0.524 M parameters and 0.866 G FLOPs), achieving high inference efficiency.
2. The added cross-attention module enhances directional features through asymmetric fusion strategy. This module guides the effective injection of auxiliary branch IR and INV information through RGB visual features. Compared with traditional simple concatenation or weighted summation, it can more accurately capture complementary pathological information between different branches.
3. The framework, as a multitasking structure, introduces a self-supervised generator to perform adversarial training. By generating interference samples, the classifier is forced to learn more fundamental pathological features, reducing reliance on surface artifacts.
4. Auxiliary classification heads are deployed at the end of each branch extractor, enforcing the discriminative ability of single branch features through deep supervision loss.

2. Related Work

Early skin lesion diagnosis systems primarily relied on traditional CNN architectures, such as ResNet, Inception, and DenseNet. These models achieved significant accuracy on standard HAM10000 or ISIC datasets by learning the color, texture, and structural features of images. However, these models are typically designed for large-scale servers, with a huge number of parameters and high inference latency. To improve the models' ability to capture global contextual information, vision transformers (ViTs) and their variants have been introduced into the field of medical imaging.

Since its introduction, MobileViT has been applied to skin disease diagnosis and detection by researchers worldwide. This paper summarizes some of the published research results in skin lesion image classification by researchers in this field. In 2017, a joint team from the Department of Electrical Engineering and Dermatology at Stanford University, USA [6], published research in *Nature* detailing the training of a model based on a deep CNN using 129,000 clinical images covering 2032 diseases. In binary classification tasks

of keratinocyte carcinoma vs. benign seborrheic keratosis, and malignant melanoma vs. benign nevus, they achieved diagnostic performance comparable to 21 certified dermatologists, demonstrating, for the first time, the great potential of deep learning in automatic skin cancer classification. In 2022, research in this field focused on model architecture optimization and attention-mechanism integration. The Department of Electronic Design at Mid Sweden University and the Department of Industrial Engineering at the University of Salerno, Italy [7] collaborated to propose a CNN–transformer hybrid model, combining the focal loss functions to alleviate the class imbalance problem in the ISIC 2018 dataset, effectively improving the generalization ability of skin lesion classification; the School of Mechanical Engineering at Quzhou University and the School of Mechanical Engineering at Hangzhou Dianzi University, China [8], proposed SEACU-Net, integrating squeeze-and-excitation (SE) layers and attention ConvLSTM into the U-Net architecture, optimizing the accuracy of skin lesion segmentation; the School of Information Science and Engineering at Yunnan University [9] enhanced the feature-extraction capabilities of lesion images through the combination of dense dilated spatial pyramid pooling and attention mechanisms; and the School of Information Engineering and Automation at Kunming University of Science and Technology [10] proposed a classification model based on the DSception module and attention mechanism, providing a new feature-extraction approach for skin disease recognition. From 2023 to 2025, research further expanded to model fusion and multimodal technologies. A team proposed a fusion model based on DenseNet201 and ConvNeXt_L [11], incorporating efficient channel-attention modules, achieving accuracies of 96.54% and 95.29% on a private acne dataset and the public HAM10000 dataset, respectively. Another team constructed an integrated model of EfficientNetV2S and Swin-Transformer [12] achieving an accuracy of 99.10% on the HAM10000 dataset after data augmentation. The SkinSavvy2 system developed by Skin Savvy [13] in the United States integrates EfficientNetV2, Swin Transformer, and GPT models, combining personalized factors to provide classification results and care recommendations, enhancing the clinical applicability of the technology [14].

Recent research in dermoscopic image analysis has increasingly shifted toward resource-efficient transformer architectures. To capture intricate pathological details, Xu et al. [15] introduced Skinformer, which learns statistical texture representations specifically for skin lesion analysis. To address the computational bottlenecks of traditional Vision Transformers, Mehta and Rastegari [16] proposed MobileViTv2, leveraging separable self-attention to achieve high diagnostic performance on mobile-class hardware. Furthermore, Touvron et al. [17] demonstrated that distillation-based structural optimizations can significantly improve data efficiency, a crucial factor when training transformers on specialized medical datasets. The challenge of class imbalance in dermatological datasets has prompted the development of sophisticated generative and augmentation techniques. You et al. [18] proposed class-aware generative transformers to stabilize decision boundaries in long-tailed medical distributions. Complementing these architectural gains, Khan et al. [19] developed a mobile-friendly transformer specifically for skin lesion diagnosis, while Su et al. [20] introduced a GAN-based data augmentation method to mitigate the scarcity of minority lesion classes in multi-class classification tasks. To ensure feasibility for on-device medical diagnostics, recent studies have focused on robustness and extreme parameter efficiency. Zhang et al. [21] developed DermViT, a diagnosis-guided vision transformer tailored for efficient skin lesion classification. To further optimize models for edge deployment, Walczak et al. [22] proposed BitMedViT, which utilizes ternary quantization to enable medical AI assistants on resource-constrained devices. Finally, comparative analyses of hybrid transformer–CNN frameworks [23] have highlighted their ability to balance local feature extraction with global context modeling on the HAM10000 benchmark.

Although the above studies have made progress in their respective fields, there is currently a lack of a unified framework that can simultaneously address the triple challenges of edge deployment efficiency, multimodal complementary fusion, and feature robustness. High-performance ViT models have high computational costs, while lightweight models are usually limited to single-modal input or employ simple fusion mechanisms. Therefore, this paper proposes SCGViT, a low-latency diagnostic system that integrates the lightweight MobileViT, learnable cross-attention fusion, and CGAN-enhanced multi-task learning into a single framework.

3. Proposed Framework

3.1. Overview of the Overall Architecture

The proposed SCGViT framework aims to solve the two bottlenecks faced by automatic diagnosis of skin lesions in the edge computing environment: (1) the data scarcity and long-tail distribution that compromise the model's sensitivity to minority classes and (2) the representational insufficiency of unimodal RGB data, which struggles to encapsulate the multifaceted characteristics of pathological structures.

MobileViT has shown a good ability to achieve a balance between detection accuracy and computational efficiency in single model recognition. MobileViT adopts a multi-stage feature-extraction structure, which includes a stem layer, multi-stage feature-extraction module, global pooling layer, and classification layer from input to output, providing a solid foundation for solving lightweight and low-latency issues. On the basis of this architecture, SCGViT introduces CGANs, SE, and cross attention to systematically implement three key aspects: pseudo-multimodal input and feature extraction, adversarial training, and asymmetric fusion strategy for directional feature enhancement. As shown in Figure 1, the SCGViT framework consists of three interconnected stages.

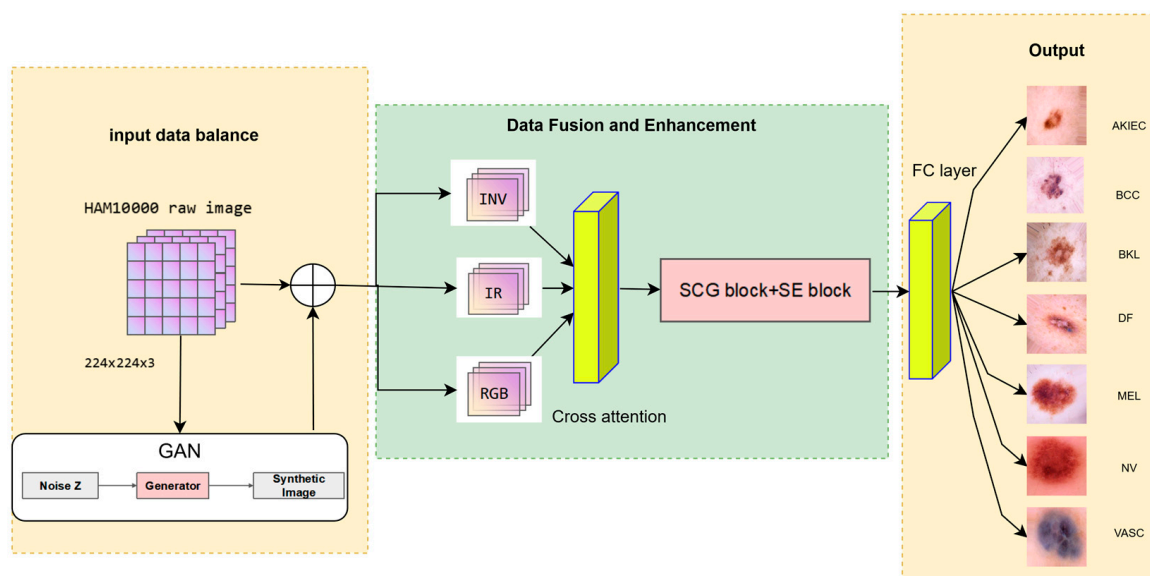


Figure 1. Overall architecture of the proposed SCGViT framework for skin lesion classification.

Firstly, the input image is processed through the CGAN module and then combined with the original image for preliminary image sample equalization. Secondly, three branch feature-extraction paths are constructed to simulate the original visual RGB, INV, and IR, and cross attention is used to achieve cross data information exchange. Finally, SCG block and SE block are used to enhance the representation ability of key pathological regions.

3.2. Residual CGAN for Generative Data Balancing

The HAM10000 dataset exhibits a significant “long-tail” distribution [15], where the majority classes tend to dominate the loss function, potentially leading to biased classification. To address this issue, we designed a generative model based on residual learning. Unlike traditional over-sampling or focal loss, which often lead to overfitting by repeatedly presenting identical minority samples or overemphasizing hard pixels, our CGAN-based approach generates novel feature representations by simulating the latent manifold of minority classes. This mechanism serves as a form of implicit data augmentation that enhances the structural diversity of the training set, rather than merely adjusting class-specific weights. The detailed structure is shown in Figure 2.

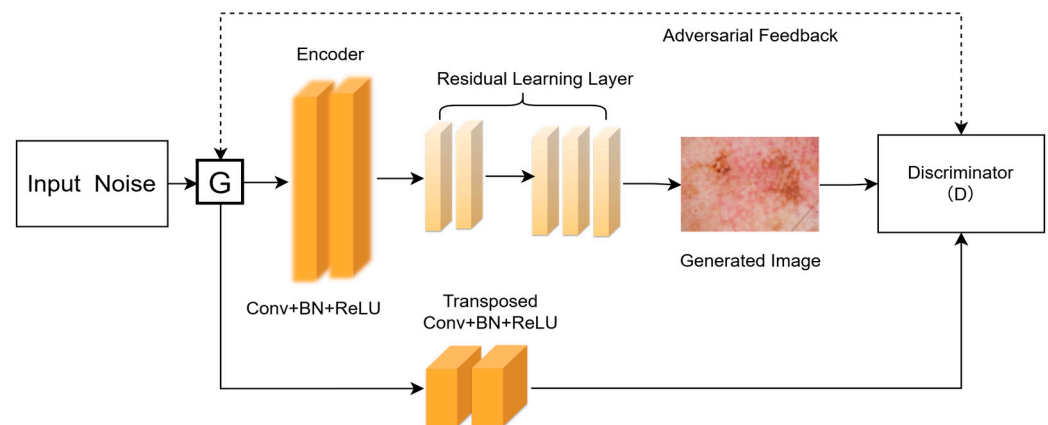


Figure 2. Architecture of the proposed residual CGAN. The dashed arrows represent the adversarial feedback path from the discriminator to the generator.

Firstly, there is noise input and initial encoding. Starting from random noise (input noise), feature mapping is performed through the encoder of the generator (G). The encoder is composed of three layers stacked to convert one-dimensional noise vectors into high-dimensional feature tensors. At the same time, convolutional layers are used to capture local patterns, the BN layer is stably trained, and the ReLU layer introduces nonlinearity to provide basic feature expressions for subsequent generation. Next is residual learning and feature reconstruction. The encoded features enter the residual learning layer and are upsampled through transpose convolution (restoring spatial dimensions). At the same time, residual connections preserve shallow details (such as texture and boundaries of skin lesions) to avoid information loss in deep networks, gradually reconstructing high-dimensional features into feature maps that match the size of real skin lesion images. Finally, the reconstructed feature map is output and connected to the discriminator (D) for true false judgment. The output of the discriminator is then passed back to the generator, forcing it to continuously optimize the parameters of encoding, residual learning, and other modules. Ultimately, the generator generates samples that are consistent with the distribution of real skin lesion images, achieving data augmentation to alleviate the class imbalance problem in the skin lesion dataset. To ensure precise control over the balancing process, we employ a CGAN architecture. The number of synthetic samples generated for each minority class is determined by its specific imbalance degree: 500 for akiec, 600 for bcc, 300 for bkl, 800 for df, 400 for mel, 200 for nv, and 700 for vasc. To maintain the causal integrity of feature learning, we implement a phased training schedule, the first 10 epochs utilize only real samples to establish a reliable baseline of clinical features. Subsequently, synthetic and real samples are integrated per batch at a fixed mixing ratio of 3:7. Furthermore, a real-time artifact filtering mechanism is applied, where any synthetic sample with a quality score (derived from the discriminator) below 0.7 is discarded. We also use the Inception score to ensure

that the distribution similarity remains high and that visible artifacts are controlled within a 10% threshold. The generation results are visualized in Figure 3. Figure 3 shows that parts of synthetic samples generated by our CGAN maintain high morphological similarity to authentic clinical images across multiple classes.

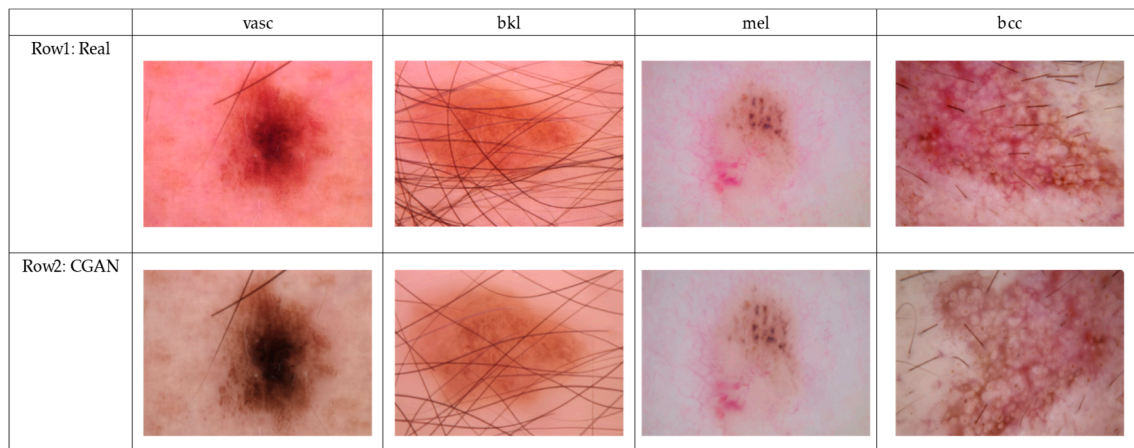


Figure 3. Representative examples of real vs. CGAN-generated skin lesion images.

Furthermore, as demonstrated in Figure 4, our real-time filtering mechanism successfully identifies and excludes low-quality samples with visible artifacts (scores < 0.7). This rigorous quality control ensures that the augmented dataset contributes genuine pathological features rather than distribution-driven noise.

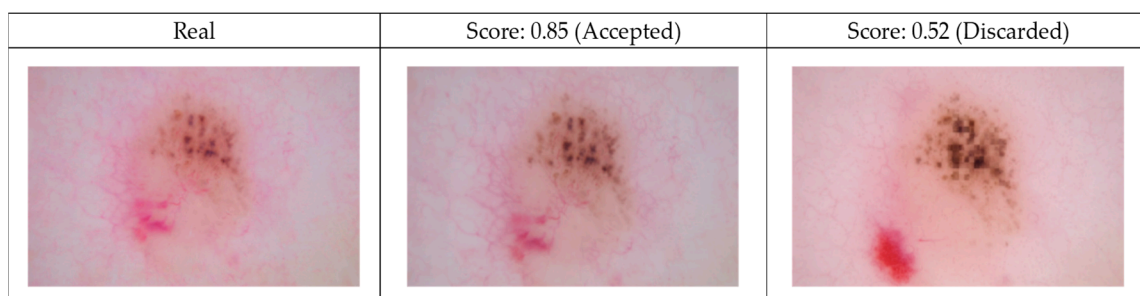


Figure 4. Visual comparison of real dermoscopic image and CGAN-generated samples with different quality scores.

The equilibrium process of overall generation can be described as a game problem, and its objective function L_{GAN} is defined as follows:

$$\min_G \max_D V(D, G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

where $V(D, G)$ is the value function that quantifies the game state between the two, $D(x)$ is the output of the discriminator on the real sample x , x is the real image, and z is the noise vector that follows the prior distribution p_z . Unlike traditional direct generation methods, the core of residual learning is to let the network only learn the difference between the synthesized image and the reference distribution. Residual blocks pass shallow features to deep layers through identity mapping, and their output X_{out} can be expressed as

$$X_{out} = \sigma(BN(Conv(\sigma(BN(Conv(x_{in})))))) + X_{in} \quad (2)$$

where X_{in} is the input feature map of the module, while X_{out} is the output feature map of the module. Conv is the convolution operation and BN is batch normalization used

to accelerate training, suppress overfitting, and standardize convolutional outputs. The schematic diagram is shown in Figure 5.

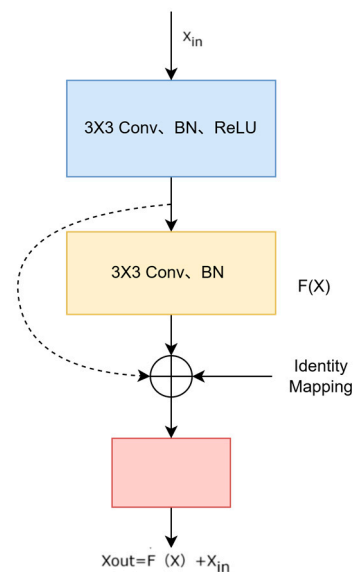


Figure 5. Detailed structure of the residual block. The dashed arrow represents the shortcut connection that performs identity mapping, allowing the original input information to be directly added to the output of the convolutional layers.

The input of the module is the feature tensor X_{in} , which is first processed through the 3×3 Conv, BN layer, and ReLU layer. After the first layer, the residual feature $F(X)$ is obtained through 3×3 convolution and BN. Identity mapping directly passes the original input X_{in} to subsequent layers, and then adds the residual feature element-by-element to the original input. Finally, ReLU activation is performed to obtain the output X_{out} of the module. By utilizing the difference between input and output, the details of the original input are preserved, while making it easier for the gradient of the deep network to be transmitted, avoiding slower parameter updates as the depth increases during training.

3.3. Multi-Branch Cross-Attention Fusion Mechanism

A single RGB image often fails to provide sufficient differentiation criteria when facing complex lighting, hair occlusion, or lesions with colors close to the skin color. To address this, we define the input space as a pseudo-multimodal configuration. The primary source is RGB dermoscopy images, we decompose them into three different feature representations: RGB, INV, and a luminance-based IR-simulated branch, to simulate a multi-perspective feature-extraction environment. This approach enables the model to learn complementary features that are typically coupled within a single RGB channel. Specifically, INV branch (F_{INV}) is generated by the linear transformation:

$$F_{INV} = L_{MAX} - F_{RGB} \quad (3)$$

where L_{MAX} represents the maximum intensity level of the color space, and F_{RGB} denotes the original pixel intensity vector. In clinical dermatology, the INV representation enhances the visual contrast between the lesion's morphological structures and the peripheral healthy tissue. By mapping the original pixels to their logical opposites, the F_{INV} branch effectively highlights the negative space of the lesion, assisting the model in capturing critical geometric features such as asymmetry and border irregularity that may be obscured by dark pigmentation in the RGB channel. The IR branch (F_{IR}) simulates the high-penetration char-

acteristics of infrared light by extracting the luminance intensity from the long-wavelength components, calculated as

$$F_{IR} = \alpha \cdot R + \beta \cdot G + \gamma \cdot B \quad (4)$$

where R , G , and B represent the pixel intensities of the red, green, and blue channels, respectively. The parameters α equals 0.298, β equals 0.587, and γ equals 0.114 are fixed weighting coefficients. IR branch (F_{IR}) is a simulated intensity representation derived from RGB channels to emphasize structural information. By applying the specified weighting coefficients (α , β , and γ), this branch serves as a structural texture enhancer. It effectively suppresses surface color variations and highlights consistent morphological patterns across different lighting conditions, providing the model with a filtered view of the lesion's internal structural density. By explicitly defining F_{INV} and F_{IR} , we force the model to perform “cross-branch alignment.” The INV branch focuses on boundary geometry, while the IR branch focuses on internal structural density. This enables the subsequent cross-attention mechanism to dynamically weigh which “modality” is more reliable for a specific lesion.

To this end, a multi-branch cross-attention fusion mechanism is added, and its detailed structure is shown in Figure 6.

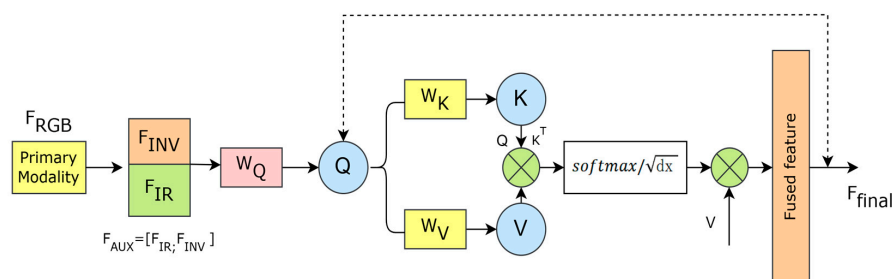


Figure 6. Detailed workflow of the multi-branch cross-attention fusion. The dashed arrows represent the cross-attention mechanism between different feature branches, illustrating the information interaction and fusion path.

The module takes the RGB main mode feature F_{RGB} as the core, first mapping it to the query vector Q through the projection matrix W_Q ; at the same time, auxiliary mode features such as infrared F_{IR} and inverse F_{INV} are concatenated into F_{Aux} , and then mapped to the key vector K and value vector V through W_K and W_V , respectively; then, the similarity between Q and K is calculated and the attention weight is obtained through Softmax, and V is weighted and summed to finally output the feature F_{final} that integrates the core information of the main mode and the complementary information of the auxiliary mode. Through this process, dynamic allocation of different branch weights is achieved: when the RGB image boundary is blurred, the attention mechanism is designed to adaptively redistribute weights, prioritizing discriminative feature maps that contribute significantly to the classification task. This semantic-guided fusion approach is more effective in suppressing noise and extracting complementary information across branches. The formula for its space is

$$Q = W_Q F_{RGB} \quad (5)$$

where the Q cross-attention query vector represents the information retrieval requirements of the main modality, W_Q is a learnable projection matrix used to map the main branch features to the query vector Q , and F_{RGB} is the RGB main branch feature, which is the core reference benchmark for fusion.

$$K = W_K F_{AUX} \quad (6)$$

where K is the key vector of cross attention, representing the information label of the auxiliary branch, used to calculate the correlation with Q ; W_K is a learnable projection

matrix used to map auxiliary branch features into key vectors K ; and F_{Aux} is an auxiliary branch concatenation feature.

$$V = W_V F_{Aux} \quad (7)$$

where V is the value vector of cross attention, representing the information content of the auxiliary branch and W_V is a learnable projection matrix used to map auxiliary branch features into a value vector V .

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

where attention (Q, K, V) is the output feature of cross attention, which is an intermediate feature that combines the information of the main and auxiliary branches, QK^T is the dot product of Q and K , used to calculate the information correlation between the main and auxiliary branches, and d_k is the dimension of the key vector K .

$$F_{final} = \sum_{m \in \{RGB, IR, Ino\}} \alpha_m \cdot F_m \quad (9)$$

where The final multi-branch fusion feature of F_{final} integrates key information from the primary and auxiliary branches and α_m is the attention weight of the m th branch, representing its importance level, the original features of the m th mode of F_m .

3.4. SCG-SE Blocks: Feature Recalibration and Global Modeling

In order to meet the real-time requirements of edge devices, this paper adopts a hybrid feature modeling scheme and embeds a channel-type attention mechanism in it. In order to coordinate local and global features, we first use a 3×3 depth separable convolution to obtain spatial local biases. Subsequently, the feature map is flattened into a sequence and long-range dependencies between pixels are constructed using the self-attention mechanism of the transformer architecture. Then, the local and global features are effectively concatenated in the channel dimension through 1×1 convolution. This hybrid design ensures that the model can recognize both small pathological spots and perceive the macroscopic distribution of lesions throughout the entire skin area. At the end of the feature stream, we introduce a weight allocation module based on feature reweighting—the adaptive channel-attention mechanism (SE). This module first compresses the feature map in spatial dimension, and through a global average pooling layer, compresses the feature map U in spatial dimension $H \times W$ into a vector z , capturing the global distribution of each channel, as shown in Figure 7.

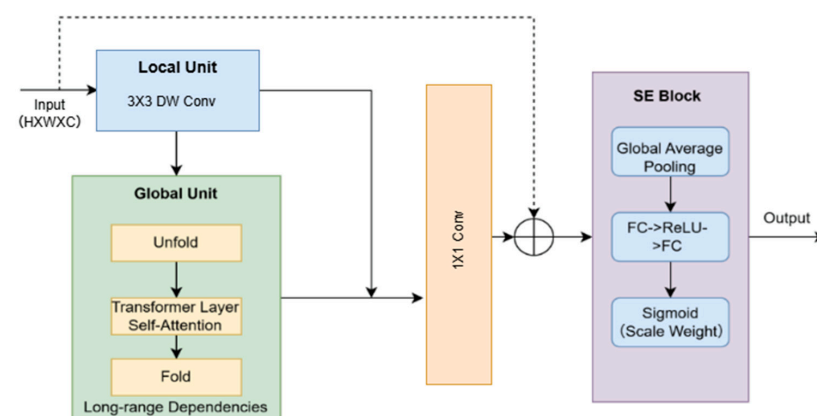


Figure 7. Internal structure of the SCG-SE blocks. The dashed arrow represents the residual skip connection that facilitates the integration of original input features with the recalibrated channel weights.

This architecture achieves feature refinement through three stages. Firstly, local spatial biases are obtained through deep convolution, global long-range dependency modeling is performed through the transformer, and final channel recalibration is achieved through squeeze-and-excitation (SE) mechanisms. The output of the above three stages achieves extremely low weight coefficients for channels with little diagnostic significance. On the contrary, for channels rich in pathological features, signal amplification will be performed. This process of feature selection and enhancement greatly improves the feature representation performance of the model without significantly increasing computational resources.

4. Experiments

4.1. Datasets

This article uses the HAM10000 dataset, which was jointly released by the Medical University of Vienna and the University of Queensland in 2018 [24]. It is a benchmark dataset for the diagnosis of pigmentary skin lesions [25]. Due to its large scale, authoritative labeling, and clinical relevance, it is widely used in skin lesion analysis and machine-learning research [26]. The distribution of its dataset is shown in Figure 8. This dataset has been used to analyze commonly used basic models, including ResNet18, CNN, Vit-Tiny, MobileViT [27], OverLoCK [28], H-CAST [29], GhostNetV2 [30], and MobileNetV3-S [31].

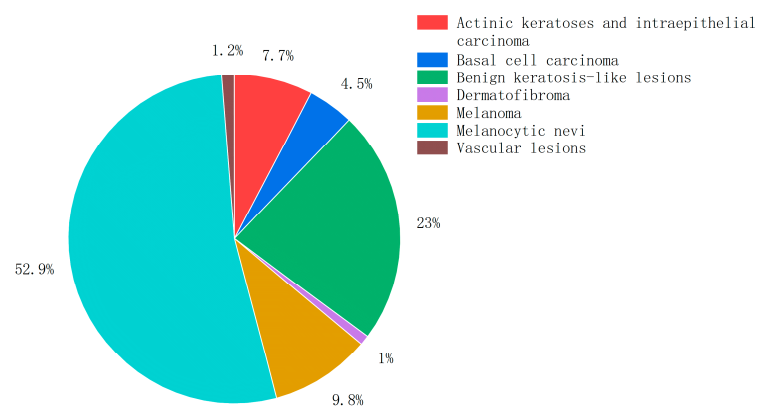


Figure 8. Pie chart showing the category distribution of the HAM10000 medical image dataset.

This dataset contains 10,015 dermatoscopy images corresponding to approximately 7000 independent lesions, covering 7 clinically important skin lesions with uneven sample distribution. Among them, nv has the highest proportion, followed by mel, while df and vasc have fewer samples [32].

4.2. Implementation Details

All experiments were conducted on a workstation equipped with an AMD EPYC 7532 CPU (32 cores) and an NVIDIA RTX 4090 GPU (24 GB VRAM). The software environment included Python 3.11, CUDA 12.1.1, and PyTorch 2.2.1. To ensure the scientific rigor of the experimental results and thoroughly avoid potential data leakage risks, the HAM10000 dataset was divided into 70% for training, 15% for validation, and 15% for testing, strictly following the lesion_id division principle. Specifically to prevent distribution leakage or performance inflation, all CGAN-based synthetic samples were generated and incorporated only within the training set. We ensured that no synthetic data derived from a specific lesion_id was present in the validation or testing sets. All CGAN configurations, including generation quantities and mixing ratios, were explicitly defined to ensure full reproducibility of the reported results. During the training phase, the resolution of the input images was uniformly adjusted to 224×224 pixels. We adopted the Adam optimizer, set the initial learning rate to 0.0001, and set the batch size to 32. The model underwent 250 epochs

to ensure sufficient convergence, while introducing an early stopping mechanism with a patience value set to 50 to prevent overfitting while maintaining the model's generalization ability. To assess the practical viability for clinical deployment, we measured the end-to-end inference latency. Following standard benchmarking protocols, we implemented a 100-run warm-up phase to eliminate hardware initialization bias. During testing, the batch size was strictly set to 1, accurately reflecting the sequential nature of real-time diagnostics. Our reported latency (3.28 ms on RTX 4090) accounts for the entire pipeline, including image resizing, normalization, model forward pass, and post-processing. This high throughput suggests the framework is well-suited for integration into edge-based diagnostic platforms.

4.3. Evaluation Metrics

We use various metrics to evaluate model performance, including precision (P), recall (R), F1-score, macro-averaged F1, number of parameters (M), floating-point operations (FLOPs, G), and inference speed (FPS). These metrics are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

where TP refers to samples that are actually positive and are also predicted as positive by the model, meaning the model made a correct prediction; FN refers to samples that are actually positive but are predicted as negative by the model, meaning the model missed a positive case; and FP refers to samples that are actually negative but are predicted as positive by the model, meaning the model made a false positive prediction.

$$\text{Macro-F1} = \frac{1}{n} \sum_{i=0}^{n-1} F1_i \quad (13)$$

First, calculate the F1 score for each category, and then take the arithmetic mean of the F1 scores across all categories, where n is the total number of categories.

Precision (P), recall (R), F1-score, and macro-averaged F1 are fundamental detection performance metrics. Precision quantifies the proportion of correctly predicted positive samples among all samples predicted as positive, while recall measures the proportion of actual positive samples that are correctly detected. F1-score and macro-averaged F1 provide a more intuitive representation of the model's performance. The number of parameters (Parameters, M), floating-point operations (FLOPs, G), and inference speed (Inference Speed, FPS) are efficiency metrics of the model. The number of parameters directly determines the storage space requirements for the model weights and is a key indicator of the model's "lightweight" nature; floating-point operations reflect the computational complexity of the algorithm, with lower FLOPs generally indicating lower computational cost; and inference speed is the most intuitive metric for measuring the real-time performance of the model, with higher FPS indicating stronger real-time processing capabilities.

4.4. Comparative Experiments on HAM10000 Dataset

To comprehensively validate the effectiveness of the SCGViT multimodal model proposed in this paper, we selected seven representative architectures as comparison benchmarks. The tested models included ResNet18, CNN, ViT-Tiny, MobileViT, OverLoCK, H-CAST, GhostNetV2, and MobileNetV3-S.

The macro-averaged F1-score was used as the core evaluation metric to objectively measure the overall performance of the models on datasets with severe class imbalance. In addition, the end-device deployment performance of the models was evaluated using parameters, FLOPs, and inference speed (FPS). All models were trained under consistent experimental conditions. The experimental results are shown in Table 1.

Table 1. Comparison of different models.

Model Name	Macro-Averaged F1 Score	Number of Parameters (M)	FLOPs (G)	Inference Speed (FPS)
CNN	0.901	1.161	1.6763	347.392
ResNet18	0.902	11.180	1.823	346.998
ViT-Tiny	0.908	2.526	0.495	341.183
SCGViT	0.973	0.524	0.866	304.439
OverLoCK	0.915	3.872	1.203	298.654
H-CAST	0.928	4.516	1.437	285.319
GhostNetV2	0.943	1.235	0.167	328.415
MobileNetV3-S	0.951	0.934	0.057	386.124

Table 1 compares the performance of various models in classification tasks. SCGViT achieves the optimal balance between high performance and lightweight design, with a macro-averaged F1 score of 0.973, only 0.524 M parameters, and 0.866 G of computation. H-CAST also performs well with an F1 score of 0.928. CNN and ResNet18 have the fastest inference speeds, but their classification accuracy is slightly lower. GhostNetV2 and MobileNetV3-S have a smaller number of parameters and computational requirements; however, their F1 scores are still lower than SCGViT. Furthermore, different models were evaluated for the classification of each of the seven types of skin diseases, the specific experimental results are shown in Table 2.

From the category-level evaluation results in Table 2, it can be seen that SCGViT comprehensively outperforms other models in the classification of 7 types of skin lesions. Its precision, recall, and F1 score are the best across all categories, especially for difficult-to-classify samples like bcc and vasc, where the performance is 0.5% and 0.8% higher than the next-best model, MobileNetV3-S, respectively. CNN and ResNet18 lag significantly behind in all three metrics across most categories. While OverLoCK, H-CAST, GhostNetV2, and MobileNetV3-S perform well in some categories, none of them completely surpass SCGViT in all metrics.

The experimental results above show that SCGViT demonstrates improved performance over traditional architectures. SCGViT outperforms existing models in both overall performance and category-specific details. In terms of overall metrics, SCGViT achieves a macro-averaged F1 score of 0.973, an improvement of over 7% compared to CNN's 0.901 and ResNet18's 0.902, and a 2.2% improvement over MobileNetV3-S's 0.951. Simultaneously, its parameter count is only 0.524 M, which is about 1/7th that of OverLoCK (3.872 M), and its FLOPs (0.866 G) are notably lower than those of H-CAST (1.437 G). At the category level, SCGViT shows targeted improvements in F1 scores for all 7 types of lesions. For the difficult-to-classify bcc, its F1 score of 0.935 is 0.5% higher than MobileNetV3-S's 0.930; for vasc, the F1 score of 0.965 is 0.8% higher than MobileNetV3-S's 0.957; even for nv where the CNN model performed relatively well, SCGViT's F1 score of 0.973 is 0.6–0.7% higher than CNN's 0.967 and ResNet18's 0.966. In summary, SCGViT offers efficiency advantages

in its lightweight model design and demonstrates superior adaptability and recognition accuracy for different samples in skin lesion classification tasks.

Table 2. Detailed comparison of specific category assessment results.

Model Name	Index	Akiec	Bcc	Bkl	Df	Mel	Nv	Vasc
CNN	Precision	0.867	0.889	0.882	0.843	0.876	0.978	0.915
	Recall rate	0.924	0.897	0.893	0.901	0.912	0.956	0.938
	F1 score	0.895	0.893	0.887	0.871	0.894	0.967	0.926
ResNet18	Precision	0.872	0.885	0.886	0.849	0.873	0.981	0.908
	Recall rate	0.918	0.903	0.898	0.897	0.916	0.953	0.937
	F1 score	0.894	0.894	0.892	0.872	0.894	0.966	0.922
ViT-Tiny	Precision	0.865	0.893	0.884	0.846	0.871	0.979	0.919
	Recall rate	0.909	0.901	0.896	0.893	0.914	0.957	0.943
	F1 score	0.886	0.897	0.890	0.869	0.892	0.968	0.931
SCGViT	Precision	0.896	0.924	0.912	0.878	0.903	0.982	0.957
	Recall rate	0.921	0.947	0.935	0.905	0.928	0.964	0.973
	F1 score	0.908	0.935	0.923	0.891	0.915	0.973	0.965
OverLoCK	Precision	0.873	0.859	0.891	0.852	0.883	0.980	0.924
	Recall rate	0.912	0.908	0.902	0.895	0.918	0.958	0.945
	F1 score	0.892	0.901	0.896	0.873	0.900	0.969	0.934
H-CAST	Precision	0.881	0.903	0.898	0.861	0.892	0.981	0.932
	Recall rate	0.918	0.921	0.915	0.902	0.924	0.961	0.952
	F1 score	0.899	0.912	0.906	0.881	0.908	0.971	0.942
GhostNetV2	Precision	0.888	0.915	0.907	0.869	0.899	0.981	0.941
	Recall rate	0.923	0.934	0.926	0.907	0.930	0.963	0.960
	F1 score	0.905	0.924	0.916	0.887	0.914	0.972	0.950
MobileNetV3-S	Precision	0.893	0.921	0.913	0.876	0.902	0.981	0.948
	Recall rate	0.927	0.940	0.931	0.911	0.935	0.962	0.967
	F1 score	0.910	0.930	0.922	0.890	0.913	0.972	0.957

4.5. Ablation Experiments

To further investigate the specific contribution of each improved module to detection performance, an ablation study was conducted on the HAM10000 dataset. Table 3 summarizes the impact of adding each module on performance. The first row represents the unmodified MobileViT baseline model, and the “√” symbol indicates that the module was included in the corresponding experiment. The individual component analysis shows their respective performance contributions. By progressively introducing a CGAN, cross attention, and the SE, we quantitatively evaluated the impact of each module on macro-averaged F1 score, number of parameters, computational complexity (FLOPs), and inference speed (FPS). The experimental results are shown in the table below:

Table 3. Comprehensive ablation experiments of SCGViT on the HAM10000 dataset.

CGAN	Cross Attention	SE	Macro-Averaged F1 Score	Number of Parameters (M)	FLOPs(G)	Inference Speed (FPS)
			0.897	0.458	0.579	398.625
√			0.928	0.486	0.699	356.814
√	√		0.959	0.512	0.782	321.573
√	√	√	0.973	0.524	0.866	304.439

As shown in Table 3, introducing a CGAN to the basic MobileViT model improved the macro-averaged F1 score from 0.897 to 0.928. To further verify the impact of the GAN module on the performance of various categories, we analyzed the recall rates of minority categories. Without using the GAN module, the recall rates of MEL and DF remained at 0.583 and 0.657, respectively. However, after introducing the CGAN module, these metrics significantly improved to 0.824 and 0.841. This improvement for specific categories indicates that the CGAN module effectively mitigates the inherent “majority class bias” in the HAM10000 dataset, thereby improving the model’s capacity to recognize rare lesions, which is demonstrated by the increased recall rates in minority categories. It also verifies that using adversarial examples generated by a generator for adversarial training can effectively force the discriminator to learn essential features with stronger discriminability and robustness. After adding the cross-attention module, the model performance further increased to 0.959. With the addition of SE, the F1 score reached a peak of 0.973. This indicates that the channel-wise attention provided by the SE Block effectively adaptively weights the feature maps based on their importance, further refining the key discriminative information. Ablation study data shows that although the number of model parameters slightly increased from 0.458 M to 0.524 M and the inference speed decreased from 398.625 FPS to 304.439 FPS with the addition of these components, the overall model still maintains a lightweight design and real-time performance. The addition of each component brought significant accuracy improvements, with a total F1 score increase of approximately 7.56%, demonstrating the effectiveness and efficiency of SCGViT in the task of assisting in skin lesion diagnosis.

5. Conclusions

This paper addresses the challenges of complex environmental interference and limited computational resources in the auxiliary diagnosis of skin lesions, proposing a lightweight model called SCGViT based on multi-branch, cross-attention fusion and adversarial learning. Through extensive experiments and evaluations on the public HAM10000 dataset, the main research conclusions are as follows:

1. Introducing the CGAN during training enables the model to better learn the intrinsic distribution of lesion features, improving its resistance to variations and noise outside the training set. Ablation experiments confirm that adversarial training significantly strengthens the robustness of the model’s identification of easily confused categories.
2. Multi-branch fusion significantly improves diagnostic accuracy. A single RGB image faces recognition bottlenecks when dealing with images with blurred edges or high noise. This paper utilizes IR and INV, employing cross attention to achieve deep interaction at the feature level, effectively extracting deep structural texture and boundary contrast information of the lesion area. In the 7-class disease classification task of HAM10000, SCGViT achieved a macro-averaged F1 score of 0.973, far exceeding other models.
3. The lightweight design optimizes the parameter distribution while ensuring high-precision performance. The final model contains only 0.52 million parameters, which is approximately 1/7th of OverLoCK model, but achieves an inference speed of 304.439 FPS.

In summary, SCGViT achieves an ideal balance between performance, scale, and efficiency, providing an efficient, reliable, and easily deployable deep learning solution for early screening of skin cancer. Future research will focus on exploring more diverse auxiliary modalities and conducting cross-center validation on larger, more heterogeneous datasets, such as ISIC 2019 and BCN20000, to further verify the model’s robustness and

generalization. Additionally, further optimization of quantization on smaller, resource-constrained embedded hardware will be pursued to enhance real-time diagnostic efficiency.

Author Contributions: Conceptualization, Z.L., C.H. and H.W.; methodology, Z.L. and H.W.; validation, C.H. and Z.L.; resources, H.W.; data curation, C.H. and Z.L.; writing—original draft preparation, Z.L.; writing—review and editing, C.H. and H.W.; visualization, Z.L.; funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Major Science and Technology Projects of Sichuan Province under Grant 2023DZX0013 and Innovative Experimental Projects of Sichuan Province under Grant 2024-451-132.

Data Availability Statement: The dermatological image data utilized in this research were obtained from the HAM10000 dataset, a comprehensive public dataset comprising 10,015 clinical images of seven common skin lesions, which is freely accessible from its official repository.

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

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2025. *CA A Cancer J. Clin.* **2025**, *75*, 40–78. [\[CrossRef\]](#)
2. Kittler, H.; Tschandl, P.; Rosendahl, C. Diagnostic accuracy of dermatoscopy for skin cancer: A systematic review and meta-analysis. *Br. J. Dermatol.* **2016**, *175*, 50–60.
3. China Electronics Standardization Institute. *Edge Computing Reference Architecture—Part 1: General Principles*; China Standards Press: Beijing, China, 2023.
4. Chen, Y.; Zhang, H.; Li, J. Multi-modal attention fusion for skin lesion diagnosis using RGB, infrared and metadata. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 3824–3833.
5. Zhang, L.; Wang, X.; Zhao, Y. GAN-based data augmentation for imbalanced skin lesion classification. *Pattern Recognit.* **2023**, *136*, 109284.
6. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Souza, A.; Pereira, D.; Costa, P. A Hybrid CNN-Transformer Model with Focal Loss for Skin Lesion Classification. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3864–3873.
8. Zhang, Y.; Li, J.; Wang, H. SEACU-Net: A novel U-Net based on squeeze-and-excitation and attention ConvLSTM for skin lesion segmentation. *Comput. Biol. Chem.* **2022**, *102*, 107586.
9. Zhu, S.; Yan, Y.; Wei, L.; Li, Y.; Mao, T.; Dai, X.; Du, R. SECA-Net: Squeezed-and-excited contextual attention network for medical image segmentation. *Biomed. Signal Process. Control.* **2024**, *97*, 106704. [\[CrossRef\]](#)
10. Wen, Y.; Dongming, Z.; Teng, F.; Zhuopu, Y.; Zhen, L. Image segmentation of skin lesions based on dense atrous spatial pyramid pooling and attention mechanism. *J. Biomed. Eng.* **2022**, *39*, 1108–1116.
11. Wei, M.; Wu, Q.; Ji, H.; Wang, J.; Lyu, T.; Liu, J.; Zhao, L. A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion. *Electronics* **2023**, *12*, 438. [\[CrossRef\]](#)
12. Won, H.S.; Chae, J.W.; Cho, H.C. Progressive Defocusing Guided Attention in a Hybrid CNN-Transformer CADx System for Skin Lesion Classification. *J. Electr. Eng. Technol.* **2025**, *20*, 1–10. [\[CrossRef\]](#)
13. Kim, H.; Kim, Y.; Song, W. SkinSavvy2: Augmented Skin Lesion Diagnosis and Personalized Medical Consultation System. *Electronics* **2025**, *14*, 969. [\[CrossRef\]](#)
14. Lee, P.; Bubeck, S.; Petro, J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *N. Engl. J. Med.* **2023**, *388*, 1233–1239. [\[CrossRef\]](#)
15. Xu, R.; Wang, C.; Zhang, J.; Xu, S.; Meng, W.; Zhang, X. Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 6008–6018. [\[CrossRef\]](#)
16. Mehta, S.; Rastegari, M. Separable self-attention for mobile vision transformers. *arXiv* **2022**, arXiv:2206.02680. [\[CrossRef\]](#)
17. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. *Int. Conf. Mach. Learn.* **2021**, *139*, 10347–10357.
18. You, C.; Zhao, R.; Liu, F. Class-aware generative adversarial transformers for medical image segmentation. *arXiv* **2022**, arXiv:2201.10737.
19. Khan, M.; Ahmad, J.; El Saddik, A.; Gueaieb, W. Skin-former: Mobile-friendly transformer for skin lesion diagnosis. In Proceedings of the 2024 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 5–8 January 2024; pp. 1–6.

20. Su, Q.; Hamed, H.N.; Isa, M.A.; Hao, X.; Dai, X. A GAN-based data augmentation method for imbalanced multi-class skin lesion classification. *IEEE Access* **2024**, *12*, 16498–16513. [[CrossRef](#)]
21. 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. [[CrossRef](#)]
22. Walczak, M.; Kallakuri, U.; Humes, E.; Lin, X.; Mohsenin, T. BitMedViT: Ternary-Quantized Vision Transformer for Medical AI Assistants on the Edge. In Proceedings of the 2025 IEEE/ACM International Conference On Computer Aided Design (ICCAD), Munich, Germany, 26–30 October 2025; pp. 1–7.
23. Hoque, M.S. Comparative Analysis of CNN, Vision Transformers, and Hybrid Models for Skin Lesion Classification Using the HAM10000 Dataset. Master's Thesis, Southern Illinois University, Edwardsville, IL, USA, 2024.
24. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [[CrossRef](#)] [[PubMed](#)]
25. Medical University of Vienna. *HAM10000 Dataset Documentation*; Medical University of Vienna: Vienna, Austria, 2018.
26. Gutman, D.; Codella, N.C.; Celebi, E.; Helba, B.; Marchetti, M.; Mishra, N.; Halpern, A. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI). *IEEE Trans. Med. Imaging* **2018**, *38*, 585–598.
27. Mehta, S.; Rastegari, M. MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12057–12070.
28. Lou, M. OverLoCK: An Overview-first-Look-Closely-next ConvNet with Context-Mixing Dynamic Kernels. *arXiv* **2025**, arXiv:2502.20087.
29. Yadav, D.P.; Sharma, B.; Chauhan, S.; Webber, J.L.; Mehbodniya, A. Dual scale light weight cross attention transformer for skin lesion classification. *PLoS ONE* **2024**, *19*, e0312598. [[CrossRef](#)]
30. Lv, D.; Zhao, C.; Ye, H.; Fan, Y.; Shu, X. GS-YOLO: A Lightweight SAR Ship Detection Model Based on Enhanced GhostNetV2 and SE Attention Mechanism. *IEEE Access* **2024**, *12*, 108414–108424. [[CrossRef](#)]
31. Dhanalaxmi, B.; Kumar, B.N.; Raju, Y.; Channapragada, R.S.R. MobileNetV3: An efficient deep learning-based feature selection and classification technique for cardiovascular disease. *J. Eng. Appl. Sci.* **2025**, *72*, 107. [[CrossRef](#)]
32. Höhn, J.; Hekler, A.; Krieghoff-Henning, E.; Kather, J.N.; Utikal, J.S.; Meier, F.; Gellrich, F.F.; Hauschild, A.; French, L.; Schlager, J.G. Skin cancer classification using convolutional neural networks with integrated patient data: A systematic review (preprint). *J. Med. Internet Res.* **2020**, *23*, e20708. [[CrossRef](#)] [[PubMed](#)]

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.