# Deeply Supervised Skin Lesions Diagnosis With Stage and Branch Attention

Wei Dai, *Graduate Student Member, IEEE*, Rui Liu, Tianyi Wu,
Min Wang, *Graduate Student Member, IEEE*, Jianqin Yin, *Member, IEEE*,
and Jun Liu, *Member, IEEE*

***Abstract*—Accurate and unbiased examinations of skin lesions are critical for the early diagnosis and treatment of skin diseases. Visual features of skin lesions vary significantly because the images are collected from patients with different lesion colours and morphologies by using dissimilar imaging equipment. Recent studies have reported that ensembled convolutional neural networks (CNNs) are practical to classify the images for early diagnosis of skin disorders. However, the practical use of these ensembled CNNs is limited as these networks are heavyweight and inadequate for processing contextual information. Although lightweight networks (e.g., MobileNetV3 and EfficientNet) were developed to achieve parameter reduction for implementing deep neural networks on mobile devices, insufficient depth of feature representation restricts the performance. To address the existing limitations, we develop a new lite and effective neural network, namely HierAttn. The HierAttn applies a novel deep supervision strategy to learn the local and global features by using multi-stage and multi-branch attention mechanisms with only one training loss. The efficacy of HierAttn was evaluated by using the dermoscopy images dataset ISIC2019 and smartphone photos dataset PAD-UFES-20 (PAD2020). The experimental results show that HierAttn achieves the best accuracy and area under the curve (AUC) among the state-of-the-art lightweight networks.**

***Index Terms*—Attention, deep supervision, disease classification, skin lesion, vision transformer.**

## I. Introduction

SKIN conditions and disorders are among the most common human diseases to affect millions of people [1], [2]. Statistical data shows that around 20% of Americans are diagnosed with malign cutaneous diseases [3]. Skin cancer, consisting of non-melanoma and melanoma, affected more than 1.5 million new cases globally in 2020. Skin cancer is estimated to be the fifth most commonly detected cancer in the U.S., with 196,060 new cases reported in 2021 [4]. The annual cost for skin cancer treatment is projected to triple from 2011 to 2030 [2]. Proactive detection and early diagnosis are critically essential to save patients. For instance, the five-year survival rate for melanoma patients could be 99% with early-stage diagnosis and treatment, whereas the survival rate is dropped to around 27% if the conditions are detected in the late stage [4].

Traditional methods for detecting skin disorders include skin cancer screening by self-examination and clinical examination. Among these methods, self-examination is the most common method for the early detection of skin diseases. Around 53% of patients with melanomas are self-examined before approaching medical experts [5]. Clinical skin examination can provide affirmative screening of skin cancers with a high detection accuracy [6]. However, clinical examinations consume a considerable amount of time for medical professionals to review a large number of dermoscopic images. The long waiting time of weeks or months could delay the treatment and result in unexpected progress of the skin conditions.

The advances in imaging technology make it possible to diagnose skin lesions by analysing optical images of the skin lesions. The imaging modalities in skin examination include reflectance confocal microscopy, total body photography, teledermatology, and dermoscopy. Among all the imaging modalities, dermoscopy is a non-invasive imaging method without reflecting light to examine the skin lesions with up to 10× magnification [6]. However, manual analysis of dermoscopic images is time-consuming, and the examination results are subjective to healthcare providers.

Due to the sophisticated features of skin lesion images, it is nontrivial to automatically detect unhealthy skin areas with satisfactory accuracy. Previously, computer-aided approaches were proposed to analyse skin lesion images. Histogram thresholding employs empirical thresholds to isolate lesions from the rest of the skin tissues [7]. Principle component analysis (PCA)

Manuscript received 25 August 2022; revised 22 February 2023, 17 May 2023, and 14 July 2023; accepted 17 August 2023. Date of publication 25 August 2023; date of current version 6 February 2024. This work was supported in part by the Research Grant Council (RGC) of Hong Kong under Grants 11217922, 11212321, and ECS-21212720, in part by Guangdong Province Basic and Applied Basic Research Fund Project under Grant 2019A1515110175, and in part by the Science and Technology Innovation Committee of Shenzhen under Grant Type-C SGDX20210823104001011. *(Corresponding author: Jun Liu.)*

Wei Dai, Rui Liu, Tianyi Wu, Min Wang, and Jun Liu are with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, SAR, China (e-mail: wei.dai@my.cityu.edu.hk; rui.liu@my.cityu.edu.hk; wu.tianyi@my.cityu.edu.hk; min.wang@my.cityu.edu.hk; jun.liu@cityu.edu.hk).

Jianqin Yin is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: jqyin@bupt.edu.cn).

The code is available at https://github.com/anthonyweidai/HierAttn.

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2023.3308697, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2023.3308697

for colour histogram was further utilised in lesion colour space clustering for segmentation [8]. Gradient vector flow was computed to extract the smoothness and compactness of curve shapes in skin lesions for separating lesions from neighbouring tissues [9]. However, the majority of these techniques require heavy human participation and cannot extricate sufficient features from a whole skin lesion image to diagnose skin diseases.

To achieve an accurate and unbiased diagnosis of skin diseases, deep learning has been applied to extract representative features and provide an end-to-end analysis of medical images. To improve the diagnostic results, researchers assembled several models, such as ResNeXt, NASNet, SENet, DenseNet121 and EfficientNet, to detect skin cancer, and the highest accuracy of these models was reported up to 94.2% and 92.6% on the ISIC2018 and ISIC2019 datasets, respectively [10]. However, the majority of reported networks, especially those constructed by combining several models, consumed a large number of computational resources and were highly time-consuming due to the increased complexity of the models.

Recently, lightweight algorithms have been reported to promote the use of deep learning in conducting direct skin cancer screening with limited computational resources in clinical computers and mobile devices. The attention mechanism is a biomimetic design module in deep learning that is widely used to value the contextual information of an object with minimal computational cost [11]. An example of the attention mechanism is depthwise separable convolution (DWSConv) which effectively reduces the models' size while keeping a competitive performance [12]. Most recently, several models leveraged CNN and vision transformer (ViT) by applying standard convolution [13] or DWSConv [14] to downsize tensors before each ViT module. Although the aforementioned techniques have attained satisfactory results in common object recognition, there has been limited investigation into the application of light deep learning models for skin lesions analysis.

This article aims to address the challenges in balancing the reliability and accuracy of skin lesion analysis with small memory storage and minimal computational cost. The major novelty and key contributions of this study are:

- We introduce a new lightweight and deeply supervised architecture with diverse attention mechanisms, namely HierAttn, to distinguish multi-class skin lesions. HierAttn achieves top-level performance with a smaller size than other competing mobile models for both the ISIC2019 and PAD2020 datasets.
- We develop a novel deep supervision method, branch attention algorithm, to learn local and global representations from coarse level to fine level of features with hierarchical pooling and aggregate the extracted hierarchical features by tensors assembling. Since the hierarchical pooling and tensor assembling do not introduce additional learnable parameters, high-level features generated by the stage attention block can be extracted without increasing the model size. Moreover, branch attention does not introduce additional loss computation, which is computation-friendly compared to the conventional deep supervision method with multiple losses.
- We propose the same channel attention (SCAttn) module based on global averaging pooling without unnecessary modification of channel-wise features. The SCAttn design effectively extracts global features and outperforms other attention methods like squeeze and excitation while avoiding the increment of the model parameters to keep a small model size
- We propose the stage attention block, consisting of a SCAttn block followed by a convolution-transformer hybrid (CTH) block, to thoroughly learn the regional and global high-level feature representations. As shown in Fig. 1, the HierAttn network involves a novel CTH block by adding a skip connection and stochastic depth to gain optimal performance.

We review related literature in Section II and explain the proposed methodology in Section III. We present and discuss the experiment results in Sections IV and V. We summarise the significant findings in Section VI.

## II. Related Work

### A. CNNs in Skin Lesions Diagnosis

Recent advances in machine learning have introduced an increasing number of deep neural networks. The deep learning models, including InceptionV3, VGG, EfficientNet, ResNet, DenseNet, etc., have been applied to classify skin lesions [3], [15], [16]. Moreover, the classification activation feature map was used to capture region of interest information for learning indicative lesion representations [17]. To learn a different degree of contextual information, Dai et al. introduced an encoder which resizes features in different scales [18]. In other studies, researchers processed three modalities of patient data, including non-image metadata, clinical and dermoscopy images, in a two-stage feature fusion network to enhance skin lesions classification performance [19]. In addition, the hybrid-fusion network (Hi-Net) was proposed to amalgamate multi-modal MR images for better visualisation of syndromes [20].

Furthermore, the models, assembling several networks (e.g., ResNeXt, NASNet, SENet, DenseNet121, and EfficientNet) in several streams or stages, were utilised in skin lesions classification [16], [21], [22]. The ISIC2019 challenge winner applied an ensemble model accumulating EfficientNet B0 to B6, SENet154, and two ResNeXt to achieve a 92.6% average classification accuracy; however, the model is comparatively large with more than 100M parameters, making it impractical for clinical and home use [16].

### B. Deeply Supervised Learning

Adding and widening layers are common methods to improve training metrics by increasing the learning parameters of networks. However, both methods increase the computational complexity and hardware requirements during training. Deeply supervised learning is frequently used to learn coarse-to-fine features from intermediate branches. Instead of adding layers to the network, deep supervision usually uses auxiliary supervision branches in critical intermediate stages during training [23]. In

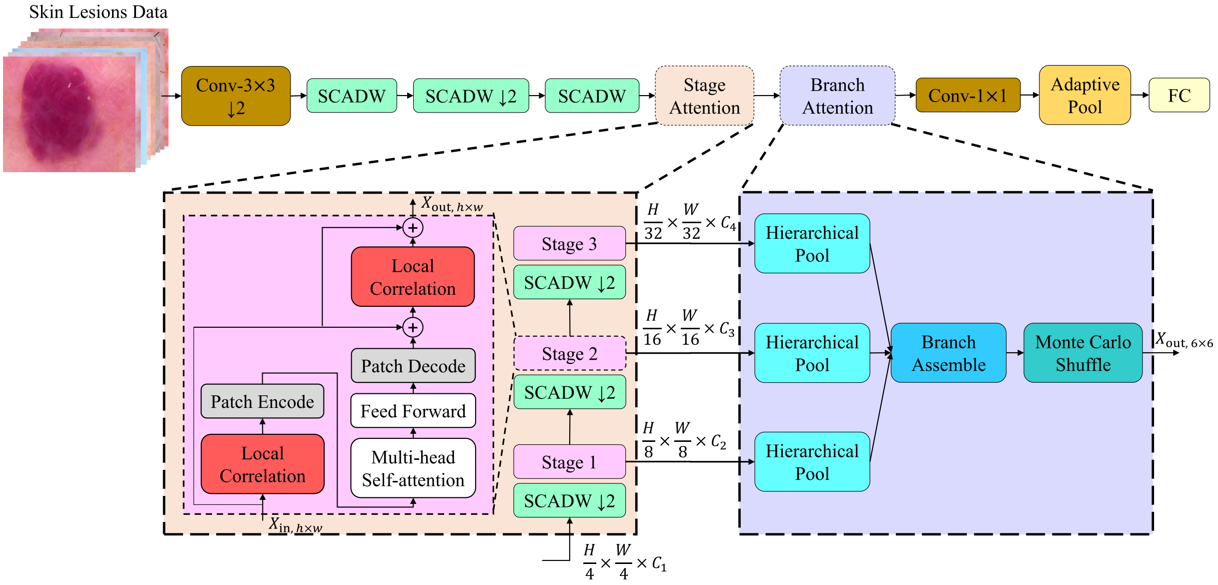


Fig. 1. HierAttn architecture. Conv-$n \times n$ represents a standard convolution, and SCADW refers to a depthwise separable convolution block with the SCAttn module. Down-sampling blocks are marked with ↓2. Stage attention has a SCADW block followed by a CTH block (pink block) that thoroughly aggregates local features and learns contextual representations. Branch attention utilises hierarchical pooling to extract global and local features steadily.

deep supervision, multilevel losses are widely used to extract feature information from stages to improve the model's performance. For instance, GoogleNet has three losses in three branches, separately [24]. Liu et al. applied deep supervision with multiple losses to improve object edges learning [25]. However, the features in various stages are fed into simple tensor operations separately. They do not have a direct correlation influence on each other.

### C. Attention Mechanisms

The attention mechanism is a biomimetic cognitive method used in diverse computer vision assignments such as image classification [14], [26], [27], [28], [29] and object segmentation [14], [30], [31], [32]. For example, reverse attention is proposed to utilise the feature mask as a cue to guide polyp segmentation [30]. Moreover, the co-attention module was utilised for encoding features from two branches of a Siamese network to increase the correlations among video frames [31]. Another example of the attention network is the SENet, which obtains global representations by global average pooling and channel-wise feature response by squeeze and excitation [26]. The convolution block attention module (CBAM) improves the SENet by using a larger kernel size to encode spatial information [27]. Coordinated attention (CoorAttn) further advances global average pooling by encoding channel relationships and long-range dependencies via average pooling along different axes [29]. However, they introduce more learnable parameters and consume more computational resources than the SENet.

### D. Vision Transformer

To improve the computational efficiency and meet the scalability requirement, self-attention mechanisms, particularly transformers, are introduced into computer vision from natural language processing. With self-attention, a ViT replaces the traditional convolution method with a transformer encoder [28]. Because the input images are directly split into patches and embedded, the ViT models are still huge, with more than 100M parameters [28]. The standard transformer has been applied to process sequences of image patches to learn the inter-patch representations. However, the original transformer method ignores the inductive biases (e.g., translation equivariance and locality) inherent to CNNs, which leads to poor performance when training with insufficient data [28]. Recently, ViT was utilised in skin lesions evaluation by appending a ViT branch to CNN-based U-shape architecture for improving long-range dependencies and contextual information extraction, but an additional branch notably increased the model weight [33].

### E. Lightweight Networks

Although knowledge distillation can effectively reduce learnable parameters, it depends on complex parameters updating between teacher and student networks and bares the comparatively enormous computational cost of the training process [37], [38], [39]. Moreover, atrous spatial pyramid pooling (ASPP) can be applied to spatially sampling features by pooling at different scales for localising segment borders as a lightweight module [40], [41], but ASPP underrates the interaction among features. Another prominent solution to achieving lightweight CNNs is the DWSConv, which replaces the standard convolution with depthwise convolution within one channel and pointwise convolution along each channel for the depletion of learnable parameters [12]. Based on the DWSConv method, MobileNetV2, MobileNetV3, EfficientNet, MnasNet, and ShuffleNet were constructed and obtained relatively satisfactory performance [42], [43], [44], [45], [46].

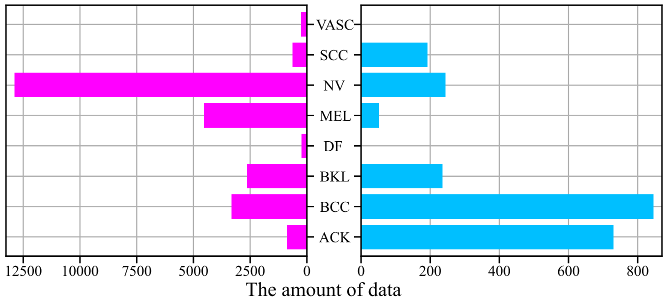


Fig. 2. Data distribution on ISIC2019 (left image) and PAD2020 (right image).

As for lightweight ViT networks, sparse attention [13], random feature attention [34], and low-rank approximation [35] were adopted to reduce ViTs' size and computational cost. To reduce the latency of splitting images, computer scientists from Apple proposed MobileViT to tackle the loss of inductive biases by taking convolution and transformer to form a hybrid block [14]. The design of convolution with transformer shows superior performance than conventional fully convolution mobile networks in terms of parameter count and inference time, as reported by [14]. Although several aforementioned networks (e.g., EfficientNet, MobileNetV2) have been utilised for diagnosing skin lesions [16], [36], the exploration of reducing the number of parameters and computational costs for mobile applications in skin lesion recognition remains limited.

## III. Proposed Methodology

### A. Skin Lesion Dataset

Two publicly available skin lesions datasets, ISIC2019 [10], [47], [48] and PAD2020 [49], are used in this research for the development and evaluation of the proposed methods. Dermoscopy and smartphones are two standard methods of capturing skin lesions images. Thus, the two datasets effectively represent current image data for the classification of skin lesions. The data distribution of the two datasets is shown in Fig. 2.

ISIC2019 dataset consists of 25,331 dermoscopy images with eight categories: actinic keratosis (ACK), basal cell carcinoma (BCC), benign keratosis (BKL), dermatofibroma (DF), melanoma (MEL), melanocytic nevus (NV), squamous cell carcinoma (SCC), vascular lesion (VASC). Three of them (ACK, BCC and SCC) belong to the non-melanoma skin cancer. Melanoma is the most severe skin cancer that is caused by uncontrolled growing cells that can produce pigment. The vascular lesions (VASC) can be either benign or malign and requires close monitoring over a period of time. The remaining categories of skin lesions are benign conditions. The PAD2020 dataset includes 2,298 skin lesion images with six classes: ACK, BCC, BKL, MEL, NV, and SCC, collected by using smartphones. PAD2020 has two fewer classes, DF and VASC, than ISIC2019 because of lacking photos of skin lesions.

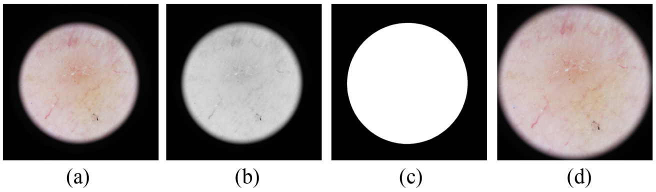

(a) (b) (c) (d)

Fig. 3. Progress in cropping image (a) original image, (b) greyed image, (c) binarised image, and (d) cropped image.

### B. Pre-Processing

*1) Image Pre-Processing:* The data among different classes from the ISIC2019 dataset has a large black area [see Fig. 3(a)], which damages the evaluation performance of deep learning models. Thus, an adaptive cropping method is taken to identify and crop these images. The original image is first turned into greyscale [Fig. 3(b)] and then binarised an adaptive threshold ranging from 50 to 255 [Fig. 3(c)]. After that, the contour of the binarised image is detected to confirm the circle location. When the value of the circle area divided by the whole image area is between 0.01 and 0.9, the region enclosed by the circle is cropped and saved [Fig. 3(d)].

*2) Data Balance:* In this study, an imbalance ratio is defined to assess the imbalance problem quantitatively. The imbalance ratio is a fraction of the number of images of the majority class over the minority class. The imbalance ratio for ISIC2019 and PAD2020 are 53.9 and 16.3, respectively. Such relatively high imbalance ratios can remarkably reduce training performance, according to previous research [50]. The data imbalance could lead to a low validation result for the minor class, though the averaged validation metrics over all classes could be high. Data balance by oversampling or undersampling for each class is a practical technique to handle the dataset with a huge imbalance ratio. Oversampling and undersampling are simultaneously utilised to balance ISIC2019 and PAD2020 data. After sampling, 2500 and 500 images are collected for each class in the ISIC2019 and PAD2020 datasets, respectively. Thus, the amount of data after balancing totals 20000 and 3000 on ISIC2019 and PAD2020, respectively, is close to the total amount of unbalanced data.

Oversampling uses horizontal flips, random crops, Gaussian blur, linear contrast, random translation, rotation, and shear on a small scale to generate new images from the old images. As for undersampling, the random selection of a fixed number of images tends to have a class-overlapping problem. Thus, we apply an adaptive data analysis method called instance hardness (IH) [51] to alleviate this adverse effect. Instance hardness is defined as:

$$\mathrm{IH}_{\mathcal{L}}\left(\langle x_i,\ y_i\rangle\right) = 1 - \frac{1}{|\mathcal{L}|}\Sigma_{j=1}^{|\mathcal{L}|} p(y_i|x_i, g_j(t,\alpha)) \tag{1}$$

where $\mathcal{L}$ is a prior with non-zero probability while treating all other learning algorithms as having zero probability, $g$ is a machining learning algorithm trained on $t$ with the hyperparameter $\alpha$, and $y_i$ is the label for data $x_i$.

Outliers and mislabelled data are expected to have high IH. Thus, IH analysis was applied in undersampling to remove those

TABLE I
DETAILED STRUCTURES OF THE PROPOSED HIERATTN ARCHITECTURE. HERE, $d$ MEANS THE DIMENSIONALITY OF THE INPUT TO THE TRANSFORMER LAYER IN THE CTH BLOCK. IN THE CTH BLOCK, KERNEL SIZE IS SET AS THREE AND PATCH HEIGHT AND WIDTH ARE SET AS TWO

| **Layer** | | **Output size** | **Repeat** | **Output Channels** | |
|---|---|---|---|---|---|
| | | | | XS | S |
| Image | | $256 \times 256$ | | | |
| Conv$3 \times 3$, $\downarrow 2$<br>SCADW | | $128 \times 128$ | 1<br>1 | 16<br>16 | 16<br>32 |
| SCADW, $\downarrow 2$<br>SCADW | | $64 \times 64$ | 1<br>2 | 24<br>24 | 48<br>48 |
| Stage 1 | Conv$3 \times 3$, $\downarrow 2$<br>CTH block | $32 \times 32$ | 1<br>1 | 48<br>$48(d = 64)$ | 64<br>$64(d = 96)$ |
| Stage 2 | Conv$3 \times 3$, $\downarrow 2$<br>CTH block | $16 \times 16$ | 1<br>1 | 64<br>$64(d = 80)$ | 80<br>$80(d = 120)$ |
| Stage 3 | Conv$3 \times 3$, $\downarrow 2$<br>CTH block | $8 \times 8$ | 1<br>1 | 80<br>$80(d = 96)$ | 96<br>$96(d = 144)$ |
| Branch attention<br>Conv$1 \times 1$ | | $8 \times 8$ | 1<br>1 | 192<br>768 | 240<br>960 |
| Pooling<br>Linear | | $1 \times 1$ | 1 | 8 or 6 | 8 or 6 |
| **# Parameters** | | | | 1.08 M | 2.14 M |

data with a high IH. This research uses the random forest as the machine learning method g, referenced from the imbalanced-learn study [52].

### C. HierAttn Architecture

The article introduces an efficient and lite Convolution-Transformer model [see Fig. 1], HierAttn, for skin lesions diagnosis. The main idea of this model is to hierarchically learn the local and global representations by utilising various attention mechanisms. Branch attention allows the network to extract various levels of features from different layers. Moreover, the SCAttn module in our new HierAttn network leverages the DWSConv by supplementing global information to improve performance. The three novel mechanisms (i.e., SCAttn, stage attention, and branch attention) consume miniature parameters or no learnable parameter, which results in the tiny size of HierAttn. Besides, the stage attention utilises an effective self-attention hybridising convolution to keep inductive biases and reduce the latency, inspired by the MobileViT architecture [14]. The attention modules are described in the following sections. Detailed structures of HierAttn are shown in Table I.

*1) Branch Attention:* Zhang et al. computed three more losses among intermediate stages to improve the feature interactions during training [53]. Instead, we propose a novel deep supervision method, brach attention (i.e., lower right in Fig. 1, more details in Fig. 4), utilising hierarchical pooling to downsize tensors from different stages, tensors assembling to improve the interactions of the downsized features and learn the hierarchy of features from the assembled tensors. Moreover, by keeping the different sizes of pooling results, hierarchical pooling learns the local representations of tensors, generating large tensor size ($C \times 5 \times 5$), and attains tensors' global representations, generating small tensor size ($C \times 1 \times 1$). The medium output tensor with size $C \times 3 \times 3$ is also designed as a buffer layer to keep both local and global features. The pooled tensors from different branches are then channel-wisely assembled. After that, the ensembled tensors are pixel-wisely assembled. At the end of branch attention, the assembled tensors are pixel-wisely shuffled by utilising the Monte Carlo method. The branch attention is applied as a particular learning stage after each stage attention block (critical learning stage). Most importantly, by using such a practical design, branch attention does not introduce an additional loss in the intermediate stage, which consumes less computational costs than the conventional deep supervision method.

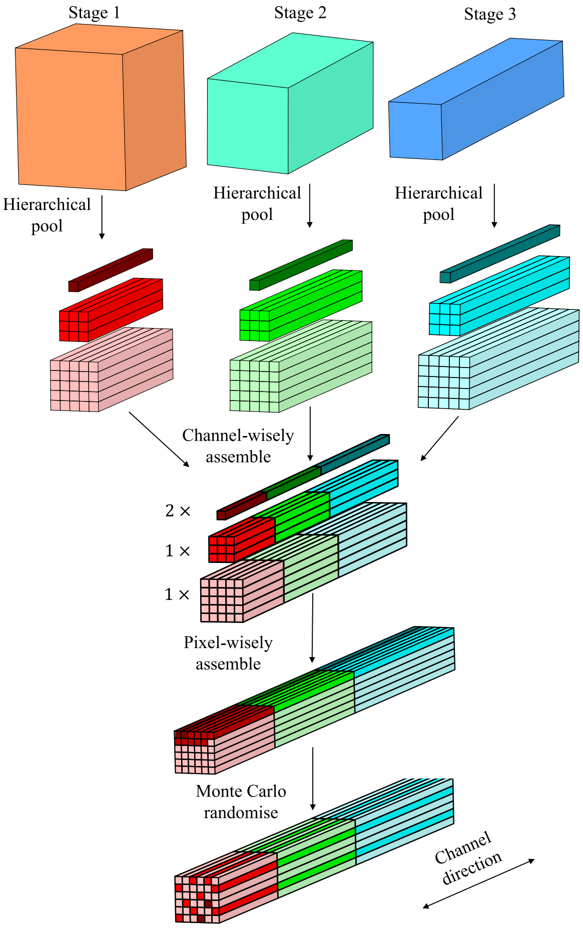


Fig. 4. Demonstration of the branch attention. From top to bottom, feature tensors from three branches are shrunk by hierarchical pooling. The pooled tensors are assembled by directions of channel and pixel and fully assembled with pixel-wise randomisation.

*2) Same Channel Attention:* The block modifies the DWS-Conv with the SCAttn technique and results in a new block structure called SCADW block (i.e., green blocks in Fig. 1). Squeeze and excitation attention (SEAttn) blocks were proposed to enhance the expressive power of the learned features after depthwise convolution in the MobileNetV3 block [46]. However, the pointwise convolution already extracts the channel-wise information, which suggests that the excitation of SEAttn is likely a redundant operation. Therefore, the SCAttn is introduced after depthwise convolution by global average pooling while maintaining the same number of learnable parameters as the DWSConv block. The SCAttn block is illustrated in Fig. 5.

*3) Stage Attention:* Each stage attention module (i.e., lower left in Fig. 1) has a SCADW block with a stride of 2 followed by a convolution-transformer hybrid block. We apply the

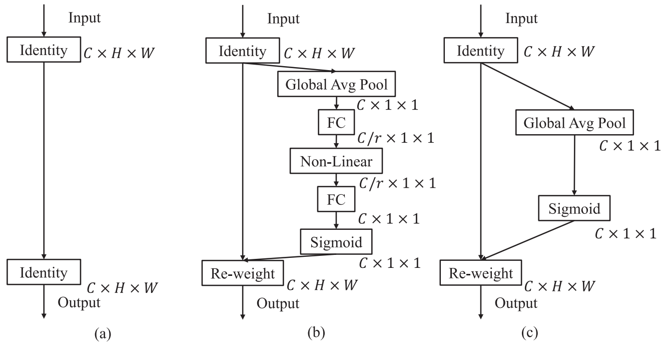


Fig. 5. Schematic comparison of (a) the original block (without attention mechanism), (b) the SEAttn block, and (c) the proposed SCAttn block.

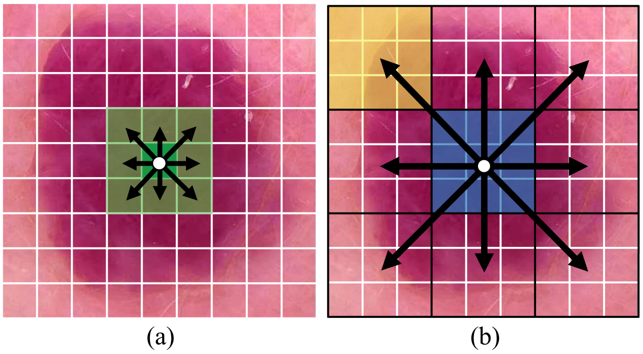


Fig. 6. Feature interactions in stage attention. (a) Local correlation extract local features and (b) Multi-head self-attention further learns global information.

convolutions and transformers simultaneously to learn the local and global representations of an input skin lesion image with fewer parameters. The convolution-transformer hybrid (CTH) block (i.e., pink blocks in Fig. 1) utilised consecutive modules to process feature maps before and after encoding. The CTH block consists of a multi-head self-attention (MHSA) followed by a multilayer perceptron feed forward (FFN) layer, which is regarded as vision transformer (ViT). As shown in Fig. 6(a), ViT learns long-range spatial dependence among encoded patches. To integrate local features, local correlation through standard convolution is applied before and after the ViT module in the CTH block, which is illustrated by Fig. 6(b). Furthermore, we also add a skip connection to link the input and output of the CTH block. The stage attention module thoroughly rearranges feature maps by downsizing feature maps, and the CTH block further extracts the processed features. Thus, each stage attention module is regarded as a **critical learning stage** in HierAttn. The CTH block bottleneck can be formulated as:

$$\begin{aligned} X &= \mathrm{LocalCorr}(X_{\mathrm{in}}) + X_{\mathrm{in}} \\ Y &= \mathrm{MHSA}(\mathrm{Encode}(X)) \\ Z &= \mathrm{Decode}(\mathrm{LayerNorm}(\mathrm{FFN}(Y))) \\ X_{\mathrm{out}} &= \mathrm{LocalCorr}(Z + X_{\mathrm{in}}) + X_{\mathrm{in}} \end{aligned} \tag{2}$$

*4) Small-Scale Attention Modules:* SCAttn in SCADW block (i.e., green blocks in Fig. 1) leverages SEAttn by keeping single-pixel attention through global averaging pooling and avoiding the modification of tensors along channel direction. It contributes to utilising an attention mechanism to extract global features after depthwise convolution without introducing any learnable parameters. Moreover, branch attention (i.e., lower right in Fig. 1) does not increase model size because hierarchical pooling and tensor assembling in branch attention are parameter-free. In addition, transformer operations lose spatial bias, increasing computational costs with a broader and deeper design to learn visual representations [14]. Therefore, ViT-Base, ViT-Large, and ViT-Huge models use the number of transformer blocks L = 12, 24, 32 and the number of embedded dimensions d = 768, 1024, 1280, respectively [28]. However, our new HierAttn only requires L = 2, 4, 3 and d = 96, 120, 144 (i.e., lower left in Fig. 1). The number of parameters for the new HierAttn and other state-of-the-art (SOTA) networks is summarised in Table II.

## IV. Experimental Results

In this section, we first evaluate HierAttn performance on the ISIC2019 and PAD2020 datasets in Section IV-B. HierAttn delivers significantly better performance than the SOTA mobile (<5M) and comparatively large networks, which can be seen in Table II and Fig. 7. In Section IV-C, we conducted ablation studies for attention mechanisms in DWSConv.

### A. Evaluation Metrics

The ISIC community recommends accuracy, precision, F1-score, specificity, ROC, and AUC to be the evaluation metrics for skin lesions classification [54]. In our experiments, top-1 accuracy (abbrev. accuracy), F1-score, and specificity are adopted to evaluate the performance of skin lesions classification (e.g., the proportion of correct values in inference results). Moreover, the receiver operating characteristic (ROC) curve and its area under the curve (AUC) are applied to demonstrate the more general performance of each model (e.g., the model's robustness and scale invariance).

Furthermore, each model's number of parameters (# parameters) with the unit of Million (M) is applied to measure its weight. Floating-point operations per second (FLOPs), computed using pytorch-OpCounter [55], and inference time are also exploited to estimate the computational complexity of the model and quantify its practicalities on skin lesions datasets. To measure the inference time, the average of 1000 iterations is calculated using an Intel Core i5-10210U CPU (1.60GHz) on a laptop, which can demonstrate the availability of HierAttn on mobile applications with slower processors.

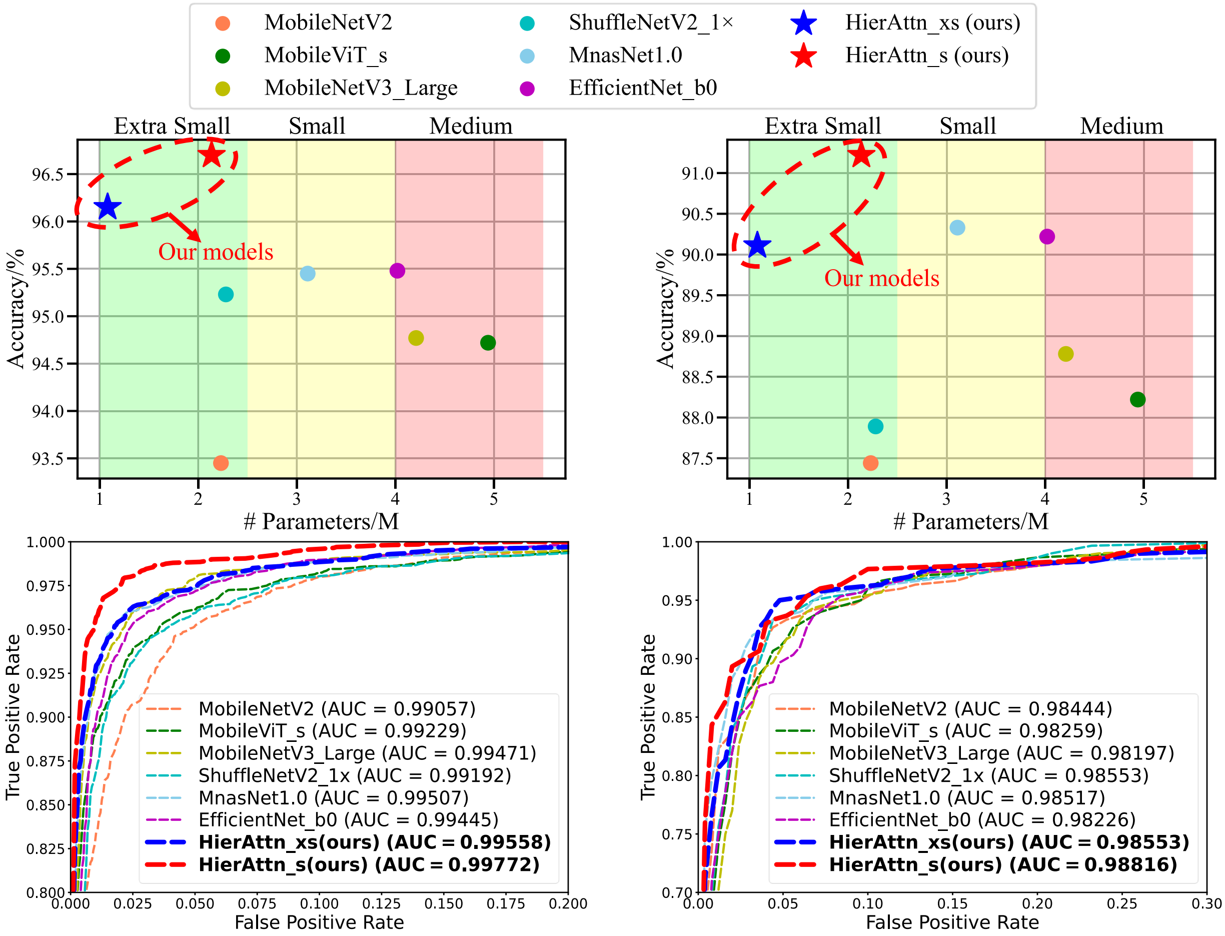
### B. Image Classification on the Skin Lesions Dataset

*1) Implementation Details:* The HierAttn network, other SOTA lite models, and three larger networks are trained and validated for 500 epochs on one RTX A4000 with a batch size of 64 images using AdamW optimiser [56] with 10-fold

TABLE II
THE CLASSIFICATION RESULTS OF DIFFERENT MODELS ON ISIC2019 AND PAD2020 DATASETS

| **Model** | **# Parameters /Million** | **FLOPs /Billion** | **Inference time/s** | **Accuracy/%** | | **Precision/%** | | **F1-score/%** | | **Specificity/%** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | ISIC2019 | PAD2020 | ISIC2019 | PAD2020 | ISIC2019 | PAD2020 | ISIC2019 | PAD2020 |
| VGG11 [58] | 128.80 | 19.56 | 0.139 | 75.95 | 70.83 | 76.76 | 71.10 | 75.68 | 70.49 | 96.56 | 96.73 |
| ResNet101 [59] | 42.52 | 20.54 | 0.187 | 92.20 | 87.67 | 92.14 | 87.67 | 92.10 | 87.62 | 98.89 | 97.53 |
| MobileNetV2 [42] | 2.23 | 0.86 | 0.042 | 93.45 | 87.44 | 93.42 | 87.36 | 93.42 | 87.35 | 99.20 | 98.20 |
| MobileViT_s [14] | 4.94 | 2.30 | 0.101 | 94.72 | 88.22 | 94.74 | 88.29 | 94.71 | 88.19 | 99.42 | 98.47 |
| MobileNetV3_Large [46] | 4.21 | 0.60 | 0.018 | 94.77 | 88.78 | 94.78 | 88.98 | 94.76 | 88.66 | 99.49 | 98.20 |
| ShuffleNetV2_1× [43] | 2.28 | 0.40 | 0.017 | 95.23 | 87.89 | 95.21 | 88.02 | 95.21 | 87.82 | 99.35 | 98.40 |
| RegNetY6.4gf [60] | 29.30 | 16.76 | 0.171 | 95.35 | 88.67 | 95.42 | 88.78 | 95.36 | 88.65 | 99.34 | 98.27 |
| MnasNet1.0 [45] | 3.11 | 0.88 | 0.034 | 95.45 | 90.33 | 95.50 | 90.46 | 95.45 | 90.30 | 98.96 | 97.53 |
| EfficientNet_b0 [44] | 4.02 | 1.08 | 0.028 | 95.48 | 90.22 | 95.46 | 90.26 | 95.47 | 90.16 | 99.49 | 98.27 |
| Ensembled Network [16] | 116.26 | 13.35 | 0.444 | 96.25 | 91.13 | 96.24 | 91.29 | 96.23 | 91.13 | 99.46 | 98.23 |
| **HierAttn_xs (ours)** | **1.08** | 0.44 | 0.023 | 96.15 | 90.11 | 96.14 | 90.32 | 96.13 | 90.10 | 99.37 | 98.33 |
| **HierAttn_s (ours)** | 2.14 | 1.32 | 0.039 | **96.70** | **91.22** | **96.69** | **91.35** | **96.69** | **91.19** | **99.51** | **98.53** |

The bold values mean the best performance.



(a)

(b)

Fig. 7. HierAttn vs. SOTA lightweight models. Classification accuracy versus the number of parameters (# parameters) and the receiver operating characteristic (ROC) curves on (a) ISIC2019 validation set and (b) PAD2020 validation set.

cross-validation and cross-entropy loss. The learning rate is ceased from 0.002 to 0.0002 during the first 30 epochs and then increased to 0.0002 utilising the cosine scheduler [57]. L2 weight decay of 0.01 is adopted. Moreover, knowledge transfer is applied to reduce the training time and improve model performance. All transferred models for training in ISIC2019 were trained in ImageNet1K, and the well-tuned models from ISIC2019 were then tuned in PAD2020. And the parameters of layers after branch attention in HierAttn are randomly initialised. Additionally, the transfer learning warm-up is applied to alleviate the negative influence of untransferred layers on transferred layers. On the first training 30 epochs, all transferred layers are frozen. After that, gradient calculation is required for all layers with learnable parameters. Furthermore, the number of parameters of each model is computed with eight classes to simplify the discussion because each model with 8 classes or 6 classes has a similar number of parameters. Meanwhile, mobile models with (1, 2.5) M, (2.5, 4) M, (4, 5.5) M parameters are regarded as "extra small", "small", and "medium", respectively.

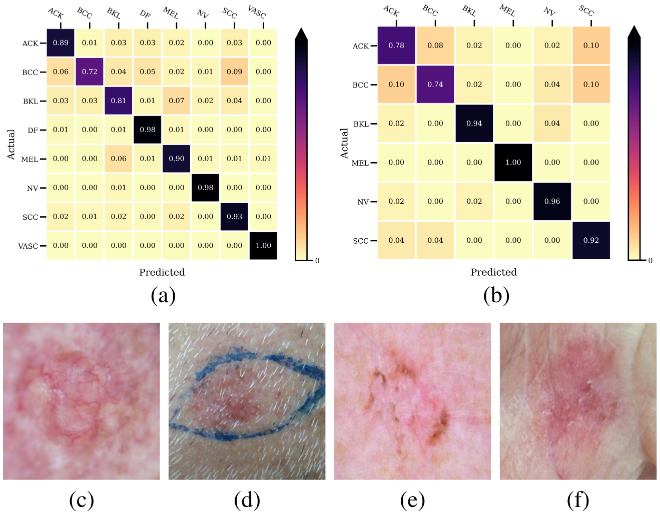
*2) Classification Results:* Fig. 7 compares HierAttn with six other lightweight networks that are also trained on the ISIC2019 and PAD2020 datasets. Detailed values are illustrated in Table II. The figures on the first row of Fig. 7 demonstrates that HierAttn networks fall in the upper left region, which means they outperform other mobile architectures with relatively small sizes. For instance, with around 2.14 million parameters, HierAttn_s outperforms MobileNetV2 by 3.25%, MobileNetV3_large by 1.93%, ShuffleNetv2_1x by 1.47%, MnasNet1.0 by 1.25%, and EfficientNet-b0 by 1.22% accuracy on the ISIC2019 validation set. Furthermore, HierAttn_s also outperforms MobileViT_s, which also contains convolution-transformer hybrid blocks, by 1.98% and 3.00% accuracy on ISIC2019 and PAD2020 datasets, respectively. Besides, HierAttn attains more than 90% accuracy on both the dermoscopy and the smartphone images of skin lesions, demonstrating that HierAttn is versatile in visually diagnosing skin lesions from various domains.

It can also be discovered in Table II that HierAttn_s achieves the highest precision, 96.69% and 91.35%, and HierAttn_xs attains the third highest precision, 96.14% and 90.32%, on ISIC2019 and PAD2020, respectively. The experimental results indicate that HierAttn can functionally distinguish more than 90% samplers who truly have skin diseases from the positive test results. Moreover, HierAttn_s secures the first place in F1-score (i.e., 96.69% on ISIC2020 and 91.19% on PAD2020), suggesting that HierAttn_s was robust and powerful in recognising skin lesions.

As shown in Table II, HierAttn_s delivers a better specificity (i.e., 99.51% on ISIC2019 and 98.53% on PAD2020) than other models, which demonstrates that HierAttn_s has the best potential to precisely discriminate the healthy people from all cases. In the real-life application, HierAttn_s is accurate and effective in recognising and comforting healthy people who suspect the misleading symptoms of skin diseases.

Furthermore, the larger models, VGG11, ResNet101, and RegNetY6.4gf, do not have satisfactory performance in detecting skin diseases. Particularly, VGG11 with 128.80M parameters achieves lower classification results than mobile models by a considerable margin (e.g., $-17.50\%$ accuracy on ISIC2019 and $-16.61\%$ accuracy on PAD2020). Although the Ensembled Network reaches the second highest performance in classification results, it contains over 100M parameters and has the highest inference time, as shown in Table II.

*3) Inference Analysis:* Table II illustrates that inference time for each image using HierAttn_xs is below 0.03s, which is 30% faster than MnasNet1.0 and only 6ms slower than ShuffleNetV2_1×. Despite the increased computational complexity from the novel self-attention and branch attention mechanisms, the HierAttn_xs model remains the second and third fastest in the FLOPs (0.44 billion) and inference time (0.023s). This time cost is acceptable for self-examination or clinical diagnostic purposes. In contrast, VGG11, ResNet101, RegNetY6.4gf, and Ensembled Network consume over 10 billion FLOPs and 100ms to recognise each image. Notably, the Ensembled Network utilises 444ms to diagnose a single image, approximately 19 times longer than HierAttn_xs. Therefore, lightweight models are more suitable for analysing skin diseases than these larger models.



(a) (b)

(c) (d) (e) (f)

Fig. 8. Confusion matrices of (a) ISIC2019 & (b) PAD2020 classification and examples of failure cases from the two datasets. (c) & (d) BCC are misclassified as (e) & (f) SCC images because of the low contrast between the diseased and healthy regions.

*4) Roc and Auc:* The graphs on the second row of Fig. 7 illustrate that all lite models on experiments have more than 0.99 AUC on the ISIC2019 and 0.98 AUC on the PAD2020 validation sets, suggesting that they possess sufficient analytical capacities to conduct multi-classes lesions classification tasks. Furthermore, the ROCs of HierAttn_s and HierAttn_xs are the first and second closest to the point (0, 1) on both ISIC2019 and PAD2020 validation sets. Moreover, HierAttn_s and HierAttn_xs have the first and the second largest AUC on both datasets among all models, for instance, 0.99772 and 0.99558 on the ISIC2019 validation set, respectively. Thus, the ROCs and AUCs show that HierAttn is the most reliable and superlative model to detect skin lesions among current conventional and advanced mobile models. Meanwhile, the same model's AUC of the PAD2020 validation set is lower than the ISIC2019 validation set, which means these models perform better on the ISIC2019 validation set (in conformity with lite models' performance in Table II).

*5) Error Analysis:* The confusion matrices of HierAttn_s on the ISIC2019 and PAD2020 classification tasks are presented in Fig. 8(a) and (b), revealing that basal cell carcinoma (BCC) is the most commonly misclassified class, often mistaken for squamous cell carcinoma (SCC). Several images of BCC were misclassified as SCC with less than 4% difference in accuracy between BCC and SCC. Because these images have low contrast between the infection area and healthy region, the texture of syndromes is similar between BCC (Fig. 8(c) and (d)) and SCC (Fig. 8(e) and (f)). Such a challenge could be further addressed by improving the ability of HierAttn to widen the inter-class distance.

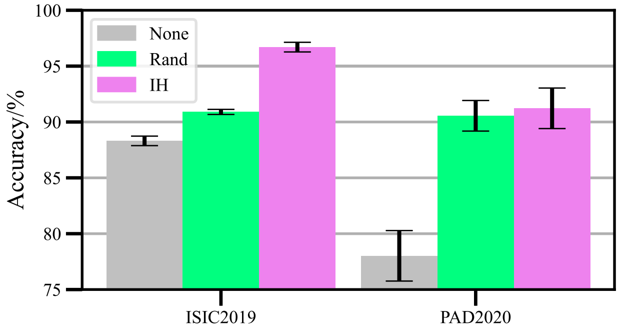


Fig. 9. Impact of different data balance methods.

TABLE III
IMPACT OF DIFFERENT ATTENTION MECHANISMS IN DWSCONV

| Attention mechanism | # Parameters/ Million | Inference time/s | Accuracy /% | AUC |
|---|---|---|---|---|
| - | 2.14 | **0.037** | 96.20 | 0.99596 |
| CBAM [27] | 2.17 | 0.043 | 95.77 | 0.99442 |
| CoorAttn [29] | 2.21 | 0.043 | 95.97 | 0.99453 |
| SEAttn [46] | 2.17 | 0.039 | 96.65 | 0.99682 |
| **SCAttn(Ours)** | **2.14** | 0.039 | **96.70** | **0.99772** |

The bold values mean the best performance.

## C. Ablation Studies

*1) Implementation Details:* In ablation studies, the model used is HierAttn_s, and the dataset used is ISIC2019, if not mentioned. Other parameters are the same as Section IV-B1.

*2) Data Balance Methods:* Fig. 9 provides a comprehensive overview of the accuracy achieved by different data balancing methods on the ISIC2019 and PAD2020 datasets. The undersampling methods, Rand (randomised sampling) and IH (instance hardness), demonstrate at least a 2.59% and 12.53% improvement in accuracy than the approach of having no data balancing (None), respectively, on ISIC2019 and PAD2020. Moreover, IH exceeds 5.9% and 0.7% accuracy than Rand in ISIC2019 and PAD2020, which means IH can more effectively sample images than Rand. The diminishing improvement in the PAD dataset could be caused by the low imbalance ratio of PAD compared to the ISIC dataset (16.3 for PAD vs. 53.9 for ISIC). The effects of alleviating negative influence on classification are more apparent for the dataset with a higher imbalance ratio. Thirdly, the error bar of ISIC2019 is shorter than PAD2020, indicating that HierAttn is more robust on ISIC2019 than PAD2020.

*3) Attention Blocks in DWSConv:* Table III shows the number of parameters, inference time and accuracy of models using different attention mechanisms in DWSConv. It can be discovered that without the attention mechanism [see Fig. 5(a)], the network only achieved only 96.20% accuracy. However, HierAttn with SCAttn achieves a notable improvement of +0.50% accuracy and +0.00176 AUC than the version without using the attention mechanism in DWSConv. Besides, HierAttn with SCAttn outperforms +0.05% accuracy and +0.00090 AUC than HierAttn with SEAttn. Besides, HierAttn with CBAM or CoorAttn exhibits a decrease of at least −0.23% accuracy and −0.00143 AUC than the version without using the attention mechanism in DWSConv, suggesting that CBAM or CoorAttn is not a suitable attention module in DWSConv for light networks.

TABLE IV
IMPACT ON THE PERFORMANCE UNDER DIFFERENT CONDITIONS (: CANCEL THE SETTING,: USE THE SETTING)

| Setting | Accuracy/% | | AUC | |
|---|---|---|---|---|
| | ✗ | ✓ | ✗ | ✓ |
| Warm-up | 96.28 | **96.70** | 0.99637 | **0.99772** |
| Stochastic depth | 96.05 | **96.70** | 0.99512 | **0.99772** |
| Skip connection | 96.47 | **96.70** | 0.99253 | **0.99772** |

The bold values mean the best performance.

*4) Transfer Learning Warm-Up:* We consider that the models are unstable after partially transferring tunable parameters and randomly initialising those layers without transferring them. Hence, we apply a warm-up technique to alleviate the influence of random weight initialisation for transferred layers. We freeze the transferred layers by stopping gradient backpropagation in the first 30 epochs. In these 30 epochs, only those layers without transferring can update their parameters. Table IV shows that the performance of the HierAttn_s with the transfer learning warm-up improves by 0.42% accuracy and 0.00135 AUC.

*5) Stochastic Depth:* Stochastic depth, also regarded as "layer dropout", is implemented in each layer with a skip connection in HierAttn. Typically, it is in all SCADW blocks with a stride of 1 and all CTH blocks. Table IV demonstrates that the stochastic depth effectively enhances the performance of HierAttn_s by 0.65% accuracy and 0.00260 AUC. Note that even without this stochastic depth, the performance of HierAttn_s delivers similar or better results than SOTA mobile models.

*6) Skip Connection of CTH Block:* We add a skip connection link in the CTH (convolution-transformer hybrid) block with a stride of 1. Moreover, we also apply stochastic depth to those modules with skip connections. Thus, we can reuse the lower lever feature and alleviate the gradient descent. Table IV shows a 0.23% accuracy and 0.00519 AUC improvement in the performance of HierAttn_s with the skip connection. It demonstrates that stochastic depth can reduce the negative influence of adding two distinctive features in the skip connection of the CTH block.

## V. DISCUSSION

From Table III, we can also discover that even without attention after depthwise convolution, HierAttn still achieves the most considerable accuracy, 96.20%, than other SOTA lightweight models on the ISIC2019 validation set. It suggests that branch attention could be preferable to obtaining information on critical stages. It also shows a high potential to use branch attention as a general method for improving the performance of different models. Applying branch attention in a model with more layers, e.g., 300, can prevent gradient descent in backpropagation.

Furthermore, additional experiments were also conducted to assess the versatility of HierAttn in analysing other medical objects, and the results indicate that HierAttn can achieve higher performance than the other SOTA mobile models. Since branch attention, SCAttn, and stage attention modules have the advantage of requiring fewer or no learnable parameters, these modules can be applied to real-time segmentation and tracking the focus of infection for broader medical applications in the future.

Mobile application for skin cancer diagnosis allows dermatologists to perform point-of-care testing. Moreover, possible patients can carry out further detection by utilising mobile applications while doing regular self-exam. HierAttn has a statistically close speed to classic mobile networks, which shows great potential to be developed on mobile devices. If the skin lesion is recognised as MEL, BCC, ACK, SCC or VASC with more than 50% possibility, users are suggested to go to a clinic or hospital to perform further diagnosis. Otherwise, it is more likely that the detected area of the skin is healthy. We expect that HierAttn can be deployed on the mobile phone to assist ordinary people in performing regular self-check in the future.

## VI. Conclusion

In this paper, we propose a HierAttn network consisting of stage attention, branch attention, and SCAttn for skin lesions diagnosis. Stage attention consists of a SCADW block for downsizing feature maps and a CTH block for effectively learning the local and global representations. Branch attention applies hierarchical pooling after each stage attention to learn local and global representations and improve the feature interactions. SCAttn directly extracts global features by using only global average pooling without redundantly operating channel-wise information. With these novel modules, HierAttn can achieve better skin lesion classification results, 96.70% accuracy and 0.9972 AUC on ISIC2019 and 91.22% accuracy and 0.98816 AUC on PAD2020 validation set, than other SOTA mobile networks. Moreover, HierAttn is the most miniature model among SOTA mobile networks, which shows great potential to be deployed in mobile devices for broader impacts on the general public.

## Acknowledgment

Wei Dai: Conceptualisation, Methodology, Validation, Writing. Rui Liu: Conceptualisation, Methodology. Tianyi Wu: Conceptualisation, Validation. Min Wang: Conceptualisation. Jianqin Yin: Writing. Jun Liu: Conceptualisation, Writing, Supervision.

## References

[1] G. P. Guy Jr., S. R. Machlin, D. U. Ekwueme, and K. R. Yabroff, “Prevalence and costs of skin cancer treatment in the U.S., 2002-2006 and 2007-2011,” *Amer. J. Prev. Med.*, vol. 48, no. 2, pp. 183–187, 2015.

[2] G. P. Guy Jr., C. C. Thomas, T. Thompson, M. Watson, G. M. Massetti, and L. C. Richardson, “Vital signs: Melanoma incidence and mortality trends and projections - United States, 1982-2030,” *MMWR. Morbidity Mortality Weekly Rep.*, vol. 64, no. 21, 2015, Art. no. 591.

[3] A. Esteva et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[4] H. Sung et al., “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[5] J. A. Avilés-Izquierdo, I. Molina-López, E. Rodríguez-Lomba, I. Marquez-Rodas, R. Suarez-Fernandez, and P. Lazaro-Ochaita, “Who detects melanoma? Impact of detection patterns on characteristics and prognosis of patients with melanoma,” *J. Amer. Acad. Dermatol.*, vol. 75, no. 5, pp. 967–974, 2016.

[6] L. J. Loescher, M. Janda, H. P. Soyer, K. Shea, and C. Curiel-Lewandrowski, “Advances in skin cancer early detection and diagnosis,” in *Seminars in Oncology Nursing*, vol. 29. Amsterdam, The Netherlands: Elsevier, 2013, pp. 170–181.

[7] M. E. Celebi, H. Iyatomi, G. Schaefer, and W. V. Stoecker, “Lesion border detection in dermoscopy images,” *Computerized Med. Imag. Graph.*, vol. 33, no. 2, pp. 148–153, 2009.

[8] F. Peruch, F. Bogo, M. Bonazza, V.-M. Cappelleri, and E. Peserico, “Simpler, faster, more accurate melanocytic lesion segmentation through MEDS,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 557–565, Feb. 2014.

[9] H. Zhou, G. Schaefer, M. E. Celebi, F. Lin, and T. Liu, “Gradient vector flow with mean shift for skin lesion segmentation,” *Computerized Med. Imag. Graph.*, vol. 35, no. 2, pp. 121–127, 2011.

[10] A. Adegun and S. Viriri, “Deep learning techniques for skin lesion analysis and melanoma cancer detection: A survey of state-of-the-art,” *Artif. Intell. Rev.*, vol. 54, no. 2, pp. 811–841, 2021.

[11] J. Shen, X. Tang, X. Dong, and L. Shao, “Visual object tracking by hierarchical attention siamese network,” *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3068–3080, Jul. 2020.

[12] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, *arXiv:1704.04861*.

[13] J. Pan et al., “EdgeViTs: Competing light-weight CNNs on mobile devices with vision transformers,” in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 294–311.

[14] S. Mehta and M. Rastegari, “MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer,” in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–26.

[15] W. Weng, J. Deaton, V. Natarajan, G. F. Elsayed, and Y. Liu, “Addressing the real-world class imbalance problem in dermatology,” in *Proc. Mach. Learn. Health*, 2020, pp. 415–429.

[16] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, “Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data,” *MethodsX*, vol. 7, 2020, Art. no. 100864.

[17] P. Tang, Q. Liang, X. Yan, S. Xiang, and D. Zhang, “GP-CNN-DTEL: Global-part CNN model with data-transformed ensemble learning for skin lesion classification,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2870–2882, Oct. 2020.

[18] D. Dai et al., “Ms RED: A novel multi-scale residual encoding and decoding network for skin lesion segmentation,” *Med. Image Anal.*, vol. 75, 2022, Art. no. 102293.

[19] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, “FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification,” *Med. Image Anal.*, vol. 76, 2022, Art. no. 102307.

[20] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao, “Hi-Net: Hybrid-fusion network for multi-modal MR image synthesis,” *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2772–2781, Sep. 2020.

[21] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, “Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification,” *Comput. Methods Programs Biomed.*, vol. 193, 2020, Art. no. 105475.

[22] M. Attique Khan, M. Sharif, T. Akram, S. Kadry, and C. Hsu, “A two-stream deep neural network-based intelligent system for complex skin cancer types classification,” *Int. J. Intell. Syst.*, vol. 37, pp. 10621–10649, 2021.

[23] L. Wang, C. Lee, Z. Tu, and S. Lazebnik, “Training deeper convolutional networks with deep supervision,” 2015, *arXiv:1505.02496*.

[24] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[25] Y. Liu and M. S. Lew, “Learning relaxed deep supervision for better edge detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 231–240.

[26] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[27] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[28] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[29] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.

[30] D. Fan et al., “Pranet: Parallel Reverse Attention Network for Polyp Segmentation,” in *Proc. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 263–273.

[31] X. Lu, W. Wang, J. Shen, D. Crandall, and J. Luo, “Zero-shot video object segmentation with co-attention siamese networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2228–2242, Apr. 2022.

[32] X. He, E. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, 2022, Art. no. 102357.

[33] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "FAT-Net: Feature adaptive transformers for automated skin lesion segmentation," *Med. Image Anal.*, vol. 76, 2022, Art. no. 102327.

[34] S. Suwanwimolkul and S. Komorita, "Efficient linear attention for fast and accurate keypoint matching," in*Proc. Int. Conf. Multimedia Retrieval*, 2022, pp. 330–341.

[35] C. Yang et al., "Lite vision transformer with enhanced self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11998–12008.

[36] J. Rashid et al., "Skin cancer disease detection using transfer learning technique," *Appl. Sci.*, vol. 12, no. 11, 2022, Art. no. 5714.

[37] S. ReiB, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9532–9542.

[38] J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. Hoi, "Distilled siamese networks for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, Dec. 2022.

[39] J. Song, Y. Chen, J. Ye, and M. Song, "Spot-adaptive knowledge distillation," *IEEE Trans. Image Process.*, vol. 31, pp. 3359–3370, 2022.

[40] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[41] Z. Zhao, S. Zhao, and J. Shen, "Real-time and light-weighted unsupervised video object segmentation network," *Pattern Recognit.*, vol. 120, 2021, Art. no. 108120.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[43] N. Ma, X. Zhang, H. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 116–131.

[44] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[45] M. Tan et al., "Platform-aware neural architecture search for mobile," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2820–2828.

[46] A. Howard et al., "Searching for MobileNetv3," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1314–1324.

[47] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

[48] M. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*.

[49] A. G. Pacheco et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," *Data Brief*, vol. 32, 2020, Art. no. 106221.

[50] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, 2018.

[51] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, no. 2, pp. 225–256, 2014.

[52] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 559–563, 2017.

[53] L. Zhang, X. Chen, J. Zhang, R. Dong, and K. Ma, "Contrastive deep supervision," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–19.

[54] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, 2022, Art. no. 102305.

[55] L. Zhu, "PyTorch-OpCounter," 2018. [Online]. Available: https://github.com/Lyken17/pytorch-OpCounter

[56] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.

[57] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1150–1210.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[60] I. Radosavovic, R. P. Kosaraju, R. P. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10428–10436.