# Attention Residual Learning for Skin Lesion Classification

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen

***Abstract*— Automated skin lesion classification in dermoscopy images is an essential way to improve the diagnostic performance and reduce melanoma deaths. Although deep convolutional neural networks (DCNNs) have made dramatic breakthroughs in many image classification tasks, accurate classification of skin lesions remains challenging due to the insufficiency of training data, inter-class similarity, intra-class variation, and the lack of the ability to focus on semantically meaningful lesion parts. To address these issues, we propose an attention residual learning convolutional neural network (ARL-CNN) model for skin lesion classification in dermoscopy images, which is composed of multiple ARL blocks, a global average pooling layer, and a classification layer. Each ARL block jointly uses the residual learning and a novel attention learning mechanisms to improve its ability for discriminative representation. Instead of using extra learnable layers, the proposed attention learning mechanism aims to exploit the intrinsic self-attention ability of DCNNs, i.e., using the feature maps learned by a high layer to generate the attention map for a low layer. We evaluated our ARL-CNN model on the ISIC-skin 2017 dataset. Our results indicate that the proposed ARL-CNN model can adaptively focus on the discriminative parts of skin lesions, and thus achieve the state-of-the-art performance in skin lesion classification.**

***Index Terms*— Attention learning, residual learning, skin lesion classification, dermoscopy images.**

## I. Introduction

SKIN cancer is one of the most common forms of cancers in the United States and many other countries, with 5 million cases occurring annually [1], [2]. Dermoscopy [3], [4], a recent technique of visual inspection that both magnifies the skin and eliminates surface reflection, is one of the essential means to improve diagnostic performance and reduce melanoma deaths [5]. Classifying skin lesions, particularly the melanoma, in dermoscopy images is a significant computer-aided diagnosis task.

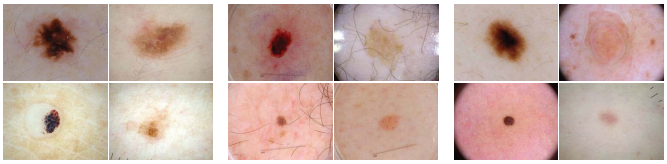


Fig. 1. Some typical samples show melanoma, nevus and seborrheic keratosis in skin lesion dermoscopy images.

A number of automated skin lesion classification methods have been proposed in the literature. Among them, deep learning solutions [8], [10], particularly those based on deep convolutional neural networks (DCNNs), have achieved significantly improved performance [28]–[30]. However, accurate classification of skin lesions remains a challenge due to three factors. First, the insufficiency of training samples limits the success of DCNNs in this task, as there is usually a small dataset in most medical imaging research, and this relates to the work required in acquiring the image data and then in annotation [6]. It is difficult for DCNNs to achieve the same success on skin lesion classification, which usually has only thousands of data, as they have done in the ImageNet Challenge [13], which has tens of millions of data. Second, the accuracy of skin lesion classification suffers from the inter-class similarity and intra-class variation [7], [30]. Skin lesion classification is much more complicated than classifying objects or scenes in natural images. As shown in Fig. 1, there is a significant visual difference among each group of four skin lesions in the same class, but visual similarities in shape and color between some lesions, which are from different classes. Such visual confusion makes it hard even for a human to distinguish the fine-grained lesion appearances without the expertise. Third, the region of a skin lesion occupies only a small part of a dermoscopy image, and most parts of the image are normal skin tissues, which are irrelevant but may have an interference with the lesion classification.

Many works [21], [22] have demonstrated that the DCNN trained for a classification task has a remarkable localization ability that can highlight the discriminative regions in images, despite being trained with only image-level labels, instead

Manuscript received December 8, 2018; revised January 8, 2019; accepted January 10, 2019. Date of publication January 21, 2019; date of current version August 30, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61771397 and Grant 61471297. The work of C. Shen was supported by the Australian Research Council (ARC) Future Fellowship. *(Corresponding author: Yong Xia.)*

J. Zhang, Y. Xie, and Y. Xia are with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: james.zhang@mail.nwpu.edu.cn; xuyongxie@mail.nwpu.edu.cn; yxia@nwpu.edu.cn).

C. Shen is with the School of Computer Science, University of Adelaide, Adelaide, SA 5005, Australia (chunhua.shen@adelaide.edu.au).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2019.2893944

of the bounding boxes of discriminative regions. Hence, we suggest strengthening the discriminative ability of a DCNN via taking advantage of its self-attention ability. Since the higher layers in a DCNN have a better ability for semantic abstraction than lower ones, it might be possible to use the feature maps obtained by higher layers as the attention mask of lower ones. Meanwhile, the residual network [23] is more suitable for small-sample learning problems than other DCNNs, such as AlexNet [48], VGG [49], and GoogLeNet [50], since it uses "shortcut connections" to skip one or more layers, and thus enable the construction of a deeper network.

In this paper, we propose an attention residual learning convolutional neural network (ARL-CNN) model for the skin lesion classification. We jointly use the residual learning mechanism to train a DCNN with a small set of dermoscopy images and a novel attention learning mechanism to strengthen the discriminative representation ability of the DCNN via enabling it to focus more on semantically meaningful parts (i.e. lesions) in dermoscopy images. The proposed attention learning mechanism makes full use of the intrinsic and remarkable self-attention ability of classification-trained DCNNs and can work well under any DCNN frameworks without adding any extra attention layers, which is critical for small-sample learning problems like skin lesion classification. From an implementation perspective, both residual learning and attention learning can be embedded in each so-called ARL block. An ARL-CNN model with an arbitrary depth can be constructed by stacking multiple ARL blocks and be trained in an end-to-end manner. We evaluated the proposed ARL-CNN model on the ISIC-skin 2017 dataset [5] which is a largest publicly available skin dermoscopy image achieve, and achieved the state-of-the-art performance (i.e., an average AUC of 0.917).

The main contributions of this paper are thus summarized as follows: (1) we propose a novel ARL-CNN model for accurate skin lesion classification in dermoscopy images, which embeds simultaneously two learning mechanisms - residual learning and attention learning. Residual learning enables the network to become deep, and attention learning helps the network focus more on semantically important regions and thus improves its ability for discriminative representation; (2) we design an effective attention mechanism which takes advantages of the intrinsic self-attention ability of DCNNs, i.e., using the feature maps obtained by a high layer as the attention mask of a low layer, instead of learning the attention mask with extra layers; and (3) we achieve the state-of-the-art skin lesion classification performance on the ISIC-skin 2017 dataset by using a single 50-layer model, which is important for computer-aided diagnosis of skin cancer.

## II. Related Work

### A. DCNN Models

In recent years, DCNNs have achieved the state-of-the-art performance in many computer vision applications, including image classification [23], [51], [54], [55], target detection [24], [25] and image segmentation [26], [27]. There are two basic operations in DCNNs - convolution and pooling. Convolutional layers apply a set of convolutional kernels to the input with the mechanism of sharing weights, which allows the DCNNs to be deeper with fewer parameters. Pooling layers use various pooling operations, such as the max-pooling, min-pooling, and average-pooling, which are indeed nonlinear down-sampling, to reduce the spatial dimension of the learning representation, number of parameters, and amount of computation.

Many successful DCNN models [23], [48]–[51] have demonstrated that their performance depends heavily on the network depth. The increased availability of large scale datasets, powerful computing devices, and computational tricks, such as the rectified linear units activation [37], dropout [40], batch normalization [38] and layer normalization [39], make it possible to design and train deep networks. However, due to degradation problems [41], [42], it is still hard to train a very deep network. The residual learning technique [23] successfully addressed this issue by using "shortcut connections", and hence enables the training of a DCNN with as many as 1,000 layers.

### B. Attention Mechanism Used in Classification

The attention mechanism is an effective technique that helps a model pay more attention to important information. It has made great progress in the cross fields of computer vision and natural language processing, such as image/video caption and visual question answering [14]–[18]. Recently, the attention mechanism has also been successfully used in DCNNs [17], [19], [20] to improve their feature representation ability in large scale image classification tasks. Wang *et al.* [19] proposed a residual attention network, which is constructed by stacking multiple attention modules in a residual network. Each attention module uses trainable layers with a bottom-up and top-down feedforward structure to learn soft weights, and then multiplies the weights with convolutional features. Hu *et al.* [20] recalibrated adaptively channel-wise feature responses by explicitly modelling the channel-wise interdependencies of convolutional features. They generated the attentive features using channel-wise multiplication between the attention weights learned by two additional fully-connected layers and the original feature map. Chen *et al.* [17] jointly used spatial and channel-wise attentions in convolutional networks. The spatial and channel-wise attention weights are generated by a neural network followed by a softmax layer, respectively. Although the attention mechanisms effectively improve the performance of deep learning models in large scale image classification tasks, the attention weights in these methods are learned by using additional learnable layers with a lot of extra parameters, which may cause not only computational costs but also overfitting on small training datasets.

### C. Skin Lesion Classification

Many skin lesion classification solutions are based on hand-crafted features, including color, texture, shape, and combined descriptors of lesions [43]–[46]. Barata *et al.* [45] proposed a global and local method to classify skin lesions by extracting a set of color and texture features and using them to train a classifier. Xie *et al.* [46] extracted the color, texture, and

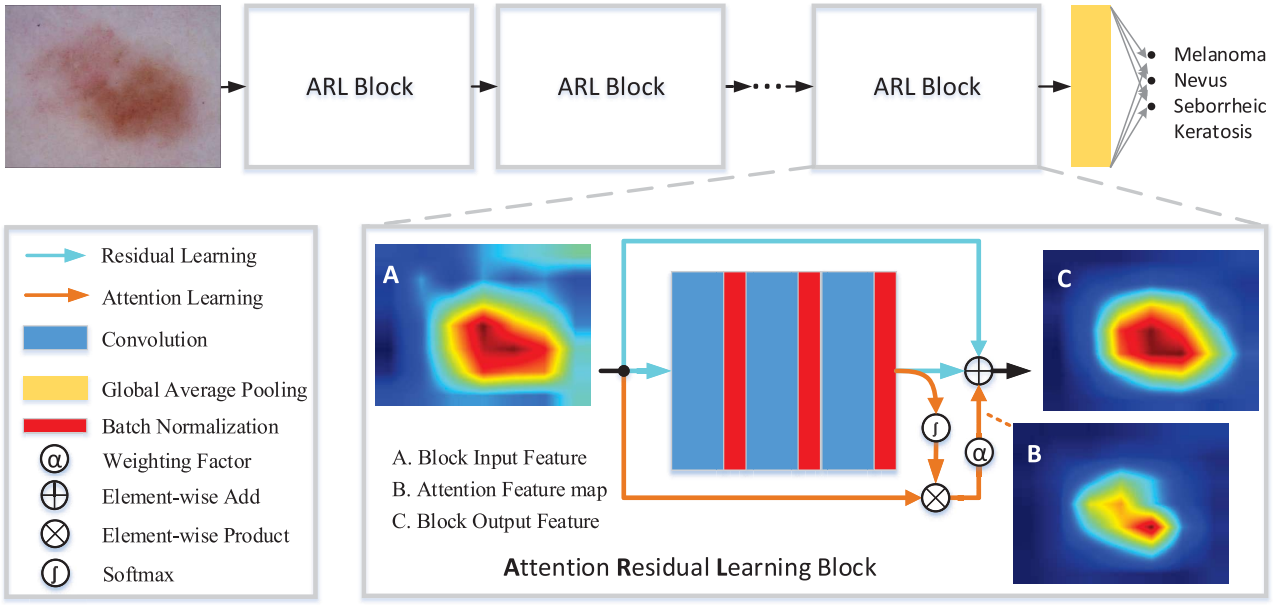


Fig. 2. Architecture of the proposed ARL-CNN model.

border descriptors of skin lesions and used an ensemble of neural networks to classify lesions.

To apply deep learning techniques directly to skin lesion classification, a straightforward solution is to collect more training data and fine-tune a powerful pre-trained DCNN [9]–[12]. Esteva *et al.* [28] trained a DCNN using 129,450 clinical images for diagnosing the most common and deadliest skin cancers and achieved the performance that matches the performance of 21 board-certified dermatologists. Ge *et al.* [29] jointly used the dermoscopy and clinical skin images to train DCNNs and demonstrated the effectiveness of cross-modality learning. Matsunaga *et al.* [31] proposed an ensemble of multiple pre-trained DCNNs with geometrically transformed images for the classification of melanoma, nevus and seborrheic keratosis. Menegola *et al.* [33] also improved the performance of skin lesion classification using pre-trained DCNNs and as much as possible training data.

To filter out the useless background, Yu *et al.* [30] leveraged very deep DCNNs for automated melanoma recognition in two steps - segmentation and classification, and found that lesion segmentation benefits the classification. Diaz [32] also designed several convolutional networks that incorporate lesion segmentation and structure segmentation into the diagnosis of skin lesions. However, since the malignancy of skin lesion usually relates only to a part of the lesion, image features extracted in the entire lesion may lead to sub-optimal results. Meanwhile, these methods often requires the ground truth boundaries of lesions being marked in training data, which makes it even more difficult to obtain a large training dataset.

## III. Method

The proposed ARL-CNN model is composed of multiple ARL blocks, a global average pooling (GAP) layer and a classification layer. In each ARL block, the residual learning mechanism is employed to address the degradation problem,

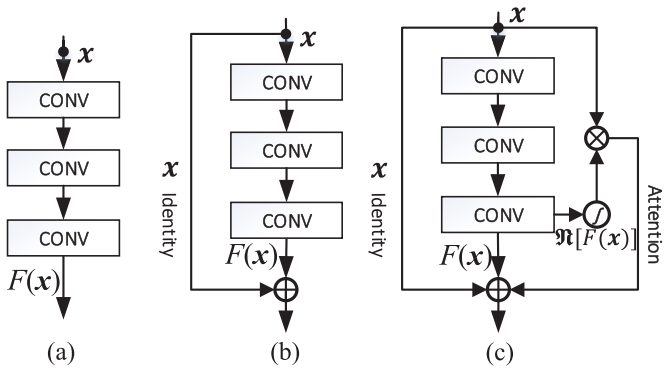


Fig. 3. Three types of learning blocks used in DCNNs. (a) Plain block. (b) Residual block. (c) ARL block.

and a novel attention mechanism is designed to strengthen the discriminative representation ability. The architecture of this model is shown in Fig. 2. We now delve into the details of this model.

### A. ARL Block

*1) Residual Learning:* Let us consider a plain DCNN block which is composed of a few stacked convolution layers, as shown in Fig. 3(a). The underlying mapping fitted by these layers is denoted as $H(\boldsymbol{x})$. These stacked layers are expected to directly approximate $H(\boldsymbol{x})$. Hence, a plain block is defined as follows

$$\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{W}) \tag{1}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent the input and output of the block, respectively, and $F(\cdot)$ represents the underlying mapping function which is learned by these stacked layers with the parameter set $\boldsymbol{W}$. In residual learning, we let these stacked layers learn a residual function $F(\boldsymbol{x}) := H(\boldsymbol{x}) - \boldsymbol{x}$, instead of directly approximating $H(\boldsymbol{x})$. The definition of residual

learning is

$$\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{W}) + \boldsymbol{x} \tag{2}$$

where the function $F(\cdot)$ represents the residual mapping learned by these stacked layers. The formulation of residual learning can be implemented by the feedforward network with shortcut connections, as shown in Fig. 3(b).

*2) Attention Learning:* The traditional attention mechanisms designed for image classification are to learn the attention weights by using extra learnable layers, such as the convolutional layers used in [19] or the fully connected layers used in [20]. Different from these solutions, we propose a novel attention learning method to strengthen the discriminative representation of the network by generating the attention weights from the classification-trained network itself, without introducing extra learnable layers. Given that higher layers have a better ability for semantic abstraction than lower ones, we suppose that the attention ability of higher layers is stronger than that of lower layers. Hence, we propose to use the more abstract feature maps produced by higher layers as the attention mask of lower layers.

We denote a set of stacked layers in a DCNN as $\{L^{i+1}, \ldots, L^{i+n}\}$, where $\{L^{i+1}\}$ is the low layer and $\{L^{i+n}\}$ is the corresponding high layer. The output feature map of $\{L^{i+n}\}$ is defined as $\boldsymbol{O}$ with a dimension of $H \times W \times C$, where $H$, $W$ and $C$ represent the height, width and number of channels, respectively. The attention mask $\boldsymbol{\varpi}$ can be generated by applying a normalization function $\mathfrak{N}$ to the feature map $\boldsymbol{O}$

$$\boldsymbol{\varpi} = \mathfrak{N}(\boldsymbol{O}) \tag{3}$$

Three normalization functions are defined as follows

$$\mathfrak{N}^S(\boldsymbol{O}) = \{m | m^c_{i,j} = \frac{e^{\boldsymbol{O}^c_{i,j}}}{\sum_{i',j'} e^{\boldsymbol{O}^c_{i',j'}}}\} \tag{4}$$

$$\mathfrak{N}^C(\boldsymbol{O}) = \{m | m^c_{i,j} = \frac{e^{\boldsymbol{O}^c_{i,j}}}{\sum_{c'} e^{\boldsymbol{O}^{c'}_{i,j}}}\} \tag{5}$$

$$\mathfrak{N}^M(\boldsymbol{O}) = \{m | m^c_{i,j} = \sigma(\boldsymbol{O}^c_{i,j})\} \tag{6}$$

where $i$, $j$ represent the spatial position, $c$ represents the channel index of $\boldsymbol{O}$, and $\sigma(\cdot)$ is the sigmoid function. The spatial attention $\mathfrak{N}^S(\cdot)$ uses a spatial softmax function and highlights the important regions in each channel. The channel attention $\mathfrak{N}^C(\cdot)$ uses a softmax function to perform the normalization in the channel space. The mixed attention $\mathfrak{N}^M(\cdot)$ uses the simple sigmoid normalization at each spatial and channel position. Since we expect the model to focus on semantic regions of skin lesions, the spatial attention $\mathfrak{N}^S(\cdot)$ is adopted and the comparison given in the Section V.A also shows that the spatial attention performs better than other two attentions in the skin lesion classification task.

Then, we use the attention mask $\boldsymbol{\varpi}$ as the control gates of input neurons $\boldsymbol{x}$ of $\{L^{i+1}\}$ which are similar to the gates used in the Highway Network [42]. The attention feature map is computed by multiplying $\boldsymbol{x}$ with $\boldsymbol{\varpi}$ on an element-by-element basis.

$$\boldsymbol{A} = \boldsymbol{\varpi} \cdot \boldsymbol{x} \tag{7}$$

This attention learning method strengthens the ability of DCNNs to focus on semantically meaningful regions without adding extra parameters, which is critical for small-sample learning tasks.

*3) ARL Block:* To train a deep DCNN with an improved ability for discriminative representation, we propose the ARL-CNN model, which embeds both residual learning and attention learning mechanisms. The architecture of an ARL block, a basic module of this model, is displayed in Fig. 3(c). It shows that we simultaneously use the identity mapping as designed in the residual block and generate an attention mask for the original input $\boldsymbol{x}$ via applying the spatial normalization $\mathfrak{N}^S$ to the feature map $F(\boldsymbol{x})$ obtained by the residual mapping. Then, we get the attention feature map through the following element-wise production

$$\boldsymbol{A} = \mathfrak{N}[F(\boldsymbol{x})] \cdot \boldsymbol{x} \tag{8}$$

The output of an ARL block is an element-wise addition of the identity map, residual feature map, and attention feature map, shown as follows

$$\boldsymbol{y} = \boldsymbol{x} + \underbrace{F(\boldsymbol{x})}_{\textit{residual feature map}} + \alpha \cdot \underbrace{\mathfrak{N}[F(\boldsymbol{x})] \cdot \boldsymbol{x}}_{\textit{attention feature map}} \tag{9}$$

where $\alpha$ is a learnable weighting factor called the scale gate that represents a trade-off between the attention feature map and other two maps. Since a well-trained DCNN has a stronger attention ability than a poorly trained DCNN, the scale gate is able to adaptively adjust the contribution of the attention feature map, avoiding the interference by a bad attention feature map obtained at the early stages of model training.

## B. ARL-CNN

An ARL-CNN model with an arbitrary depth can be easily constructed by stacking ARL blocks. We introduce a lightweight version called ARL-CNN14 and a heavyweight version called ARL-CNN50 for skin lesion classification. ARL-CNN14 has 14 learnable layers, and ARL-CNN50 has 50 learnable layers.

TABLE I gives the architectures of the ResNet14, ARL-CNN14, ResNet50 and ARL-CNN50 models. Both ResNet and ARL-CNN are stacked from a $224 \times 224 \times 3$ input layer, a $7 \times 7$ convolutional layer, a max-pooling layer, a series of residual or ARL blocks, a GAP layer and a fully-connected (FC) classification layer. The difference is that the proposed ARL-CNN model replaces the residual blocks used in ResNet with ARL blocks. Each ARL block is stacked by a fixed mode of $1 \times 1$, $3 \times 3$ and $1 \times 1$ convolutional layers followed by batch normalization layers. ARL-CNN14 can be trained from scratch for skin lesion classification, and ARL-CNN50 can be initialized by using a ResNet50, which has been trained on the ImageNet dataset, and fine-tuned with the skin lesion data. The pre-trained technique can not only improve the performance, but also reduce the training time.

TABLE I
ARCHITECTURES OF RESNET14, ARL-CNN14, RESNET50 AND ARL-CNN50 MODELS

| Layer name | Output size | ResNet14 | ARL-CNN14 | ResNet50 | ARL-CNN50 |
|---|---|---|---|---|---|
| Conv1 | 112x112x64 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 |
| Conv2_x | 56x56x256 | 3x3 max pool, stride 2 | 3x3 max pool, stride 2 | 3x3 max pool,stride 2 | 3x3 max pool,stride 2 |
| | | Residual block | ARL | [Residual block]x3 | [ARL]x3 |
| Conv3_x | 28x28x512 | Residual block,stride 2 | ARL,stride 2 | Residual block, stride 2<br>[Residual block]x3 | ARL, stride 2<br>[ARL]x3 |
| Conv4_x | 14x14x1024 | Residual block,stride 2 | ARL,stride 2 | Residual block, stride 2<br>[Residual block]x5 | ARL, stride 2<br>[ARL]x5 |
| Conv5_x | 7x7x2048 | Residual block,stride 2 | ARL,stride 2 | Residual block, stride 2<br>[Residual block]x2 | ARL, stride 2<br>[ARL]x2 |
| GAP | 1x1x2018 | Global Average Pooling | Global Average Pooling | Global Average Pooling | Global Average Pooling |
| Output | 2 | Fully Connected | Fully Connected | Fully Connected | Fully Connected |

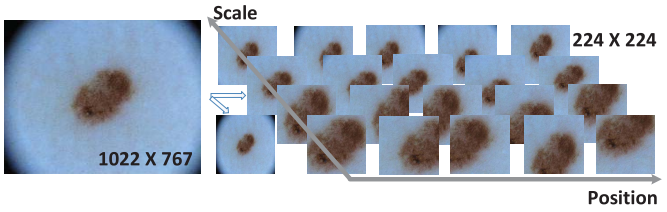


Fig. 4. An illustration of multi-scale patch extraction on dermoscopy images.

## IV. EXPERIMENTS AND RESULTS

### A. Dataset

The proposed ARL-CNN model was evaluated on the International Skin Imaging Collaboration 2017 skin lesion classification (ISIC-skin 2017) dataset [5], which is the largest skin dermoscopy image dataset publicly available, consisting of 2000 training, 150 validation, and 600 test images screened for both privacy and quality assurance. Lesions in dermoscopy images are all paired with a gold standard (definitive) diagnosis, i.e. melanoma, nevus, and seborrheic keratosis. There are two binary classification sub-tasks - melanoma classification (melanoma vs. others) and seborrheic keratosis classification (seborrheic keratosis vs. others). We also collected 1320 additional dermoscopy images, including 466 melanoma, 822 nevus images, and 32 seborrheic keratosis images, from the ISIC Archive[1] to enlarge the training dataset.

### B. Implementation

Since the proposed ARL-CNN model takes $224 \times 224$ images as input, all dermoscopy images should be shrunk to this size before they can be fed into our model. However, skin lesions occupy only a small part of an image, and shrinking the image may lead the lesions to becoming too small to be classified. To address this issue, we randomly extracted 60 rectangular image patches from the central part of each image at different scales (1/5, 2/5, 3/5, and 4/5 of original image size) on both official and extra training images, and then resized them to $224 \times 224$ using the bilinear interpolation, as shown in Fig. 4. Next, we employed online data augmentation, including random rotation ($[-10°, +10°]$), zoom (90%-110% of width and height), horizontal and vertical flips, to enlarge the training dataset.

The mini-batch SGD algorithm with a batch size of 32 was adopted as the optimizer. The learning rate was initialized to 0.01 for training ARL-CNN14 from scratch and 0.0001 for fine-tuning ARL-CNN50 with pre-trained parameters, and was reduced by half very 30 epochs. The initial weighting factor of the attention feature maps was set to 0.001 in each ARL block when fine-tuning the ARL-CNN50. The maximum epoch number was set to 100. We used the officially provided validation set to monitor the performance of our model and stopped the training process when the network fell into overfitting. In the test stage, we used the same patch extraction method to randomly crop nine patches from each test image, fed them to the trained network, and averaged the obtained scores as the predicted score of the image.

### C. Evaluation Metrics

*1) Quantitative Evaluation:* To quantitatively evaluate the proposed ARL-CNN model, we used the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) as performance metrics, which are defined as

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{10}$$

$$sensitivity = \frac{TP}{TP + FN}, \quad specificity = \frac{TN}{TN + FP} \tag{11}$$

$$AUC = \int_0^1 t_{pr}(f_{pr}) df_{pr} = P(X1 > X0) \tag{12}$$

where $TP$, $FN$, $TN$ and $FP$ represent the number of true positive, false negative, true negative and false positive, respectively, $t_{pr}$ is the true positive rate, $f_{pr}$ is the false positive rate, and $X0$ and $X1$ are the confidence scores for a negative and positive instance, respectively. The AUC value describes the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative one. The ISIC 2017 skin lesion classification challenge used the AUC value as a gold indicator, according to which all participants were ranked [5].

[1] https://isic-archive.com/

TABLE II
ARCHITECTURES OF RAN14, SENET14, RAN50 AND SENET50

| Layer name | Output size | RAN14 | SEnet14 | RAN50 | SEnet50 |
|---|---|---|---|---|---|
| Conv1 | 112x112x64 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 | Conv, 7x7, stride 2 |
| Conv2_x | 56x56x256 | Max pool | Max pool | Max pool | Max pool |
| | | Attention Module<br>Residual Block | SE Block | Attention Module<br>[Residual Block]x3 | [SE Block]x3 |
| Conv3_x | 28x28x512 | Attention Module<br>Residual Block | SE Block | Attention Module<br>[Residual Block]x4 | [SE Block]x4 |
| Conv4_x | 14x14x1024 | Attention Module<br>Residual Block | SE Block | Attention Module<br>[Residual Block]x6 | [SE Block]x6 |
| Conv5_x | 7x7x2048 | Attention Module<br>Residual Block | SE Block | [Residual Block]x3 | [SE Block]x3 |
| GAP | 1x1x2048 | GAP | GAP | GAP | GAP |
| FC | 2 | Fully Connected | Fully Connected | Fully Connected | Fully Connected |

*2) Qualitative Evaluation:* We adopted the class activation mapping (CAM) [21] to visualize the attention regions in obtained feature maps. We defined the output of a GAP layer as $f_c(i, j)$, where $c$ represents the index of channel and $i$, $j$ is the index of spatial positions, and denoted the CAM for class $k$ as $\mathbb{C}_k$. Then each element of $\mathbb{C}_k$ can be calculated as

$$\mathbb{C}_k(i, j) = \sum_c w_c^k \cdot f_c(i, j) \tag{13}$$

where $w_c^k$ is the weight corresponding to class $k$ for the channel $c$. Each spatial element $\mathbb{C}_k(i, j)$ reflects the contribution of the spatial position $(i, j)$ to the classification of the input into the category $k$. We can visualize the CAM to validate which part of the input image plays an important role in the classification.

## D. Comparing to Baseline and Attention Methods

Since the proposed ARL-CNN model uses both residual learning and attention learning, we compared it to the corresponding ResNet, which is a baseline, and two state-of-the-art attention models - RAN [19] and SEnet [20]. Our ARL-CNN50 can be easily initialized by transferring the parameters from a pre-trained ResNet50 to it since they have the same parameter structure. However, both RAN and SEnet have a lot of extra parameters, which cannot be initialized by using a pre-trained ResNet. To make a fair comparison, we evaluated the lightweight versions (with 14 learnable layers) of these models and trained them from scratch with the same parameter settings, including the SGD optimizer, initial learning rate, decay of learning rate, and maximum epoch number. Architectures of the RAN and SEnet used for this study were shown in TABLE II. The designs of "Squeeze-and-Excitation" (SE) blocks in SEnet and attention modules in RAN was adopted from [19] and [20].

In TABLE III, we compared the proposed ARL-CNN model, including the lightweight ARL-CNN14 and heavyweight ARL-CNN50, with the baseline ResNet model and state-of-the-art attention models in the melanoma classification and seborrheic keratosis classification. The second column in this table gives the number of model parameters. It shows that our ARL-CNN model has almost the same number of parameters as the baseline ResNet, but both RAN and SEnet have much more parameters than the baseline due to the additional learnable attention layers. The increased number of parameters makes it difficult to train a very deep model for skin lesion classification, which is a small-sample learning task. Other columns in this table give the classification performance in two sub-tasks. First, we compare four lightweight models which were trained from scratch. It reveals that three attention models, including RAN14, SEnet14, and ARL-CNN14, have substantially improved performance over the baseline ResNet14 model. Moreover, our ARL-CNN14 achieved the highest AUC, ACC, Sensitivity and Specificity in the melanoma classification and highest AUC, ACC, Specificity and second highest Sensitivity in the seborrheic keratosis classification among all lightweight models. Then, we compared the heavyweight ARL-CNN50 model with the baseline ResNet50 and other heavyweight attention models. Note that the ARL-CNN50 model and RAN50 model were initialized with the corresponding parameters from ResNet50, which has been well trained in the ImageNet classification dataset. The additional layers designed for attention learning in RAN50 had to be trained from scratch. Besides, the SEnet50 model was initialized from the ImageNet pre-trained model provided in [20]. It shows that, compared to lightweight models, a deeper architecture and the pre-training technique contribute significantly to the classification performance in both sub-tasks. However, the RAN50 model did not produce much improvement in AUC, since the additional layers in RAN50 were trained from scratch to learn attention weights, which may be inaccurate due to the insufficient training data. Compared to these attention methods, the proposed ARL-CNN model can achieve much better performance, particularly an AUC of 0.875 in the melanoma classification and an AUC of 0.958 in the seborrheic keratosis classification.

## E. Comparing to Challenge Records

In TABLE IV, we compared the performance of the proposed ARL-CNN50 model to six top-ranking performances in the ISIC-2017 skin lesion classification challenge

TABLE III

COMPARISON OF THE PROPOSED ARL-CNN MODEL WITH THE BASELINE MODELS (RESNET14 AND RESNET50), AND STATE-OF-THE-ART ATTENTION METHODS (RAN14, SENET14, RAN50 AND SENET50) IN MELANOMA AND SEBORRHEIC KERATOSIS CLASSIFICATIONS. THE MODELS WITH "*" WERE INITIALIZED WITH THE IMAGENET PRE-TRAINED MODEL, AND OTHERS WERE RANDOMLY INITIALIZED. THE BEST PERFORMANCES ACHIEVED BY LIGHTWEIGHT AND HEAVYWEIGHT MODELS WERE HIGHLIGHTED IN "BLACK BOLD" AND "RED BOLD", RESPECTIVELY

| Methods | Params ($\times 10^7$) | Melanoma Classification | | | | Seborrheic Keratosis Classification | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUC | ACC | Sensitivity | Specificity | AUC | ACC | Sensitivity | Specificity |
| ResNet14 [23] | 0.8 | 0.732 | 0.748 | 0.538 | 0.799 | 0.820 | 0.711 | 0.800 | 0.696 |
| RAN14 [19] | 1.2 | 0.767 | 0.762 | **0.615** | 0.797 | 0.852 | 0.758 | **0.833** | 0.745 |
| SEnet14 [20] | 1.1 | 0.758 | 0.757 | 0.598 | 0.795 | 0.847 | 0.727 | 0.811 | 0.712 |
| ARL-CNN14 | 0.8 | **0.777** | **0.778** | **0.615** | **0.818** | **0.875** | **0.763** | 0.822 | **0.753** |
| ResNet50* [23] | 2.3 | 0.857 | 0.838 | 0.632 | 0.888 | 0.948 | 0.842 | 0.867 | 0.837 |
| RAN50* [19] | 3.9 | 0.849 | **0.850** | 0.624 | **0.906** | 0.942 | 0.862 | **0.878** | 0.859 |
| SEnet50* [20] | 2.6 | 0.861 | 0.848 | 0.624 | 0.903 | 0.952 | 0.863 | 0.856 | 0.865 |
| ARL-CNN50* | 2.3 | **0.875** | **0.850** | **0.658** | 0.896 | **0.958** | **0.868** | **0.878** | **0.867** |

TABLE IV

COMPARISON BETWEEN THE PERFORMANCE OF OUR ARL-CNN50 MODEL AND THE TOP SIX ISIC 2017 CHALLENGE RECORDS. FOR EACH PERFORMANCE METRIC, THE HIGHEST AND SECOND HIGHEST VALUES WERE HIGHLIGHTED IN "RED BOLD" AND "BLACK BOLD", RESPECTIVELY. NOTE THAT THE AVERAGE AUC OF BOTH CLASSIFICATION SUB-TASKS IS THE GOLD EVALUATION METRIC, ACCORDING TO WHICH ALL PARTICIPANTS WERE RANKED

| Methods | External data | Ensembles | Melanoma Classification | | | | Seborrheic Keratosis Classification | | | | Average AUC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AUC | ACC | Sensitivity | Specificity | AUC | ACC | Sensitivity | Specificity | |
| Ours | 1320 | N | **0.875** | 0.850 | **0.658** | 0.896 | **0.958** | 0.868 | **0.878** | 0.867 | **0.917** |
| #1 [31] | 1444 | Y | 0.868 | 0.828 | **0.735** | 0.851 | 0.953 | 0.803 | **0.978** | 0.773 | **0.911** |
| #2 [32] | 900 | N | 0.856 | 0.823 | 0.103 | **0.998** | **0.965** | 0.875 | 0.178 | **0.998** | 0.910 |
| #3 [33] | 7544 | Y | **0.874** | **0.872** | 0.547 | 0.950 | 0.943 | 0.895 | 0.356 | 0.990 | 0.908 |
| #4 [34] | 1600 | Y | 0.870 | **0.858** | 0.427 | 0.963 | 0.921 | **0.918** | 0.589 | 0.976 | 0.896 |
| #5 [36] | 1341 | Y | 0.836 | 0.845 | 0.350 | **0.965** | 0.935 | 0.913 | 0.556 | 0.976 | 0.886 |
| Ours | 0 | N | 0.859 | 0.837 | 0.590 | 0.896 | 0.951 | 0.908 | 0.778 | 0.931 | 0.905 |
| #6 [35] | 0 | N | 0.830 | 0.830 | 0.436 | 0.925 | 0.942 | **0.917** | 0.700 | **0.995** | 0.886 |

leaderboard [31]–[36]. Almost all the methods listed in TABLE IV (including those reported in [31]–[34] and [36]) were trained with external dermoscopy images to boost their classification performance. Especially, Menegola *et al.* [33] trained the deep model with up to 7,500 external images. Besides, the ensemble strategy was employed in [31], [33], [34], and [36] for an extra performance gain. Actually, these methods cannot be compared directly with each other due to the differences in the training dataset and whether it is an ensemble or not. However, these reported results on the ISIC-2017 challenge dataset can, to some extent, reflect the state-of-the-art performance in the skin lesion classification task. To compare our model to the state of the art while keeping this comparison informative enough, we provided the number of external training data and whether using ensemble learning as a reference in TABLE IV.

It shows that our ARL-CNN50 model, which were trained on the ISIC-2017 training dataset and 1320 additional dermoscopy images, achieved the highest AUC and the second highest sensitivity in melanoma classification, the second highest AUC and sensitivity in seborrheic keratosis classification. Although [32] achieved the highest AUC and specificity in seborrheic keratosis classification, this solution has an extremely low sensitivity. According to the ranking rule of the challenge, our ARL-CNN50 achieved an average AUC of 0.917 in two sub-tasks, which is higher than the top-ranking performance listed in the leaderboard and is, to our knowledge, the best skin lesion classification performance on the ISIC-skin 2017 dataset. More importantly, our model achieved the state-of-the-art performance using only a single network with 50 learnable layers, which requires less computation resources and training time than ensemble models.

Meanwhile, we also compared our model to the one presented in [35]. Both models use neither ensemble learning nor additional training data. In this scenario, our ARL-CNN50 model attained an average AUC of 0.905, which is noticeably higher than that reported in [35].

### F. Visualization of CAM

The proposed ARL-CNN model shows an excellent performance in skin lesion classification, substantially better than the performance of their ResNet counterparts (see TABLE III). We suppose that the performance gain is mainly attributed

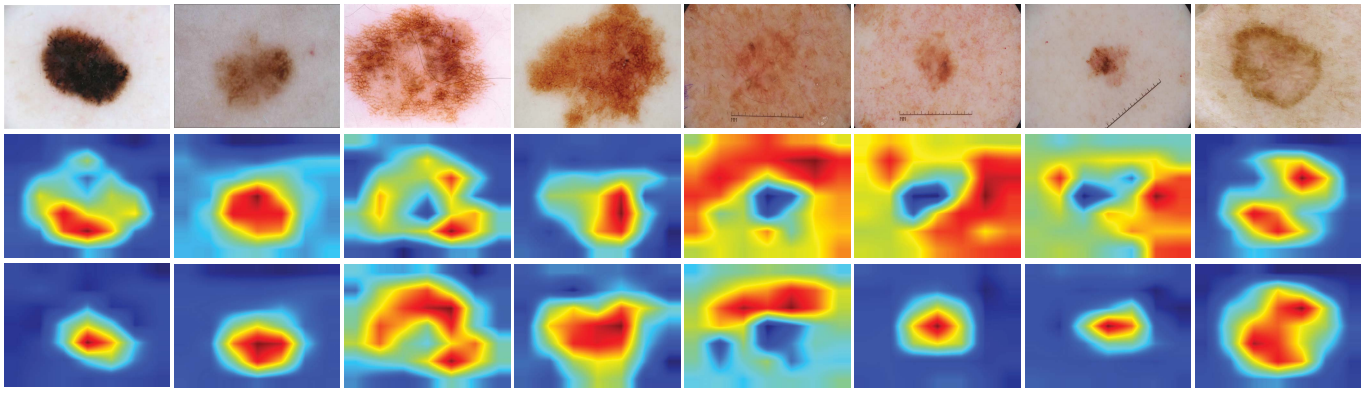

Fig. 5. Visualization of dermoscopy images (top row) and the corresponding CAMs obtained by ResNet50 (middle row) and our ARL-CNN50 (bottom row).

TABLE V
PERFORMANCE OF ARL-CNN50 IN TWO SUB-TASKS WHEN USING DIFFERENT ATTENTION NORMALIZATION METHODS

| Methods | Melanoma Classification | | | | Seborrheic Keratosis Classification | | | |
|---|---|---|---|---|---|---|---|---|
| | AUC | ACC | Sensitivity | Specificify | AUC | ACC | Sensitivity | Specificify |
| Channel-wise softmax | 0.861 | 0.847 | 0.607 | **0.905** | 0.952 | 0.852 | 0.867 | 0.849 |
| Mixed Sigmoid | 0.851 | 0.832 | 0.641 | 0.878 | 0.951 | **0.898** | 0.811 | **0.913** |
| Spatial-wise softmax | **0.875** | **0.850** | **0.658** | 0.896 | **0.958** | 0.868 | **0.878** | 0.867 |

to the use of attention learning, which enables a DCNN to focus more on semantically meaningful parts of lesions and thus strengthens the network's ability to learn discriminative representation. To validate this, we visualized the CAMs obtained by ResNet50 and ARL-CNN50 in Fig. 5. It shows that the attention regions learned by both models, i.e. the highlights in CAMs, have different positions and concentrations. Compared to ResNet50, our ARL-CNN50 model shows a stronger attention ability that highlights the discriminative lesion parts instead of background tissues in dermoscopy images, especially in the samples given in the $6^{th}$ and $7^{th}$ columns.

Although highlighting background regions, a powerful ResNet50 model could rely on its data fitting ability to easily remember those examples during the training process [47]. However, the generalization ability of such model is poor. We hope that a classification model could pay more attention to the regions, which have better discriminatory power, instead of focusing on irrelevant normal tissues. With a stronger attention ability, the proposed ARL-CNN model makes more reliable and accurate classification based on discriminative regions. Hence, it may explain why the proposed ARL-CNN model has a better classification performance than the corresponding ResNet model.

## V. DISCUSSION

### A. Attention Mask Normalization

In the proposed ARL-CNN model, we use the spatial softmax, which normalizes a soft mask in the spatial domain, to generate the soft attention mask $\boldsymbol{\varpi}$ from the high-layer feature map $\boldsymbol{O}$. Certainly, other mask normalization methods can also be used. In Table V, we compared the classification performance of ARL-CNN50 in two sub-tasks when using different attention soft mask normalization methods, including the channel-wise softmax, mixed sigmoid, and spatial-wise softmax. It shows that the spatial softmax method achieved the best performance in most metrics.

### B. Attention Regions Learned From Different Tasks

In Fig. 6, we displayed eight dermoscopy images and the corresponding CAMs learned by ARL-CNN50 for melanoma classification and seborrheic keratosis classification, respectively. It reveals that, if the skin lesion is small (see left four images), the CAMs learned for both tasks show similar attention regions, which highlight the skin lesions; otherwise (see right four images), the CAMs learned for both tasks demonstrate different highlights. Therefore, the proposed ARL-CNN model is able to adaptively focus on task-related semantic regions when the skin lesion is large.

### C. Weighting Factor of Attention Learning

As shown in Eq. (9), the output of each ARL block is a weighted sum of the identity map, residual feature map, and attention feature map. The weighting factor $\alpha$, which is also a learnable parameter, represents the trade-off between the attention feature map and other two maps. We displayed the attention feature maps produced by the low (13-layer), middle (31-layer) and high layers (49-layer) of the proposed ARL-CNN50 model in the left column of Fig. 7. It shows that a higher layer has a stronger attention ability than a lower layer. In the right column of Fig. 7, we plotted the variation of the learned value of $\alpha$ versus the epoch during the training process. Generally, the weighting factor $\alpha$ increases during training in all three cases. However, a high layer may have a

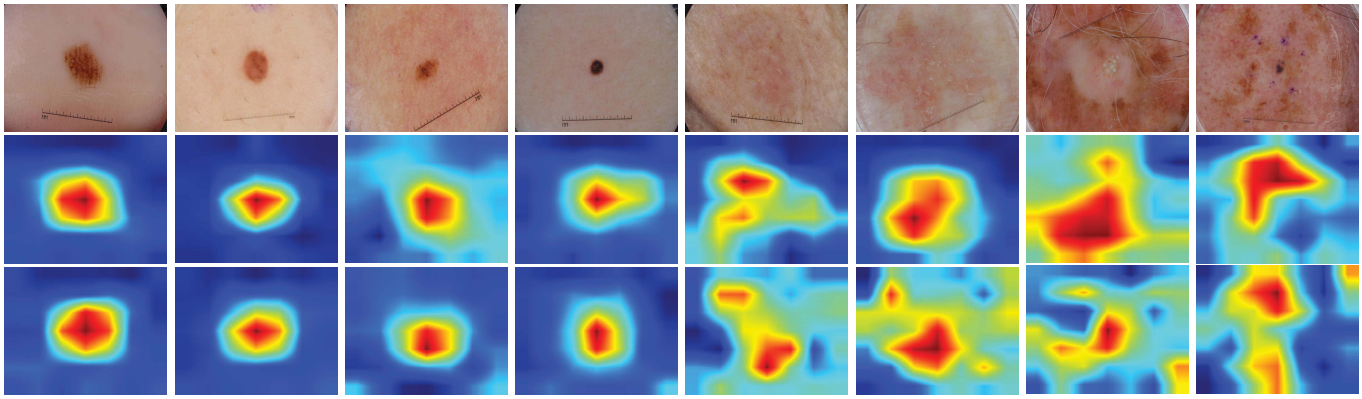

Fig. 6. Eight dermoscopy images (top row) and the corresponding CAMs obtained by ARL-CNN50 when applying it to melanoma classification (middle row) and seborrheic keratosis classification (bottom row), respectively.

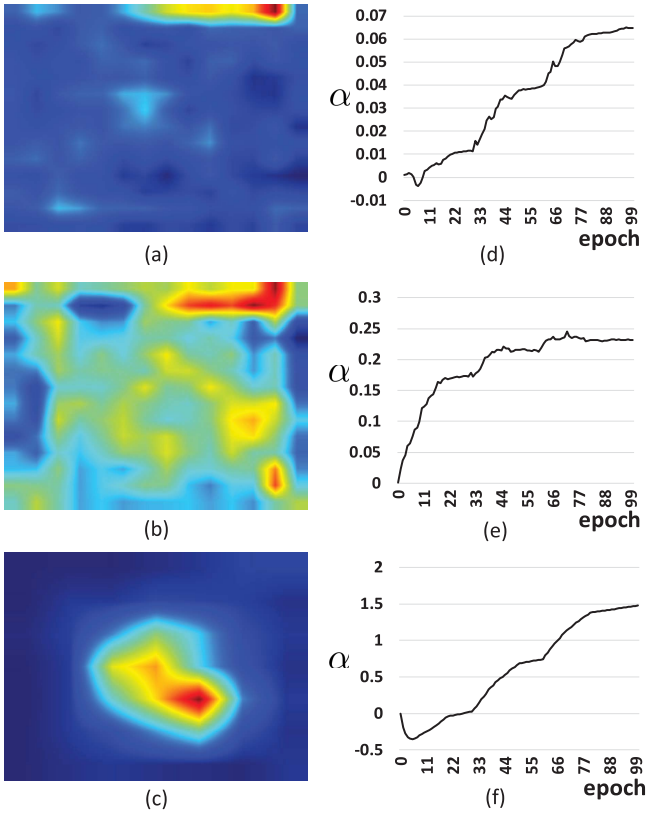


Fig. 7. Visulization of the feature maps in low, middle and high layers and the corresponding curves of weighting factors during the training process: (a) feature map in the $13^{th}$ layer, (b) feature map in the $31^{st}$ layer, (c) feature map in the $49^{th}$ layer, (d) curve of weighting factors in the $13^{th}$ layer, (e) curve of weighting factors in the $31^{st}$ layer, and (f) curve of weighting factors in the $49^{th}$ layer.

large weighting factor $\alpha$, and accordingly the attention feature map learned by a high layer makes more contributions to the classification process.

### D. Effect of Lesion Segmentation

In the above experiments, the lesions were not segmented before applying the images to the proposed ARL-CNN model and directly patches were extracted. To evaluate the effect of lesion segmentation, we also compared the baseline model

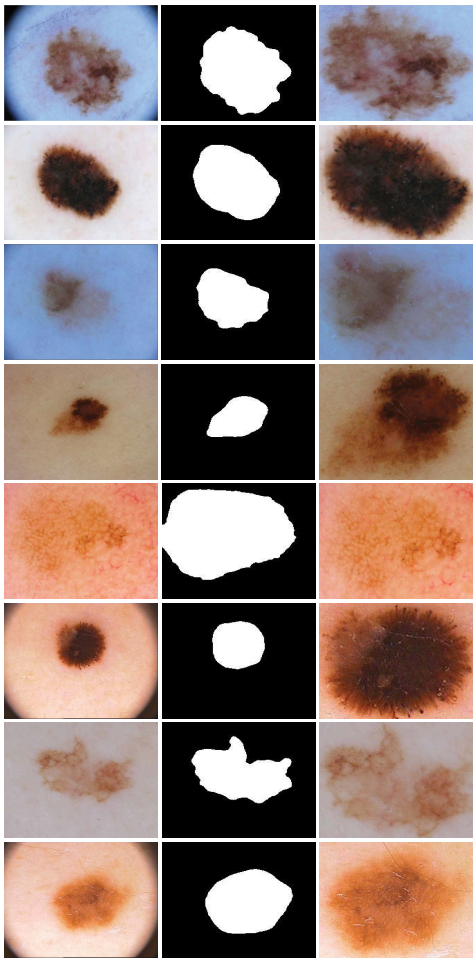

Fig. 8. Eight examples of (left column) dermoscopy images and the corresponding (middle column) segmentation masks and (right column) ROIs cropped according to segmentation masks.

ResNet50 and our proposed ARL-CNN50 model, both of which were trained with the segmented skin lesions - region of interest (ROI). To obtain an accurate segmentation mask of a lesion, we trained a state-of-the-art segmentation model deeplabV3+ [52] by using the ISIC 2017 Skin Lesion Segmentation dataset [5]. Some segmentation examples are shown in Fig. 8. We randomly extracted image patches in the ROI at different scales as introduced in section IV.B. In TABLE VI,

TABLE VI
CLASSIFICATION PERFORMANCE OF RESNET50 AND ARL-CNN50 WITH/WITHOUT SEGMENTATION. (M: MELANOMA CLASSIFICATION, SK: SEBORRHEIC KERATOSIS CLASSIFICATION)

| Methods | Segm | AUC of M | AUC of SK | Average AUC |
|---|---|---|---|---|
| ResNet50 | × | 0.857 | 0.948 | 0.903 |
| ResNet50 | ✓ | 0.864 | 0.955 | 0.910 |
| ARL-CNN50 | × | 0.875 | 0.958 | 0.917 |
| ARL-CNN50 | ✓ | 0.876 | 0.96 | 0.918 |

we compared the AUC obtained by the ResNet50 model and our ARL-CNN50 model with and without segmentation in the melanoma classification and seborrheic keratosis classification. It shows that training the ResNet50 model based on the segmented ROI can effectively resolve the background noise and generated more discriminative features for better classification, resulting in an improvement of the average AUC from 0.903 to 0.910. However, training the proposed ARL-CNN50 model based on the segmented ROI only slightly improved the average AUC from 0.917 to 0.918. The reason can be attributed to the effective attention learning mechanism which enables our ARL-CNN model to focus more on semantically meaningful parts and to achieve the comparable performance without segmentation. Considering that the pixel-wise segmentation has little effect on classification, but a rather high computational cost, we did not segment lesions before applying the images to the proposed model.

### E. Attention Ability to the Artifacts

The classification of skin lesions may suffer from the presence of artifacts, including natural hairs and artificial air bubbles. Some skin lesions may be partly obscured or covered by these artifacts, as shown in Fig. 9. To evaluate the effect of our attention mechanism on these artifacts, we visualized the CAMs obtained by the proposed ARL-CNN50 model for these skin lesions in Fig. 9. It shows that the attention is still paid to the skin lesions, instead of hairs, bubbles, or other artifacts. The robustness of learned attention to these artifacts is significantly important for the accurate lesion classification.

### F. Normalization of Luminance and Color

We exploited the gray-world color constancy algorithm [53] to normalize the luminance and color in dermoscopy images. As shown in Fig. 10, the top row is original dermoscopy images, and the second row is corresponding normalized images. We trained the proposed ARL-CNN50 model with these normalized images and achieved an average AUC of 0.918. It shows that our ARL-CNN model is able to achieve the comparable performance without normalizing the luminance and color, which shows a good generalization ability on diversified images.

### G. Robustness Analysis

We evaluated the proposed ARL-CNN50 model and the baseline ResNet50 model 10 times independently

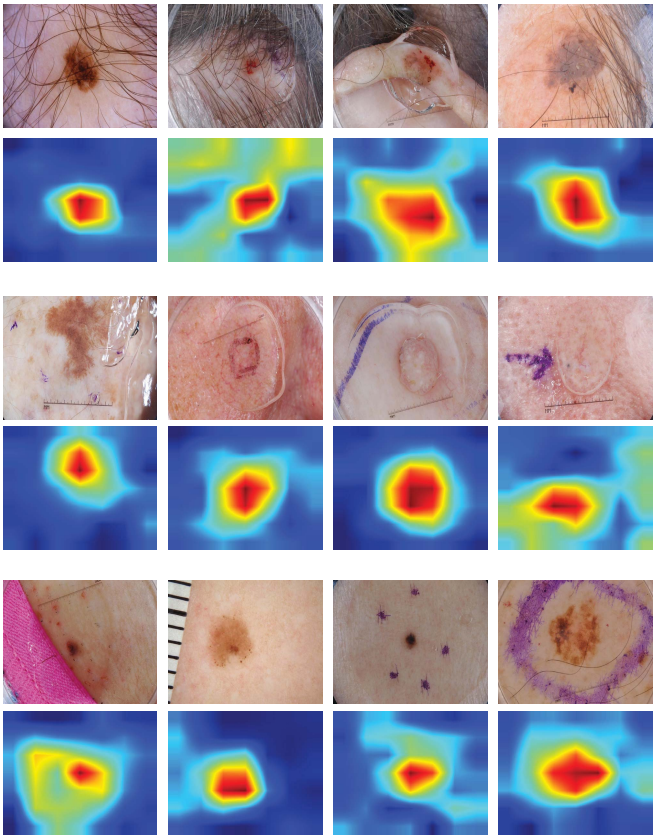

Fig. 9. Visualization of dermoscopy images (above) which contains some artifacts, such as hair and bubbles, and the corresponding CAMs (bellow) obtained by the proposed ARL-CNN50 model.

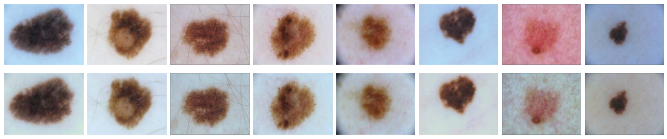

Fig. 10. Some typical examples of dermoscopy images normalized by using color constancy algorithm. Top row: original images; bottom row: normalized images.

($m = n = 10$) and obtained 10 average AUC values for each model, which were listed in Table VII. We assumed that the average AUC values of ResNet50 and ARL-CNN50 are random variables $X$ and $Y$, respectively, each following a Gaussian distribution, i.e. $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and $\sigma_1^2 = \sigma_2^2$. We adopted the independent two-sample $t$-test to determine whether the ARL-CNN50 significantly improves the classification performance. The hypotheses to be tested are $H_0 : \mu_1 \geq \mu_2$ versus $H_1 : \mu_1 < \mu_2$. Given the significance level $\alpha = 0.01$, $t_{0.01}(10 + 10 - 2) = 2.552$, and we have a rejection domain $W = \{t \leq -2.552\}$. According to Table VII, we have

$$\begin{cases} \overline{x} = 90.26, \quad \overline{y} = 91.66 \\ s_1^2 = \dfrac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2 = \dfrac{0.0224}{9} \\ s_2^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2 = \dfrac{0.0064}{9} \end{cases} \tag{14}$$

TABLE VII
AVERAGE AUC VALUES (%) OF ARL-CNN50 AND RESNET50 OBTAINED IN 10 INDEPENDENT TESTS

| Test Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| ResNet50 | 90.3 | 90.4 | 90.3 | 90.1 | 90.3 | 90.5 | 90.2 | 90.4 | 90.1 | 90.0 |
| ARL-CNN50 | 91.7 | 91.7 | 91.6 | 91.8 | 91.7 | 91.5 | 91.6 | 91.6 | 91.7 | 91.7 |

Then,

$$s_w = \sqrt{\frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2}} = \sqrt{\frac{0.0224 + 0.0064}{10+10-2}} = 0.04 \tag{15}$$

and the value of statistic $t_0$ is

$$t_0 = \frac{\overline{x} - \overline{y}}{s_w\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{90.26 - 91.66}{0.04\sqrt{\frac{1}{10} + \frac{1}{10}}} = -78.2624 < -2.5176 \tag{16}$$

Since $t_0$ belongs to the rejection domain $W$, we reject the hypothesis $H_0$. Therefore, comparing to ResNet50, the proposed ARL-CNN50 model improves the classification performance and the improvement is statistically significant.

### H. Computational Complexity

In our experiments, it took about 30 hours to train the proposed ARL-CNN50 model with one NVIDIA GTX Titan XP GPU. The bulk of the time was consumed during the off-line training. However, using the trained model to classify a test image is relatively fast, taking less than 0.2 second (0.02 second per patch) on average. The fast online testing suggests that our approach could be used in a routine clinical workflow.

## VI. CONCLUSION

In this paper, we propose the ARL-CNN model for skin lesion classification in dermoscopy images, which jointly uses the residual learning and a novel attention learning mechanisms to improve the discriminative representation ability of DCNNs. The novel attention learning mechanism is designed to use the feature maps learned by high layers to generate the attention maps for low layers. We evaluated our model on the ISIC-skin 2017 dataset. Our results show that the proposed ARL-CNN model can adaptively focus on the discriminative parts of skin lesions, and thus achieve the state-of-the-art performance in skin lesion classification. Our future work includes the investigation of the unsupervised attention learning and fine-grained skin lesion classification.

## ACKNOWLEDGMENT

We acknowledge the efforts devoted by the International Skin Imaging Collaboration (ISIC) to collect and share the skin lesion classification database for the evaluation of computer-aided approaches to skin lesion classification in dermoscopy images. The first two authors' contribution was made when visiting The University of Adelaide.

## REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30, 2016.
[2] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, "Incidence estimate of nonmelanoma skin cancer (Keratinocyte Carcinomas) in the U.S. population, 2012," *JAMA Dermatol.*, vol. 151, no. 10, pp. 1081–1086, Oct. 2015.
[3] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, Mar. 2002.
[4] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Oct. 2015, pp. 118–126.
[5] N. C. F. Codella *et al.* (2017). "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)." [Online]. Available: https://arxiv.org/abs/1710.05006
[6] J. Weese and C. Lorenz, "Four challenges in medical image analysis from an industrial perspective," *Med. Image Anal.*, vol. 33, pp. 44–49, Oct. 2016.
[7] Y. Song *et al.*, "Large margin local estimate with applications to medical image classification," *IEEE Trans. Med. Imag.*, vol. 34, no. 6, pp. 1362–1377, Jun. 2015.
[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[10] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[11] Z. Zhou, J. Shin, L. Zhang, S. Gurudu, M. Gotway, and J. Liang, "Fine-tuning convolutional neural networks for biomedical image analysis: Actively and incrementally," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4761–4772.
[12] J. Zhang, Y. Xia, Y. Xie, M. Fulham, and D. D. Feng, "Classification of medical images in the biomedical literature by jointly using deep and handcrafted visual features," *IEEE J. Biomed. Health Infom.*, vol. 22, no. 5, pp. 1521–1530, Sep. 2018, doi: 10.1109/JBHI.2017.2775662.2017.
[13] J. Deng, W. Dong, R. Socher, L. J. Li, L. Kai, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[14] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2015, pp. 2048–2057.
[15] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
[16] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 3107–3115.
[17] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6298–6306.
[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.
[19] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6450–6458.
[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.
[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 618–626.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2980–2988.
[25] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3007.
[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[27] 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. 2017.
[28] A. Esteva *et al.*, "Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 546, no. 7660, p. 686, Jun. 2017.
[29] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *Proc. MICCAI*, Sep. 2017, pp. 250–258.
[30] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.
[31] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga. (Mar. 2017). "Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble." [Online]. Available: https://arxiv.org/abs/1703.03108
[32] I. G. Díaz. (Mar. 2017). "Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions." [Online]. Available: https://arxiv.org/abs/1703.01976
[33] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. Avila, and E. Valle. (Mar. 2017). "RECOD titans at ISIC challenge 2017." [Online]. Available: https://arxiv.org/abs/1703.04819
[34] L. Bi, J. Kim, E. Ahn, and D. D. Feng. (Mar. 2017). "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks." [Online]. Available: https://arxiv.org/abs/1703.04197
[35] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su. (Mar. 2017), "A novel multi-task deep learning model for skin lesion segmentation and classification." [Online]. Available: https://arxiv.org/abs/1703.01025
[36] T. DeVries and D. Ramachandram. (Mar. 2017). "Skin lesion classification using deep multi-scale convolutional neural networks." [Online]. Available: https://arxiv.org/abs/1703.01402
[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2010, pp. 807–814.
[38] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2015, pp. 448–456.
[39] J. L. Ba, J. R. Kiros, and G. E. Hinton. (Jul. 2016). "Layer normalization." [Online]. Available: https://arxiv.org/abs/1607.06450
[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[41] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5353–5360.
[42] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2015, pp. 2377–2385.
[43] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.
[44] M. E. Celebi *et al.*, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, Sep. 2007.
[45] C. Barata, M. Ruela, M. Francisco, T. Mendonça, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Sys. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014.
[46] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble mode," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.
[47] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learn. Rep.*, 2017.
[48] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 1106–1114.
[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Rep.*, 2015.
[50] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[51] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2261–2269.
[52] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.
[53] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 3, pp. 1146–1152, May 2015.
[54] Y. Xie *et al.*, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT," *IEEE Trans. Med. Imag.*, to be published, doi: 10.1109/TMI.2018.2876510.2018.
[55] J. Zhang, Y. Xie, Q. Wu, and Y. c, "Skin lesion classification in dermoscopy images using synergic deep learning," in *Proc. MICCAI*, Sep. 2018, pp. 12–20.