Automated skin lesion segmentation using multi-scale feature extraction scheme and dual-attention mechanism

G Jignesh Chowdary
Vellore Institute of Technology
Chennai, India
jigneshchowdary@gmail.com

Ganesh V S N Durga Yathisha
V R Siddhartha Engineering College
Vijayawada, India
gyathisha6807@gmail.com

Suganya G
Vellore Institute of Technology
Chennai, India
suganya.g@vit.ac.in

Premalatha M
Vellore Institute of Technology
Chennai, India
premalatha.m@vit.ac.in

Abstract—Segmenting skin lesions from dermoscopic images is essential for diagnosing skin cancer. But the automatic segmentation of these lesions is complicated due to the poor contrast between the background and the lesion, image artifacts, and unclear lesion boundaries. In this work, we present a deep learning model for the segmentation of skin lesions from dermoscopic images. To deal with the challenges of skin lesion characteristics, we designed a multi-scale feature extraction module for extracting the discriminative features. Further in this work, two attention mechanisms are developed to refine the post-upsampled features and the features extracted by the encoder. This model is evaluated using the ISIC2018 and ISBI2017 datasets. The proposed model outperformed all the existing works and the top-ranked models in two competitions.

Index Terms—Deep learning, Dermoscopic image segmentation, Feature refinement, UNet, ISIC2018, ISBI2017.

I. INTRODUCTION

Skin cancer is one of the leading causes of cancerous death around the world. It was estimated that nearly 13,29,779 new cases of skin cancer were diagnosed in 2018 [44]. Melanoma is the deadliest form of skin cancer, responsible for most skin cancer deaths. In 2018 [44], the incidence of Melanoma was estimated to be 287700 with 60700 death cases. In spite of the high mortality rate of malignant Melanoma, it was observed that early-stage diagnosis could reduce the death rate and enhance the survival rate by five years on average for 95% of the patients. But in later stages, the survival rate is as low as 15%, even with advanced medications and treatment procedures [7]. Since Melanoma involves the pigmentation of lesions on the skin's surface, dermatologists can detect them through visual inspection. But this doesn't guarantee accurate and early-stage diagnosis in all cases. Besides this conventional method, Dermoscopy is the advanced non-invasive procedure for diagnosing Melanoma in the early stages. Dermoscopy eliminates the skin reflections and enhances the visualization ability of the deep skin, thereby enabling dermatologists/ oncologists to diagnose Melanoma which the human eye cannot see. Previous studies have shown that Dermoscopy enhanced the diagnostic accuracy of the conventional procedure [38]. But the manual examination of these dermoscopic images is a non-reproducible and time-consuming process due to the complexity of the lesions and an immense number of images [17]. Manual examinations often contain considerable human error, resulting in a faulty diagnosis [9], [29], [45]. For several cases, it is even difficult for experienced medical professionals to diagnose Melanoma from dermoscopic images due to the varying characteristics of the tumor. There was a great interest in developing Computer-Aided-Diagnostic (CAD) systems for assisting medical professionals in clinical evaluation [1]. One of the important components for developing such CAD systems is the automatic segmentation of lesions from dermoscopic images so that these regions can be used for further analysis [19]. Designing such automated segmentation methods is challenging due to the variations in shape and size, irregular lesion boundaries, and minimum contrast difference between the lesion and the skin. Researchers have worked a lot to resolve these issues. Early segmentation approaches employed edge detection, region growing, and optimum thresholding methods. Saez et-al. [41] used an edge-based-level-set algorithm for the segmentation of the lesion. Grana et al. [22] utilised the Catmull-Rom spline approach to predict the lesion boundary after calculating the lesion slope regularity and slope numerically. Lesions with less contrast difference between the skin can be better diagnosed by combining several threshold-based methods than single-threshold approaches [15]. For finding the optimal colour channel for lesion segmentation, some research works [20], [43] used cluster-based histogram threshold and colour space analysis. These traditional approaches require several hyper-parameters to be fine-tuned for achieving high segmentation performance.

With the rapid developments of the convolutional neural networks (CNN) for medical image segmentation [3], several studies employed them for skin lesion segmentation. Ghafoorian et-al. [21] proposed a multi-branch DCNN for extracting multi-scale context features, but this network is too shallow to extract high discriminative features. With the development of batch-normalization [26], and residual [24], the problem of network degeneration and the vanishing gradient is solved, resulting in the networks going deeper. Yu et al. [49] reported that the deep architectures could extract high discriminative features for skin lesion segmentation, but these networks neglected global features as they are focussed on local context, restraining them from achieving more accurate results with deep architectures.

Recently, attention mechanisms have become popular in deep learning to extract global features for accurate segmentation. In the study [35], the attention mechanism coupled with the popular UNet architecture is used to select the discriminative features for different organs with varying characteristics like size and shape by weighting the different channels. But these single attention approaches fail in cases of lesions with complex characteristics.

In this work, we proposed an UNet based segmentation framework that employs a novel multi-scale feature extraction module for extracting highly discriminative deep features for skin lesion segmentation. In addition to the multi-scale feature extraction module, we also employed a dual attention mechanism for refining the post-upsampled features and the features extracted by the decoder. This work is experimented using the ISBI2017 and ISIC2018 datasets. The experimental results reported that the proposed model achieved better segmentation performance than the existing state-of-the-art works.

II. RELATED WORKS

Over the last decade, several researchers have proposed computerized approaches for the segmentation of skin lesions. These methods can be categorized into active contour models, region growing and splitting, thresholding, clustering, and supervised learning like Celebi et al. [16] employed the ensemble of four thresholding-based methods for the estimation of the lesion boundary. Peruch et al. [37] proposed a novel two-stage approach named Mimiking-Expert-Dermatologists-Segmentation (MEDS) to segment the skin lesions. In the first stage, the Principle Component Analysis (PCA) is employed to project the skin lesion image onto the first principle component of the color-histogram, and in the second stage, the thresholding method is applied to mimic the cognitive procedure of the human dermatologist by clustering the pixels based on color into non-lesion and lesion regions. These thresholding-based approaches mainly depend on the histogram of the image color; thus, these methods suffer when the images contain a significant amount of bubbles, hair, or other unwanted structures. To overcome the effect of such distractions, Zhou et al. [52] employed the classic GVF Snake algorithm and included a mass density function into the optimization objective functional, which may be solved via mean shift estimation.

However, this optimization procedure requires a huge amount of computation for achieving convergence. Sadri et al. [40] employed a fixed-grid wavelet network for segmentation. In this work orthogonal least squares method was employed to supervisely optimize the network topology and for the calculation of the network weights. Xie et al. [48] coupled the generic algorithm with the self-generating neural network for the accurate segmentation of skin lesions. All these supervised and the combination of unsupervised and supervised approaches require particularised domain knowledge as they rely on hand-crafted features.

With the rapid developments in deep learning, deep convolutional neural networks (DCNN) have become more popular in solving computer vision tasks [11], [12]. These DCNN's have the ability to extract highly discriminative hierarchical features from raw images, thereby eliminating the need for hand-crafted features. Besides the success of these models in natural image classification tasks [31], they have also shown promising results in medical imaging like the diagnosis of Mitosis from histopathology images [13], skin cancer classification [18], and image registration [34]. In addition to medical image classification, the DCNN's have shown excellent results in the segmentation of tumors from MRI images [36], left ventricle from cardiac MRI images [4], amongst others. With this motivation, these methods are also used for skin lesion segmentation from non-dermoscopic images [27]. But all these segmentation approaches employed patch-passed classification, in which the input image is divided into several patches, and then each patch is classified/predicted as outside the target or within the target. This technique only integrates limited contextual information included in the patch because each patch only depicts a local region of the image. The contextual information can be enhanced by enlarging the parches, but large patches result in the loss of fine details, which play a major role in final segmentation. The sliding window approach can be employed to integrate both global and local contextual information in segmentation. However, this approach is highly computationally expensive and is not efficient due to overlapping patches.

Due to the limitations in patch-based and sliding window approaches, few researchers employed DCNN's with global dermoscopic images for skin lesion segmentation. Yu et al. [49] developed a neural network with an embedded residual module to enhance the feature extraction capability of the model for efficient skin lesion segmentation. Yuan et al. [51] proposed a new loss function based on Jaccard Similarity for optimizing the segmentation task. Sarkar et al. [42] used the combination of four pre-trained networks in the encoder for extracting discriminative features and proposed a novel loss function based on softmax. These pre-trained networks are extended by pyramid-pooling modules. For a stronger feature representation, Alom et al. [2] presented a novel recursive-residual layer to extract features based on cyclic convolutions. Even though these methods have shown satisfactory performance by enhancing the capability to extract local features, they fail to extract required global features for significantly

higher performance. Recently by combining multiple backbones, Jahanifar et al. [28] proposed an ensemble approach for the segmentation of skin lesions. But the ensemble of multiple models will increase the number of parameters, and such models require more run-time for network convergence. Hence such an approach is complex and difficult to be employed in a clinical scenario. Traditional DCNN's ignore long-range dependencies as they don't have any smart mechanism to guide feature selection. Wang et al. [47] used a non-local block based on an attention mechanism to represent long-range pixel-wise relationships in order to get long-range dependencies. Hu et al. [25] employed a $Squeeze_and_Excitation$ (SE) block for modeling global context information, which re-calibrated channel dependencies by scaling distinct channels. However, these single attention mechanisms are incapable of dealing with the challenges in skin lesion segmentation due to the complex and variant characteristics of lesions. In this work, a new feature extraction module and dual attention mechanism are employed for faster and efficient segmentation of skin lesions.

III. MATERIALS AND METHODS

A. Residual Multi-Scale Module (RMSM)

In this module, multiple convolutional layers with varying kernel sizes are used to extract multi-scale information for each pixel. This multi-scale method enhances the segmentation performance, as the large scale provides more spatial information, whereas smaller scales give more comprehensive information about each pixel's immediate neighbors [6]. Moreover, the residual connection makes learning easier for the network. The structure of the proposed RMSM is shown in Figure 1. In this module, the batch-normalization layer is used after each convolutional layer except for the bottleneck layers to avoid the problem of vanishing gradients while retaining convolutional layers. The ourput of the RMSM can be computed using Equation 1.

$$Out_{RMSM} = I_n \oplus Cov_{1\times1}(L_{BN}(Cov_{1\times1}(I_n)) \\ \oplus L_{BN}(Cov_{3\times3}(I_n)) \oplus L_{BN}(Cov_{r\times5})(I_n))) \quad (1)$$

Where in Equation1, Out_{RMSM} represents the output feature map of the RMSM, I_n represent the input feature map, L_{BN} represent the batch normalization layer, $oplus$ represents the concatenation, and $Cov_{1\times1}$, $Cov_{3\times3}$, $Cov_{5\times5}$ represent convolutional layers with respective kernel sizes.

B. Attention mechanisms

1) Decoder feature refinement attention module (DF-RAM): Feature concatenation between the encoder and decoder at each stage of the network is the topological structure of the UNet. The feature maps from the encoder are combined with upsampled feature maps of the decoder for the better localization of the intended segmentation targets [39]. But, not every visual representation of the encoded features helps for accurate segmentation. Besides this, the semantic gap between
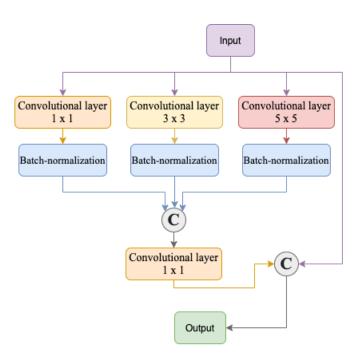

Fig. 1. Multi-scale feature extraction module with residual connection (RMSM).

the decoder and encoder may hinder the segmentation performance. So an attention module (DF-RAM) is employed for the refinement of encoder features before concatenation. The DR-RAM refines features using both channel and spatial attention; the structure of DF-RAM is shown in Figure 4. This module requires a $Decoder\_Feature(DF)$ and $Skip\_Feature(SF)$, where $Decoder\_Feature$ refers to the feature map from the last decoder block (or feature map from the previous encoder block) and $Skip\_Feature$ refers to the feature map from the encoder block at the same level fed through the skip connection. The DF is denoted by $D\epsilon\mathbb{R}^{N_d \times L_d \times K_d}$ and the SF is represented as $S\epsilon\mathbb{R}^{N_s \times L_s \times K_s}$.

In comparison to SF, DF contains semantic information contained in the channel dimension. So in this work, a Max-Pooling operation followed by a Multi-layer perceptron is used for creating the channel-attention map $CH_A(S)\epsilon\mathbb{R}^{N_s \times 1 \times 1}$. The size of the output of the MLP is smaller than the input; this removes the irrelevant information from the channel dimension. The channel attention is presented in Equation2.

$$CH_A(S) = \sigma(MLP(P_{GAP}(D))) \quad (2)$$

In Equation2, σ represents the sigmoid activation, and P_{GAP} represents the Global Average Pooling.

In spatial attention, both DF and SF are used. Firstly a convolutional layer with 1×1 filter is employed to reduce the channel dimensions of SF and DF. Then the squeezed DF is upsampled to complement the size of SF for channel-wise concatenation. These concatenated features are passed through a set of two convolutional layers with variant kernel sizes for producing spatial attention $SP_A(S)\epsilon\mathbb{R}^{1 \times L_s \times K_s}$. The spatial attention is presented in Equation5.
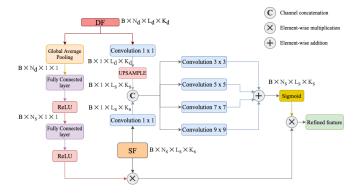
Fig. 2. Proposed Decoder Feature Refinement Attention Module (DF-RAM). It uses SF and EF for refining the encoder features. Here B represents the batch size, N represents the number of channels, L, and K represent the height and width of the feature maps.
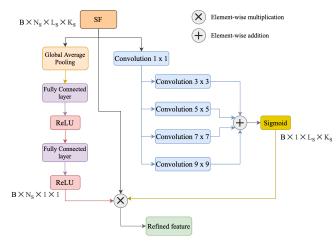

Fig. 3. Proposed Encoder Feature Refinement Module (EF-RAM). It uses SF to refine the post-upsampled feature maps. Here B denotes the batch size, N denotes the number of channels, L, and K denotes the height and width of the feature maps.

$$SP_A(S) = \sigma(Con_{3\times3}([S_r, D_r]) + Con_{5\times5}([S_r, D_r]) + \\ Con_{7\times7}([S_r, D_r]) + Con_{9\times9}([S_r, D_r])) \quad (3)$$

Here $D_r = Up\_sample(Conv^r_{1\times1}(D))$, $S_r = Conv^r_{1\times1}(S)$

In Equation3, r is the reduce ratio and is set to 16 in this work, $Conv_{3\times3}$, $Conv_{5\times5}$, $Conv_{7\times7}$, $Conv_{9\times9}$ represent the convolution operations with respective $kernel\_sizes$, and the $Conv^r_{1\times1}$ is employed to squeeze the channel dimension. Finally, element-wise multiplication is employed for combining spatial and channel attention for generating fused attention $DF(S)$, shown in Equation 4.

$$DF(S) = S \otimes CH_A(S) \otimes SP_A(S) \quad (4)$$

In Equation 4, $\otimes$ represents the *element wise multiplication.*

2) Encoder feature refinement attention module (EF-RAM): For semantic segmentation, the visual representation extracted by the encoder needs to be upsampled for making dense predictions. Interpolation and transposed convolutions are the two approaches for image upsampling, although each has its own set of limitations. Compared to interpolation, transposed convolutions enhance the model capacity as they are trainable and offer non-linearity to the segmentation models. However, improper fine-tuning of hyper-parameters increases the grid effects, and this becomes more complex when more than one transposed convolutions are stacked. Thus, in this work, bi-linear interpolation with the following convolution is employed. As interpolation is non-trainable, noise and other irrelevant information are possible during the upsampling process. To overcome this problem, an attention mechanism is introduced in this work. The attention mechanism refines the feature maps that have been upsampled in both spatial and channel dimensions, as shown in Figure 3. The spatial attention is denoted by $SP_A(S)\epsilon\mathbb{R}^{N\times1\times1}$ and the channel attention is represented by $CH_A(S)\epsilon\mathbb{R}^{N\times1\times1}$.The channel and spatial attentions can be computed using equations 5 and 6.

$$CH_A = \sigma(MLP(P_{GAP}(S)))) \quad (5)$$

$$SP_A(S) = \sigma(Con_{3\times3}(Con^r_{1\times1}(S)) + Con_{5\times5}(Con^r_{1\times1}(S)) + \\ Con_{7\times7}(Con^r_{1\times1}(S)) + Con_{9\times7}(Con^r_{1\times1}(S))) \quad (6)$$

In Equation 5 and 6, r is the reduce ratio and is set to 16, P_{GAP} represents the Global Average Pooling, σ denotes the sigmoid activation, $Con_{3\times3}$, $Con_{5\times5}$, $Con_{7\times7}$, and $Con_{9\times9}$ represents the convolutional operations with respective kernel sizes. And $Con^r_{1\times1}(S)$ is used for squeezing the channel dimension. The fused attention $EF(s)$ is computed

The fused attention $EF(S)$ is computed by combining $SP_A(S)$ and $CH_A(s)$ through element wise multiplication $\otimes$, shown in Equation 7

$$EF(S) = SP_A(S) \otimes CH_A(S) \otimes S \quad (7)$$

C. Residual Multi-Scale dual attention Unet

In this work, a 4 level architecture of Residual Multi-scale Dual Attention UNet is proposed to segment skin lesions from dermoscopic images. This network consists of three paths: encoder, decoder, and bridge. The encoder path extracts the compact representations from the input images, the decoder path recovers the extracted representations to the pixel-wise segmentation, and the bridge connects the encoder and decoder paths. The encoder path consists of three RMSM modules. In each RMSM module, a stride of 2 is applied to the last convolutional layer for downsampling the feature map. The decoder path consists of three RMSM modules, and before each module, there is an upsampling of the feature map from the previous module and concatenation from the RMSM module in the encoder path at the corresponding level. The bridge consists of one RMSM module. In this network, DF-RAM and EF-RAM modules are inserted for enhancing the segmentation performance. The DF-RAM module is used to refine the feature extracted by the encoder and minimize the

segmentation gap and is inserted before the concatenation. And the EF-RAM module is inserted in the decoder path after each RMSM module for refining the upsampled feature maps in both spatial and channel dimensions. At the end of the decoder path, a 1×1 convolutional layer followed by a sigmoid activation layer is used for producing the segmented output.

D. Loss function

Since skin lesions have irregular lesion boundaries and variant shapes, a comprehensive loss function is needed to assist the model during the training process for achieving a rapid and stable convergence. In this regard, dice loss and focal loss are combined to form the loss function $Loss_{seg}$. The loss functions are shown in Equations 8-9.

$$Dice_{loss} = 1 - 2\frac{1 + 2P_{seg}Y_{seg}}{1 + P_{seg} + Y_{seg}} \quad (8)$$

$$Focal_{loss} = -Y_{seg}(1 - P_{seg})^{\gamma}log(P_{seg}) \\ -P_{seg}^{\gamma}(1 - Y_{seg})log(1 - P_{seg}) \quad (9)$$

In equations 8 and 9, the P_{seg} and Y_{seg} represents the predicted segmentation map and the ground-truth. The γ is the focussing parameter and is set as 2 as it has shown satisfactory results in [5], [33], [53]. The combined loss function used in this work is shown in Equation 10.

$$Loss_{seg} = Dice_{loss} + Focal_{loss} \quad (10)$$

E. Training

In this work, two famous publicly available datasets, namely ISBI2017 [14] and ISIC2018 [46], are used for training and validating the model. The skin lesion images had variant sizes in both datasets, so they are normalized into 224 x 224 resolution using the bicubic interpolation algorithm. The ISBI2017 and ISIC2018 consisted of 2000 and 2594 images. As small datasets create the problem of overfitting, data augmentation methods are employed to generalize the model. The employed data augmentation operations include Horizontal and Vertical flip, sharpening, and rotation (randomly from 0 to 90 degrees). By this procedure, four times the samples of the original datasets are generated for training the model. In our experiments, the maximum epoch number is set to 250, and Adam [30] optimizer is chosen with a batch size of 16. The learning-rate is set to 0.0003, and this continues to drop by one-tenth after every 20 epochs.

IV. EXPERIMENTS AND RESULTS

A. Datasets

As mentioned, two publicly available datasets, namely ISBI2017 [14] and ISIC2018 [46] datasets are used in this work. The ISBI2017 dataset consists of a total of 2750 dermoscopic images divided into three sets: training set (2000 images), test set (600 images), and validation set (150 images). And the ISIC2018 dataset contains 2594 images divided into two sets: training set (2000 images) and test set (594 images). With data augmentation methods, the training images

in ISBI2017 and ISIC2018 datasets are increased to 8000 images each.

B. Performance metrics

In this work, five evaluation metrics, including accuracy (ACC), specificity (SPE), recall (REC), dice coefficient (DC), and Jaccard Similarity Index (JSI), are used for validating the proposed work. These metrics are formulated in Equations 11-15.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$SP = \frac{TN}{FP + TN} \quad (12)$$

$$REC = \frac{TP}{FP + TN} \quad (13)$$

$$DC = \frac{2|G \cap Y|}{|G| + |Y|} \quad (14)$$

$$JSI = \frac{|G \cap Y|}{|G \cup Y|} \quad (15)$$

In equations 11-13, the TP and TN represent the number of pixels correctly predicted as lesion and non-lesion regions, FP and FN represent the number of pixels wrongly predicted as lesion and non-lesion regions. And in equations 14 and 15, G and Y are ground-truth and prediction.

C. Comparison with state of the art models

In this work, the performance reported by the proposed model is compared with the state-of-the-art methods, including UNet [39], CE-Net [23], Attention-UNet [35], DeepLab V3+ [10], SLSDeep [42], and R2U-Net [2]. These methods are implemented and applied to the test sets of ISBI2017 and ISIC2018 datasets using the same augmentation methods and computing environment for a fair comparison. For visual comparison, the segmentation results reported by the state-of-the-art methods and the proposed model on complex cases from both the datasets are shown in Figure 5. It was observed that UNet is unable to accurately predict the lesion boundaries of the complex cases. With the use of pyramid pooling for enlarging the receptive fields and the combination of end-point error and log-likelihood in the loss function, the SLSDeep obtained better performance than the UNet. With the usage of recurrent, residual modules, the R2U-Net outperformed the SLSDeep model. The Attention U-Net used attention gates to aid the segmentation network in differentiating the target pixel from the background, and it performed better than earlier research. The CE-Net also reported better segmentation results using a combination of residual-multi-kernel pooling and an atrous convolution module. By employing dual attention mechanisms for enriching the decoder and up-sampled feature maps, the proposed model reported enhanced segmentation results than the existing methods. As shown in the third column of Figure 5, the predicted boundary by the proposed model is
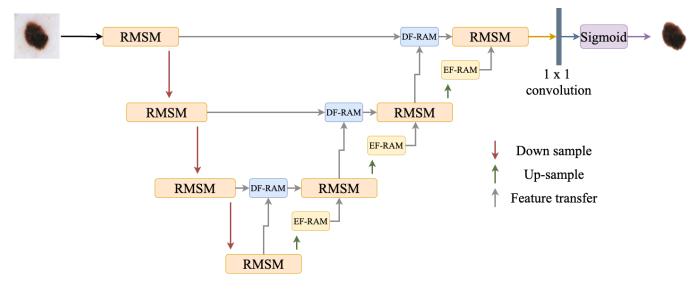
Fig. 4. The proposed segmentation network with Residual Multi scale Feature Extraction module and dual attention mechanism.

Dataset	Method	AC	SP	REC	DC	JSI
ISBI2017	U-Net	88.73	87.84	79.72	78.45	71.58
	CE-Net	95.42	95.38	87.59	89.21	80.61
	Attention-UNet	94.76	94.03	87.16	87.75	78.49
	DeepLab V3+	93.24	94	86.64	87.42	78.05
	R2U-Net	93.12	93.89	86	86.73	77.95
	SLSDeep	92.34	92.17	85.75	86.30	76.82
	Proposed method	**97.5**	**96.94**	**94.29**	**91.16**	**83.83**
ISIC2018	U-Net	86.84	88.57	87.56	85.39	78
	CE-Net	95.75	94.32	94.41	90.26	84.1
	Attention-UNet	94.93	92.59	93.31	89.45	83
	DeepLab V3+	94.56	96.8	90.7	89.67	82.16
	SLSDeep	93.19	91	89.18	89.38	83.67
	R2U-Net	93.79	93.64	91.48	90	83.7
	Proposed method	**95.92**	**97**	**95.37**	**91.52**	**85.41**

Method	DC	JSI	REC
UNet	90.4	77.6	84.92
RMSM Unet	91.29	80	87.62
RMSM UNet + DF-RAM	93.45	81.74	91
RMSM UNet + EF-RAM	92.98	82.53	90.9
RMSM UNet + DF-RAM +EF-RAM (Proposed)	**94.53**	**84.19**	**94.9**

D. Ablation Study

In this work, several ablation studies were conducted to demonstrate the utility of each module used in this work. In this regard, the performance of the proposed method is compared with baseline models including UNet, UNet with Residual Multi-scale Module (RMSM UNet), RMSM UNetwith DF-RAM, and RMSM UNet with EF-RAM. These methods are evaluated on the validation set of ISBI2017. The performance comparison is shown in Table II.

The UNet has several convolutional layers stacked at each level for extracting efficient features and has a skip connection, which is used to transmit information from the corresponding encoder and decoder blocks. This structural schema of the UNet acts as the backbone for the proposed work and does not have any specific mechanism for lesion segmentation, hence resulting in less segmentation JSI and DC. The concept of RMSM is to aggregate feature maps of different convolutional layers with varying kernel sizes, thus making the net wider and capable of learning more discriminative features. The residual connection is added to overcome the problem of network degradation and saturation. By adding the RMSM module, the RMSM UNet is formed, which outperformed the UNet model. The attention mechanisms DF-RAM and EF-RAM are added to refine the feature maps from the decoder and post-upsampled features. From Table II, it is observed

much similar to the ground truth than the previously proposed methods, especially for complex cases with complicated color distribution and irregular boundaries.

Besides the visual comparisons, the performance comparison in terms of AC, SP, REC, DC, and JSI on the ISBI2017 and ISIC2018 datasets are tabulated Table I. It is observed that DeepLabV3+, SLSDeep, and R2U-Net reported superior performance than the UNet in terms of accuracy. The CE-Net and Attention U-Net further enhanced the segmentation accuracy by an average of 1.85% and 0.78% on ISBI2017 and ISIC2018 datasets. The proposed model reported higher performance in terms of AC, SP, REC, DC, and JSI than the existing works on both datasets, indicating the effectiveness of the proposed dual attention multi-scale model for skin lesion segmentation. As UNet is widely used for medical image segmentation, the proposed model achieved an accuracy improvement of 8.77% and 9.08% on the test sets of ISBI2017 and ISIC2018 datasets.
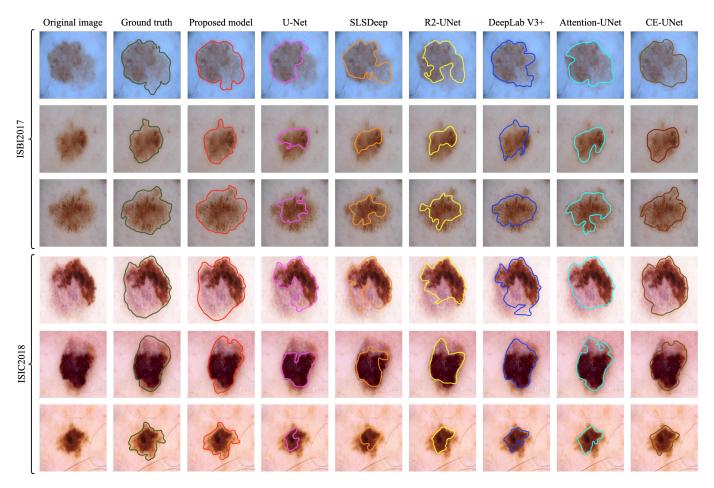
Fig. 5. Comparison with the state-of-the-art methods on complex images from ISBI2017 and ISIC2018 datasets.

that RMSM UNet with EF-RAM reported higher JSI than the RMSM UNet, and RMSM UNet with DF-RAM. In contrast, the RMSM UNet with DF-RAM reported higher DC than the RMSM UNet and RMSM UNet with EF-RAM. And there is a small degradation of recall when using RMSM UNet with EF-RAM than the RMSM UNet with DF-RAM. This shows that the absence of any of the attention mechanisms will hinder the segmentation results. The inclusion of both EF-RAM and DF-RAM mechanisms in RMSM UNet yields the best performance on all the experiments. The performance increase is more than the sum of the performance boosts of each module. These experiments indicate that the introduction of EF-RAM and DF-RAM modules enhances performance mutually.

V. DISCUSSION

From the presented comparative experiments and the ablation studies, it was observed that even for complex cases of skin lesions, the proposed model reported satisfactory results based on the use of multi-scale feature extraction and a dual-attention framework. From Figure 5 and Table I, it can be seen that the proposed model reported superior performance to the attention mechanism-based models. In addition to the comparison with the state-of-the-art works, we also compared

the performance of the proposed model with the competition leaderboard presented in the Table III. The JSI metric is used to rank the methods [8], [32], [50] in the ISBI2017 challenge. Compared with the first ranked method [50], the proposed model reported an enhancement of 8.23% in the JSI metric. For the ISIC2018 competition, the methods were also ranked based on the JSI metric. It is observed that the proposed work reported superior performance in terms of JSI and enhanced the segmentation JSI by 5.21% than the first ranked method (MaskRcnn2+segmentation). On the ISIC2018 leaderboard, it was observed that several methods employed an ensemble-based approach for improvements in segmentation accuracy. These ensemble-based approaches consume more time for training, which makes them difficult to be deployed in a clinical setting. Compared with these approaches, the proposed model requires only 17 seconds to segment each image, which is very less than the existing methods.

The ability of the proposed work in segmenting the most challenging cases is shown in Figure 6. These cases include dermoscopic images with extremely irregular lesion boundaries, and hair in the lesion region, and the state-of-the-art methods failed on these cases, whereas the proposed model successfully segmented the lesion. The multi-scale feature ex-

TABLE III
COMPARISON WITH THE TOP THREE METHODS FROM THE
COMPETETION LEADER BOARD

Competition	Method	JSI	AC	DC
ISBI2017	[50]	75.6	93.4	84.9
	[32]	76.2	93.2	84.7
	[8]	76	93.4	84.4
	Proposed model	**83.83**	**97.5**	**91.16**
ISIC2018	MaskRcnn2+segmentation *	80.2	94.2	89.8
	Ensemble_with_CRF_v3 *	79.9	94.5	90.4
	Automatic Skin Lesion Segmentation by DCNN *	79.9	94.3	90
	Proposed model	**85.41**	**95.92**	**91.52**

Note: The results in this table are obtained directly from the competition leader board. And the methods mentioned with * are not citable.
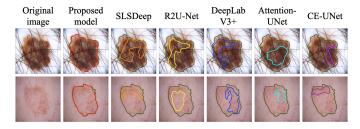

Fig. 6. Comparison with the state-of-the-models on challenging cases. Orange, yellow, cyan, pink, and red contours indicate the predicted segmentation results by the state of the art models and the proposed model, and green indicates the ground truth.

traction with the dual-attention feature refinement enabled the proposed model to deal with these challenges. The proposed model has the potential to be used in real-world scenarios for the segmentation of skin lesions from dermoscopic images. The employed dual-attention mechanism can be used in similar applications where the target region has variant characteristics of color and shape, like skin lesions.

VI. CONCLUSION

In this work, a deep learning approach based on UNet is presented to segment skin lesions from dermoscopic images. Unlike the standard UNet, a new multi-scale feature extraction module is employed for extracting discriminative features to deal the challenge of skin lesion segmentation, which replaces the convolutional layers in the UNet. For enhancing the segmentation performance, a dual-attention mechanism is employed for refining the post-upsampled features and the features extracted by the encoder. This attention mechanism employs both channel and spatial attention for feature refinement. Different from the existing work, this model does not employ any post-processing or pre-processing procedures. This model is evaluated using two publicly available ISIC2018 and ISBI2017 datasets. Experimental results indicate that the proposed model outperformed the existing methods dicussed in the literature.

REFERENCES

[1] Nabila Abraham and Naimul Mefraz Khan, *A novel focal tversky loss function with improved attention u-net for lesion segmentation*, 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 683–687.

[2] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari, *Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation*, (2018).

[3] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan, *Medical image analysis using convolutional neural networks: a review*, Journal of medical systems **42** (2018), no. 11, 1–13.

[4] MR Avendi, Arash Kheradvar, and Hamid Jafarkhani, *A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri*, Medical image analysis **30** (2016), 108–119.

[5] Kanza Azhar, Fiza Murtaza, Muhammad Haroon Yousaf, and Hafiz Adnan Habib, *Computer vision based detection and localization of potholes in asphalt pavement images*, 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2016, pp. 1–5.

[6] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, *Segnet: A deep convolutional encoder-decoder architecture for image segmentation*, IEEE transactions on pattern analysis and machine intelligence **39** (2017), no. 12, 2481–2495.

[7] Charles M Balch, Jeffrey E Gershenwald, Seng-jaw Soong, John F Thompson, Michael B Atkins, David R Byrd, Antonio C Buzaid, Alistair J Cochran, Daniel G Coit, Shouluan Ding, et al., *Final version of 2009 ajcc melanoma staging and classification*, Journal of clinical oncology **27** (2009), no. 36, 6199.

[8] Lei Bi, Jinman Kim, Euijoon Ahn, and Dagan Feng, *Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks*, arXiv preprint arXiv:1703.04197 (2017).

[9] M Emre Celebi, Hitoshi Iyatomi, Gerald Schaefer, and William V Stoecker, *Lesion border detection in dermoscopy images*, Computerized medical imaging and graphics **33** (2009), no. 2, 148–153.

[10] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, *Encoder-decoder with atrous separable convolution for semantic image segmentation*, Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[11] G Jignesh Chowdary et al., *Class dependency based learning using bi-lstm coupled with the transfer learning of vgg16 for the diagnosis of tuberculosis from chest x-rays*, arXiv preprint arXiv:2108.04329 (2021).

[12] _____, *Machine learning and deep learning methods for building intelligent systems in medicine and drug discovery: A comprehensive survey*, arXiv preprint arXiv:2107.14037 (2021).

[13] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, *Mitosis detection in breast cancer histology images with deep neural networks*, International conference on medical image computing and computer-assisted intervention, Springer, 2013, pp. 411–418.

[14] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al., *Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)*, 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, 2018, pp. 168–172.

[15] M Emre Celebi, Hassan A Kingravi, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, Randy H Moss, Joseph M Malters, James M Grichnik, Ashfaq A Marghoob, Harold S Rabinovitz, et al., *Border detection in dermoscopy images using statistical region merging*, Skin Research and Technology **14** (2008), no. 3, 347–353.

[16] M Emre Celebi, Quan Wen, Sae Hwang, Hitoshi Iyatomi, and Gerald Schaefer, *Lesion border detection in dermoscopy images using ensembles of thresholding methods*, Skin Research and Technology **19** (2013), no. 1, e252–e258.

[17] Holly C Engasser and Erin M Warshaw, *Dermatoscopy use by us dermatologists: a cross-sectional survey*, Journal of the American Academy of Dermatology **63** (2010), no. 3, 412–419.

[18] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, *Dermatologist-level classification of skin cancer with deep neural networks*, nature **542** (2017), no. 7639, 115–118.

[19] Harald Ganster, P Pinz, Reinhard Rohrer, Ernst Wildling, Michael Binder, and Harald Kittler, *Automated melanoma recognition*, IEEE transactions on medical imaging **20** (2001), no. 3, 233–239.

[20] Rahil Garnavi, Mohammad Aldeen, M Emre Celebi, George Varigos, and Sue Finch, *Border detection in dermoscopy images using hybrid thresholding on optimized color channels*, Computerized Medical Imaging and Graphics **35** (2011), no. 2, 105–115.

[21] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, IWM Van Uder, Frank-Erik de Leeuw, Elena Marchiori, Bram van Ginneken, and Bram Platel, *Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation*, 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 1414–1417.

[22] Costantino Grana, Giovanni Pellacani, Rita Cucchiara, and Stefania Seidenari, *A new algorithm for border description of polarized light surface microscopic images of pigmented skin lesions*, IEEE Transactions on Medical Imaging **22** (2003), no. 8, 959–964.

[23] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu, *Ce-net: Context encoder network for 2d medical image segmentation*, IEEE transactions on medical imaging **38** (2019), no. 10, 2281–2292.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Deep residual learning for image recognition*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[25] Jie Hu, Li Shen, and Gang Sun, *Squeeze-and-excitation networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[26] Sergey Ioffe and Christian Szegedy, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, International conference on machine learning, PMLR, 2015, pp. 448–456.

[27] M Hossein Jafari, Ebrahim Nasr-Esfahani, Nader Karimi, SM Reza Soroushmehr, Shadrokh Samavi, and Kayvan Najarian, *Extraction of skin lesions from non-dermoscopic images for surgical excision of melanoma*, International journal of computer assisted radiology and surgery **12** (2017), no. 6, 1021–1030.

[28] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, Ali Gooya, and Nasir Rajpoot, *Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations*, arXiv preprint arXiv:1809.10243 (2018).

[29] Guillod Joel, Schmid-Saugeon Philippe, Guggisberg David, Cerottini Jean Philippe, Braun Ralph, Krischer Joakim, Saurat Jean-Hilaire, and Kunt Murat, *Validation of segmentation techniques for digital dermoscopy*, Skin Research and Technology **8** (2002), no. 4, 240–249.

[30] Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, *Imagenet classification with deep convolutional neural networks*, Advances in neural information processing systems **25** (2012), 1097–1105.

[32] Yuexiang Li and Linlin Shen, *Skin lesion analysis towards melanoma detection using deep learning network*, Sensors **18** (2018), no. 2, 556.

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, *Focal loss for dense object detection*, Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[34] Shun Miao, Z Jane Wang, and Rui Liao, *A cnn regression approach for real-time 2d/3d registration*, IEEE transactions on medical imaging **35** (2016), no. 5, 1352–1363.

[35] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., *Attention u-net: Learning where to look for the pancreas*, arXiv preprint arXiv:1804.03999 (2018).

[36] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva, *Brain tumor segmentation using convolutional neural networks in mri images*, IEEE transactions on medical imaging **35** (2016), no. 5, 1240–1251.

[37] Francesco Peruch, Federica Bogo, Michele Bonazza, Vincenzo-Maria Cappelleri, and Enoch Peserico, *Simpler, faster, more accurate melanocytic lesion segmentation through meds*, IEEE Transactions on Biomedical Engineering **61** (2013), no. 2, 557–565.

[38] Christina Ring, Nathan Cox, and Jason B Lee, *Dermatoscopy*, Clinics in Dermatology (2021).

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-net: Convolutional networks for biomedical image segmentation*, International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[40] Amir Reza Sadri, Maryam Zekri, Saeed Sadri, Niloofar Gheissari, Mojgan Mokhtari, and Farzaneh Kolahdouzan, *Segmentation of dermoscopy images using wavelet networks*, IEEE Transactions on Biomedical Engineering **60** (2012), no. 4, 1134–1141.

[41] Aurora Saez, Carmen Serrano, and Begona Acha, *Model-based classification methods of global patterns in dermoscopic images*, IEEE transactions on medical imaging **33** (2014), no. 5, 1137–1147.

[42] Md Mostafa Kamal Sarker, Hatem A Rashwan, Farhan Akram, Syeda Furruka Banu, Adel Saleh, Vivek Kumar Singh, Forhad UH Chowdhury, Saddam Abdulwahab, Santiago Romani, Petia Radeva, et al., *Slsdeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks*, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 21–29.

[43] Philippe Schmid, *Segmentation of digitized dermatoscopic images by two-dimensional color clustering*, IEEE Transactions on Medical Imaging **18** (1999), no. 2, 164–171.

[44] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal, *Cancer statistics, 2019*, CA: a cancer journal for clinicians **69** (2019), no. 1, 7–34.

[45] Nivedita Singh and Shailendra K Gupta, *Recent advancement in the early detection of melanoma using computerized tools: An image analysis perspective*, Skin Research and Technology **25** (2019), no. 2, 129–141.

[46] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler, *The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific data **5** (2018), no. 1, 1–9.

[47] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, *Non-local neural networks*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[48] Fengying Xie and Alan C Bovik, *Automatic segmentation of dermoscopy images using self-generating neural networks seeded by genetic algorithm*, Pattern Recognition **46** (2013), no. 3, 1012–1019.

[49] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng, *Automated melanoma recognition in dermoscopy images via very deep residual networks*, IEEE transactions on medical imaging **36** (2016), no. 4, 994–1004.

[50] Yading Yuan, *Automatic skin lesion segmentation with fully convolutional-deconvolutional networks*, arXiv preprint arXiv:1703.05165 (2017).

[51] Yading Yuan, Ming Chao, and Yeh-Chi Lo, *Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance*, IEEE transactions on medical imaging **36** (2017), no. 9, 1876–1886.

[52] Huiyu Zhou, Xuelong Li, Gerald Schaefer, M Emre Celebi, and Paul Miller, *Mean shift based gradient vector flow for image segmentation*, Computer Vision and Image Understanding **117** (2013), no. 9, 1004–1016.

[53] Hasib Zunair and A Ben Hamza, *Patch efficient convolutional network for multi-organ nuclei segmentation and classification*.