Expert Systems With Applications 183 (2021) 115433

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

SLSNet: Skin lesion segmentation using a lightweight generative adversarial network

Md. Mostafa Kamal Sarker [a,b,*], Hatem A. Rashwan [b], Farhan Akram [c,i], Vivek Kumar Singh [b], Syeda Furruka Banu [b], Forhad U.H. Chowdhury [f], Kabir Ahmed Choudhury [g], Sylvie Chambon [d], Petia Radeva [a,h], Domenec Puig [b], Mohamed Abdel-Nasser [b,e]

[a] *Departament de Matemàtiques i Informàtica, University of Barcelona, 08007 Barcelona, Spain*
[b] *Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*
[c] *Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, 127788 Dubai, United Arab Emirates*
[d] *Institut de Recherche en Informatique de Toulouse, University of Toulouse, 31000 Toulouse, France*
[e] *Department of Electrical Engineering, Aswan University, 81528 Aswan, Egypt*
[f] *Dhaka Medical College Hospital, 1209 Dhaka, Bangladesh*
[g] *Warwick Medical School, University of Warwick, Coventry CV4 7HL, UK*
[h] *Computer Vision Center, University of Barcelona, 08193, Barcelona, Spain*
[i] *Mil-kin Inc., 2F, 2-6-1, Otemachi, Chiyoda-ku, Tokyo, Japan*

ARTICLE INFO

Keywords:
Skin lesion segmentation
Deep generative adversarial network
1-D kernel factorized network
Position attention
Channel attention

ABSTRACT

The determination of precise skin lesion boundaries in dermoscopic images using automated methods faces many challenges, most importantly, the presence of hair, inconspicuous lesion edges and low contrast in dermoscopic images, and variability in the color, texture and shapes of skin lesions. Existing deep learning-based skin lesion segmentation algorithms are expensive in terms of computational time and memory. Consequently, running such segmentation algorithms requires a powerful GPU and high bandwidth memory, which are not available in dermoscopy devices. Thus, this article aims to achieve precise skin lesion segmentation with minimum resources: a lightweight, efficient generative adversarial network (GAN) model called SLSNet, which combines 1-D kernel factorized networks, position and channel attention, and multiscale aggregation mechanisms with a GAN model. The 1-D kernel factorized network reduces the computational cost of 2D filtering. The position and channel attention modules enhance the discriminative ability between the lesion and non-lesion feature representations in spatial and channel dimensions, respectively. A multiscale block is also used to aggregate the coarse-to-fine features of input skin images and reduce the effect of the artifacts. SLSNet is evaluated on two publicly available datasets: ISBI 2017 and the ISIC 2018. Although SLSNet has only 2.35 million parameters, the experimental results demonstrate that it achieves segmentation results on a par with the state-of-the-art skin lesion segmentation methods with an accuracy of 97.61%, and Dice and Jaccard similarity coefficients of 90.63% and 81.98%, respectively. SLSNet can run at more than 110 frames per second (FPS) in a single GTX1080Ti GPU, which is faster than well-known deep learning-based image segmentation models, such as FCN. Therefore, SLSNet can be used for practical dermoscopic applications.

1. Introduction

According to the World Health Organization (WHO),[1] in 2018 there were 1.04 million cases of melanoma in worldwide. Over the last decades, the number of patients affected by melanoma or non-melanoma skin cancers has increased rapidly (Apalla, Nashan, Weller, & Castellsague, 2017). With the growth of artificial intelligence, computer vision, and image analysis techniques, computerized non-invasive

* Corresponding author.
E-mail addresses: mdmostafakamalsarker@ub.edu (M.M.K. Sarker), hatem.abdellatif@urv.cat (H.A. Rashwan), farhan@vim.cau.ac.kr (F. Akram), vivekkumar.singh@urv.cat (V.K. Singh), syeda.samia@hotmail.com (S.F. Banu), drmarufsomc@gmail.com (F.U.H. Chowdhury), K.Choudhury@warwick.ac.uk (K.A. Choudhury), sylvie.chambon@toulouse-inp.fr (S. Chambon), petia.ivanova@ub.edu (P. Radeva), domenec.puig@urv.cat (D. Puig), egnaser@gmail.com (M. Abdel-Nasser).

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

https://doi.org/10.1016/j.eswa.2021.115433
Received 10 July 2020; Received in revised form 14 May 2021; Accepted 10 June 2021
Available online 17 June 2021

dermatology has become essential for the early detection of malignant melanoma to increase the survival rate and reduce the cost of diagnosis/treatment (Esteva, Kuprel, Novoa, Ko, et al., 2017). Computer-aided diagnosis (CAD) systems based on skin lesion delineation methods can help dermatologists to analyze the images captured by digital dermatoscopes. Existing skin lesion delineation methods face such challenges as (1) the vast diversity in the color, shape, texture, size, irregularity and fuzziness of the boundaries of lesions, (2) the presence of blood vessels and hairs, and (3) the low contrast between skin tissues (Al-Masni, Al-antari, Choi, Han, & Kim, 2018).

Several approaches have been presented in the literature - histogram thresholding, clustering, and supervised segmentation methods - to mitigate the challenges mentioned above. A comprehensive survey of traditional segmentation techniques was presented in Celebi et al. (2015). However, these approaches yield inaccurate segmentation results with skin lesions having ambiguous boundaries (Celebi et al., 2015). These traditional segmentation techniques also require different pre-processing algorithms to be applied to improve the inspected images, such as hair removal and contrast enhancement. With the tremendous progress in machine learning, and particularly in deep learning, many skin lesion segmentation approaches have been introduced to improve the accuracy of skin lesion segmentation. For instance, the SLSDeep model was proposed in Sarker, Rashwan, Akram, et al. (2018) to segment skin lesions by using feature pyramid pooling. In Al-Masni et al. (2018), a full resolution convolutional network (FrCN) was proposed to directly learn the full resolution visual content of the input images without pre-processing. And a generative adversarial network (GAN) with an improved loss function, called SegAN, was also introduced in Xue, Xu, and Huang (2018) for learning semantic features of skin lesions in multiscale image representations.

Although existing deep learning-based skin segmentation methods provide acceptable precision, they have hundreds of millions of parameters that make them unsuitable for practical applications. Therefore, using them in clinical settings, especially with dermatoscopy devices with limited computational and memory resources is a challenge. For example, a variety of mobile dermatoscopic devices have been developed for analyzing skin lesions, such as DermLite (3Gen Inc, USA), MoleScope II (MetaOptima Technology Inc, Canada), and HandyScope (FotoFinder Systems, Germany). These devices use a smartphone with a special lens. In this regard, such lightweight image-based segmentation models as Paszke, Chaurasia, Kim, and Culurciello (2016), have been used for skin lesion segmentation but have proved to be less accurate than state-of-the-art lesion segmentation methods. Therefore, there is a need to develop lightweight skin lesion segmentation models that have accuracy rates similar to the state-of-the-art. This article proposes SLSNet, a lightweight GAN model for segmenting melanoma in dermoscopic images. SLSNet extracts low-level skin lesion-relevant features with multiscale convolutional networks and uses a 1-D kernel factorized network (Romera, Alvarez, Bergasa, & Arroyo, 2018) to minimize the computational cost. It also exploits the position and channel attention mechanisms to promote skin lesion feature representations. The main contributions of this work can be summarized as follows:

- We propose SLSNet, which is an efficient, lightweight and fully automated skin lesion segmentation model which has a low computational cost and segmentation that is precise enough to compete with state-of-the-art models.
- A multiscale aggregation mechanism is added to SLSNet to extract relevant features of skin lesions at different scale representations and cope with lesion shape variability. The use of traditional 2D convolution networks increases the number of parameters. Therefore, 1-D kernel factorized networks are exploited instead of the 2D convolution networks to minimize the training parameters.
- We use position and channel attention mechanisms (Fu, Liu, Tian, Fang, & Lu, 2018) to capture the correlation between the spatial and channel features and enhance the ability to discriminate between lesion and non-lesion feature representations.
- We use binary cross-entropy, the Jaccard index, and L_1-loss to formulate a loss function that addresses the challenges accompanied by artifacts existing in dermoscopic images.

The article is organized as follows. Section 2 discusses recent skin lesion segmentation methods based on classical computer vision and deep learning techniques. The model's architecture and the experimental results are explained in Sections 3 and 4, respectively. Finally, Section 5 concludes and suggests some ongoing and future lines of this research.

2. Related work

Several fully-automated/semi-automated skin lesion segmentation approaches have been presented throughout the last decade based on classical computer vision, machine learning, and deep learning techniques. Below, we discuss the most common skin lesion segmentation methods and summarize them in Table 1.

2.1. *Classical computer vision-based approaches*

In the context of skin lesion image analysis, numerous computer vision methods have been used: for example, image thresholding (Mahmoud, Abdel-Nasser, & Omer, 2018), active contour (Silveira et al., 2009), region growing (Rahman, Alpaslan, & Bhattacharya, 2016)), and unsupervised learning, e.g., clustering (Agarwal, Issac, Dutta, Riha, & Uher, 2017). For instance, adaptive thresholding and region growing methods were used in Rahman et al. (2016) to segment skin lesions, and then feed the segmented regions into a support vector machine algorithm to detect the type. These approaches yield acceptable results when the skin lesions' boundaries are clearer and more distinctive than the boundaries of the background objects (regions of interest (ROIs) are considered foreground objects). However, boundaries are often not distinctive and clear because of the presence of hair in the dermoscopic image and its low contrast. Contour-based methods, such as adaptive snake and active contours (Silveira et al., 2009) are examples of methods that cannot discriminate between skin lesions and healthy skin when the boundaries of lesions are unclear. The precision of such methods also degrades with the change in pigment or the presence of hair. Similarly, clustering-based methods are also not efficient with complex dermoscopic images. In general, hand-crafted methods, such as active contours and clustering require manual tuning for many parameters to achieve acceptable accuracy. However, efficient hand-crafted methods could be deployed on devices with limited resources.

2.2. *Deep learning-based approaches*

Deep learning techniques, especially convolutional neural networks (CNNs), have been applied to such computer vision tasks as image segmentation (Guo, Liu, Georgiou, & Lew, 2018), object detection (Zhao, Zheng, Xu, & Wu, 2019) and image classification (Rawat & Wang, 2017). The fully convolutional network (FCN) (Long, Shelhamer, & Darrell, 2015) made the initial breakthrough as a deep learning approach in image segmentation using the encoder and decoder framework. Lateef and Ruichek (2019) review variations of encoder–decoder networks for various image segmentation tasks. For instance, Bi et al. (2019) proposed a skin lesion segmentation model based on the FCN architecture that achieved an intersection-over-union (IoU) score of 77.73% with the ISBI 2017 dataset. Al-Masni et al. (2018) introduced a full-resolution convolutional network (FrCN) for skin lesion segmentation and achieved an IoU score of 77.11% with the ISBI 2017 dataset. Yu, Chen, Dou, Qin, and Heng (2017) proposed a fully convolutional residual network (FCRN) that includes many deep layers and uses multiscale contextual features to segment skin lesions. This network yields detailed and accurate segmentation results but loses accuracy when the melanoma is highly heterogeneous (Adegun & Viriri, 2019). Ronneberger, Fischer, and Brox (2015) proposed the U-Net

Table 1
Summary of skin lesion segmentation methods. The unreported information is indicated with dashes (–) in the referred literature.

Reference	Dataset	Data availability	Methods/Architectures	Pre-processing/Data augmentation	Post-processing
Classical computer vision					
Silveira et al. (2009)	Hospital Pedro Hispano (HPH)	Private	Active contour	Filtering and smoothing	Morphological operation
Rahman et al. (2016)	ISBI 2016	Public	Region growing	Color conversion and binarization	–
Agarwal et al. (2017)	DermIS and DermQuest	Public	K-means clustering	Color conversion and filtering	Thresholding
Deep learning (Encoder–Decoder)					
Yu et al. (2017)	ISBI 2016	Public	FCRN	Rotating, adding noise and flipping	Thresholding
Li, Yu, Chen, Fu, and Heng (2018)	ISBI 2017	Public	U-Net	Flipping, rotating and scaling	–
Bissoto et al. (2018)	ISIC 2018	Public	U-Net	Flipping, scaling, rotating and illuminating	Thresholding and hole filling
Vesal, Ravikumar, and Maier (2018)	ISBI 2017	Public	SkinNet	Rotation, flipping, color shifting and scaling	–
Sarker et al. (2018)	ISBI 2016 and 2017	Public	SLSDeep	–	–
Bi et al. (2019)	ISBI 2016, 2017 and PH2	Public	FCN	Cropping and flipping	Thresholding and hole filling
Al-Masni et al. (2018)	ISBI 2017 and PH2	Public	FrCN	Rotating and color conversion	–
Mishra and Daescu (2019)	ISIC 2018	Public	Mask-RCNN	Flipping. rotating, scaling and blurring	Superpixel clustering
Unver and Ayan (2019)	ISBI 2017 and PH2	Public	YOLO and GrabCut	Hair removal	Morphological operation
Ma, Wu, Sun, Yu, and Liu (2019)	ISBI 2017	Public	LCASA-Net	Color distortion, flipping and cropping	–
Hartanto and Wibowo (2020)	ISBI 2017	Public	MobileNet v2 and Faster R-CNN	Region of interest selection	–
Deep learning (GAN)					
Xue et al. (2018)	ISBI 2017	Public	SegAN	Random cropping and flipping	–
Bisla, Choromanska, Stein, Polsky, and Berman (2019)	ISBI 2017 and ISIC 2018	Public	U-Net and GAN	Random masking	Hole filling
Izadi, Mirikharaji, Kawahara, and Hamarneh (2018)	DermoFit	Public	GAN with UNet-Critic	Random cropping and flipping	–
Lei et al. (2020)	ISBI 2016, 2017 and ISIC 2018	Public	DAGAN	Flipping and rotating	–

model for biomedical image segmentation using very little data. They introduced a concept called *skip connection*, which extracts features from each encoder layer and concatenates them with the corresponding decoder layer. Skip connection helps suppress the singularities inherent in the loss of deep CNNs. U-Net performs well on a variety of biomedical image segmentation tasks so several U-Net-based models have been proposed for skin lesion segmentation.

For instance, a self-ensemble U-Net model has been proposed in Li et al. (2018) using rotation and flipping transformations as well as a consistent scheme to improve the effect of regularization for pixel-level predictions. With the ISBI 2017 test dataset, (Li et al., 2018) achieved an IoU score of 79.87%. Bissoto et al. (2018) used pre-processing techniques to eliminate the noise from dermoscopic images, and then they fed the filtered images to a U-Net model to segment skin lesions, achieving an IoU score of 72.8% with the ISIC 2018 dataset. Vesal et al. (2018) used densely connected convolution layers for skin lesion segmentation (so-called SkinNet), which yielded an IoU score of 76.7% with the ISBI 2017 test dataset.

In the literature, object detection models such as YOLO-v3 and Mask-RCNN based on ResNet101 or ResNet50 have about 60 and 40 million parameters, respectively. These models have been used in skin lesion segmentation systems to detect the lesion region or provide initial segmentation results. Mishra and Daescu (2019) proposed a two-step method for segmenting skin lesion images acquired by both dermoscopic and cellphone devices with a special lens. In the first step, the Mask-RCNN based on ResNet152 is used to obtain an initial segmentation. In the second step, the initial segmentation is fed into a superpixel segmentation method, which yields accurate skin lesion segmentation. Since Mishra and Daescu (2019) exploit the ResNet152 backbone that has 60.2 million parameters, their model is not suitable for devices with limited resources. In addition, Sarker et al. (2018) proposed an end-to-end deep learning model, SLSDeep, for skin lesion segmentation, in which dilated residual convolution layers with a pyramid pooling network have been used to extract contextual features from multiscale representations. To improve the boundaries of the segmented lesions, they introduced a combination of negative log-likelihood and endpoint error functions as a loss function. SLSDeep obtained an IoU score of 78.2% with the ISBI 2017 dataset.

In turn, MobileNet, which is based on depth-wise separable convolutions, has been tested with various applications, including object detection, fine-grain classification, face attributes, and large scale geo-localization. MobileNet v1 (Howard et al., 2017) has around 4.24 million parameters, while MobileNet v2 (Sandler, Howard, Zhu, Zhmoginov, & Chen, 2018) is more efficient and powerful than MobileNet v1 with only 3.47 million parameters. Sae-Lim, Wettayaprasit, and Aiyarak (2019) used MobileNet for skin image classification. It should be noted that object detectors based on MobileNet v2 backbone, such as Faster R-CNN (Hartanto & Wibowo, 2020), reduce the number of parameters. However, the segmentation accuracy is lower. Furthermore, YOLO and the GrabCut algorithm have been combined in Unver and Ayan (2019) for skin lesion segmentation to yield a Dice score of 84.26% and a Jaccard score of 74.81% with the ISBI 2017 dataset. Ma et al. (2019) proposed a lightweight context-aware self-attention model called LCASA-Net for skin lesion segmentation. LCASA-Net has 0.49 million parameters and achieves a Dice score of 87.90% and Jaccard score of 80.90% with the ISBI 2017 dataset. Ma et al. (2019) used a deep neural network architecture (i.e., ENet) for real-time skin lesion segmentation, and achieved a Dice score of 82.70%and a Jaccard score of 74.10% with the ISBI 2017 dataset.

Other studies have used pruning, quantization, coding and knowledge distillation techniques to reduce the complexity of state-of-the-art deep learning models. For instance, Han, Mao, and Dally (2016) introduced a deep compression method to compress deep neural networks with pruning, trained quantization, and Huffman coding. They succeeded in reducing the storage of AlexNet by 35× (from 240 MB to 6.9 MB), VGG16 by 49× (from 552 MB to 11.3 MB) without any noticeable loss in classification accuracy with the ImageNet dataset. Two compressing techniques were also introduced in Polino, Pascanu, and Alistarh (2018) by using distillation and quantization methods. In Zhou et al. (2016), parameter gradients were randomly quantized to low bitwidth numbers before they were fed into convolutional layers. With AlexNet, this approach achieved 46% top-1 accuracy with the ImageNet dataset. In Tung and Mori (2019), a knowledge distillation approach was proposed for training a neural network by (1) following the supervision of a trained teacher network and (2) improving the accuracy of the model through similarity preserving loss function. Rastegari, Ordonez, Redmon, and Farhadi (2016) proposed two efficient networks, namely binary-weight-networks and XNOR-networks, to reduce the memory cost of convolutional operations and network size by 32×, so that these networks could be executed on CPU in real-time. Bethge, Bartz, Yang, Chen, and Meinel (2020) introduced the MeliusNet model, which uses binary weights and activations instead of the standard 32 bit floating-point values. MeliusNet achieved results comparable to MobileNet-v1 in terms of accuracy, model size, and the number of operations. It should be noted that such compressing techniques can degrade the performance of deep learning models since some constraints yield biased models, and a proper structural constraint can be difficult to find (Cheng, Wang, Zhou, & Zhang, 2017).

2.3. GAN-based approaches

Several methods based on GANs have been proposed for skin lesion segmentation. For instance, Xue et al. (2018) proposed a GAN-based model that depends on ResNet blocks with skip connections. With the ISBI 2017 test dataset, it achieved an IoU score of 78.50%. Bisla et al. (2019) proposed a two-stream deep convolutional generative adversarial network with the ResNet50 model to jointly segment and classify skin lesions. They used several pre-processing techniques to remove the artifacts from the skin images. Bisla et al. (2019) presented an end-to-end deep learning system for lesion segmentation and classification based on networks specialized in data purification and augmentation. The system proposed in Bisla et al. (2019) achieved IoU scores of 77.00% and 70.20%, respectively, with ISBI 2017 and ISIC 2018 datasets. Izadi et al. (2018) introduced a GAN-based UNet-Critic model for skin lesion segmentation, achieving IoU scores of 81.20% on the DermoFit dataset. Recently, a dual adversarial GAN (Lei et al., 2020) with dual discriminators achieved IoU scores of 87.10%, 77.10% and 82.40% on ISBI 2016, 2017 and ISIC 2018 datasets, respectively.

2.4. Attention mechanism-based approaches

Several studies in the literature have used attention mechanisms to enhance the performance of deep CNNs (Chen, Yang, Wang, Xu, & Yuille, 2016; Mnih, Heess, Graves, et al., 2014; Wang et al., 2017). Attention mechanisms help deep CNNs to pay more attention to the features with more enriched information. In particular, the attention mechanism can guide deep learning models to neglect irrelevant information and focus on more discriminant regions of the image by emphasizing relevant feature associations in both channels and spatial spaces. Besides, the attention module helps deep CNNs to efficiently integrate local and global features. In Chen et al. (2016), an attention mechanism for softly weighing the multiscale features was proposed to capture the key features at different scales and positions. Schlemper et al. (2018) proposed a generalized self-gated soft-attention mechanism to enable the convolutional layers to contextualize local features. This mechanism can be combined with existing deep learning segmentation or classification models while adding few trainable parameters. Oktay et al. (2018) proposed an attention gate model that determines lesion structures of different shapes and sizes. It promotes the task-relevant salient features and ignores irrelevant regions in the input images. Besides, Mnih et al. (2014) presented a visual attention mechanism, which finds the salient regions and processes them at high resolution.

Most skin lesion segmentation methods, as mentioned above, use deep learning models with a massive number of parameters, so they cannot be used for real-time applications running on low-resources devices. To address this issue, we propose SLSNet, which is a lightweight and efficient GAN-based model. In general, the number of parameters of any segmentation model depends on several factors, and in particular the number of convolutional layers used, the input layer's image size, and the fully connected layers. Unlike other models, SLSNet includes a novel layer, *factorized-attention module* (FAM), which combines two branches: residual 1-D factorized kernel convolution (factorized layer) and the channel attention module. FAM computes the convolutions in such a way that it reduces the overall number of computational parameters. We also use a multiscale aggregation mechanism to extract relevant skin lesion features at different scales to segment skin lesions of various shapes and sizes. Both the position attention module (PAM) and channel attention module (CAM) are used to encode contextual information into local features. These modules help SLSNet to accurately distinguish skin lesions from healthy skin by using localized texture information. In other words, CAM and PAM can facilitate the model's training process since they encourage the model to learn skin-lesion-relevant features and do not increase the number of parameters of the training model.

3. Methodology

In this section, we describe the network architecture and its layers in detail. The baseline network is the conditional GAN (cGAN) that mitigates the collapse mode problem. The pix2pix cGAN model proposed in Isola, Zhu, Zhou, and Efros (2017) has been used to solve several medical image segmentation and classification problems. In short, cGAN includes two main networks: the generator G and discriminator D. G contains encoder and decoder networks (i.e., auto-encoder architecture). On the one hand, G can be trained to learn a mapping function from domain A to domain B, where domain A is the dermoscopic images and domain B is the segmented skin lesions. In the G network, the encoder comprises 19 convolutional layers while the decoder comprises 7 deconvolutional layers. On the other hand, D can be used to compare the resulting segmentation masks with real segmented images (i.e., ground truth). Fig. 1 depicts the architecture of SLSNet, which has G and D networks as pix2pix. A multiscale block is used in SLSNet to aggregate the coarse-to-fine features of dermoscopic images and alleviate false detections caused by artifacts. SLSNet is explained in detail in the subsections below.

3.1. Encoder network

The multiscale block is shown in Fig. 1(left). It firstly generates three scaled images from each input image with three ratios: 1/8, 1/4, and 1/2 of the original image size. This mechanism helps SLSNet to be invariant to image resolution by dealing with objects and images of different scales. Then, four 3×3 convolutional filters followed by four CAMs are used to capture visual feature dependencies on channel dimensions. Note that the multiscale mechanism generates scale-invariant filters that help segment tiny skin lesions properly. The sizes of the four resulting features maps are $128 \times 128 \times 16$, $64 \times 64 \times 16$, $32 \times 32 \times 16$, and $16 \times 16 \times 16$, from bottom to top, respectively. Afterwards, the three lower-scale features (i.e., $64 \times 64 \times 16$, $32 \times 32 \times 16$, and $16 \times 16 \times 16$) are upsampled to the same size of the original input image by using the bilinear interpolation method. Finally, the feature maps are averaged to produce the output of the multiscale block ($128 \times 128 \times 16$ feature map). It should be noted that the multiscale

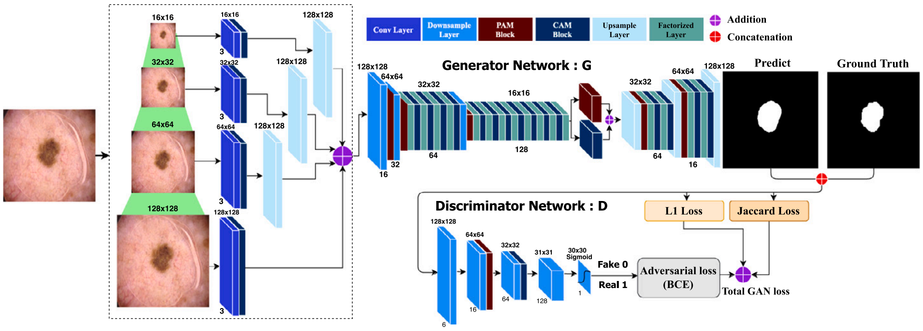


Fig. 1. SLSNet architecture. (top) the generator network and (bottom) the discriminator network. Conv, PAM, CAM, BCE and GAN stand for Convolutional, Position Attention Module, Channel Attention Module, Binary Cross Entropy and Generative Adversarial Network, respectively.

block helps the encoder to extract low-level features in different scales to cope with shadows, and that the resulting feature maps are created in both spatial and channel domains.

As shown in Fig. 1, the sixteen feature maps resulting from the multiscale block are entered into two downsampling-attention layers, each of which includes a convolutional layer followed by a (2,2) max-pooling and a position attention module (PAM) to help capture the spatial features. The resulting feature maps are fed into four factorized-attention modules (FAMs), where each FAM comprises a 1-D kernel factorized layer followed by a CAM. The output of the four FAMs is fed into a downsampling layer and a PAM. The feature maps are passed to eight FAMs. The output of the eighth FAM is entered into a 1-D kernel factorized layer followed by two branches: one branch includes CAM while the other includes PAM. The outputs of the two branches are summed to obtain visual features independent of the position and channel dimensions.

3.1.1. Channel attention module (CAM)

The feature maps consist of a set of channels, each one of which can be considered to be a class-specific response representing high-level features. However, many semantic responses (i.e., channels) are correlated with each other. Consequently, CAM, which exploits inter-dependencies among channel maps, can highlight inter-dependent feature maps and update the feature representation of specific interpretation. Thus, a CAM is used here to explicitly model inter-dependencies among channels. Fig. 2 shows the architecture of CAM. The channel attention map $\mathbf{X} \in \mathbb{R}^{C\times C}$ is calculated directly from the original features $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$, where C, H and W are the channels, height, and width of the input image, respectively. $\mathbf{A}$ is reshaped to $\mathbb{R}^{C\times N}$, where $N = H \times W$ is the number of features. Then, matrix multiplication between $\mathbf{A}$ and the transpose of $\mathbf{A}$ is performed. In order to generate the channel attention map $\mathbf{X} \in \mathbb{R}^{C\times C}$, a softmax function is applied as follows:

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)}, \tag{1}$$

Here, x_{ji} determines the ith channel impact on the jth channel. Matrix multiplication is performed between the transpose of $\mathbf{X}$ and $\mathbf{A}$. The result of the multiplication is reshaped to $\mathbb{R}^{C\times H\times W}$, multiplied by a scale parameter γ and then an element-wise addition is performed with $\mathbf{A}$ to produce the final output $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$:

$$E_j = \gamma \sum_{i=1}^{C} (x_{ji} A_i) + A_j, \tag{2}$$

Here, the final feature of each channel is the weighted sum of the features of all channels and original features. It can model the long-range semantic dependencies between feature maps, as shown in (2). Hence, CAM can highlight class-dependent feature maps and boost discriminative features that cannot be produced by the convolution layers.

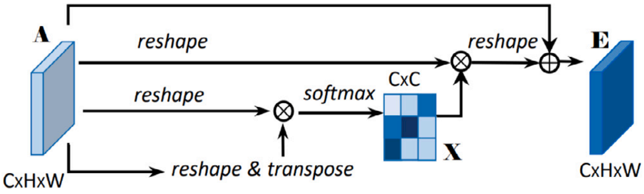


Fig. 2. The architecture of the channel attention module (CAM).

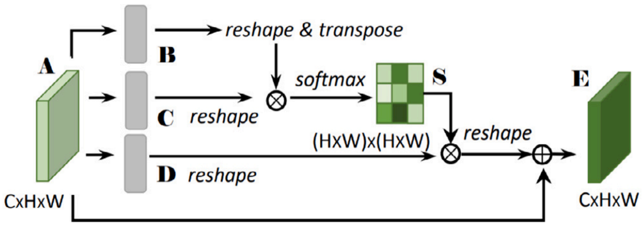


Fig. 3. The architecture of the position attention module (PAM).

3.1.2. Position attention module (PAM)

Discriminant feature representations are essential for accurate skin lesion segmentation, which can be achieved by capturing long-range contextual information. The position attention module (PAM) can encode a comprehensive series of contextual information into local features. Note that the spatial context is refined by aggregating the spatial features. Therefore, we use PAMs in SLSNet to model strong contextual links over local feature descriptions. As can be seen in Fig. 3, the local feature maps $\mathbf{A} \in \mathbb{R}^{C\times H\times W}$ are fed into a convolution layers with a batch normalization and ReLU to produce two feature maps: $\mathbf{B}$ and $\mathbf{C}$, where $\{\mathbf{B}, \mathbf{C}\} \in \mathbb{R}^{C\times H\times W}$. Then, $\mathbf{B}$ and $\mathbf{C}$ are reshaped to $\mathbb{R}^{C\times N}$, where $N = H \times W$. The transpose of $\mathbf{C}$ and $\mathbf{B}$ are multiplied and the resulting features are then fed into a softmax function S to estimate the spatial attention map $\mathbf{S} \in \mathbb{R}^{N\times N}$:

$$s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)}, \tag{3}$$

Here, s_{ji} refers to the ith position's contact on jth position. The softmax function S attempts to find the correlation between two spatial positions in the input feature maps. As shown in Fig. 3, the feature maps $\mathbf{A}$ are fed into a convolutional layer with batch normalization and ReLU to produce a new feature map $\mathbf{D} \in \mathbb{R}^{C\times H\times W}$. Then, $\mathbf{D}$ is reshaped to $\mathbb{R}^{C\times N}$. The transposes of $\mathbf{S}$ and $\mathbf{D}$ are multiplied, and then the output of PAM, $\mathbf{E} \in \mathbb{R}^{C\times H\times W}$, is computed as follows:

$$E_j = \eta \sum_{i=1}^{N} (s_{ji} D_i) + A_j, \tag{4}$$

Here, η is a scale parameter (Zhang, Goodfellow, Metaxas, & Odena, 2018). The output of PAM, $\mathbf{E}$, at each position is a weighted sum of

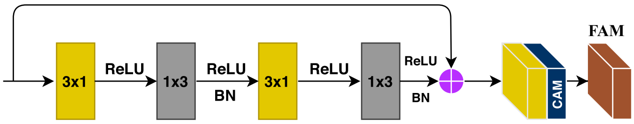


Fig. 4. The architecture of the Factorized-Attention Module (FAM). ReLU, BN and CAM stand for Rectified Linear Unit, Batch Normalization and Channel Attention Module respectively.

the features of all neighbors of original features. This analysis demonstrates that PAM can produce a global contextual representation and selectively aggregate the context according to the spatial attention map generated.

3.1.3. Factorized-attention module (FAM)

Fig. 4 depicts the architecture of the factorized-attention module (FAM) generatedused in this study to reduce the computation complexity. FAM comprises 1-D kernel factorized layers, residual connection and CAM. Assume that $\mathbf{W} \in \mathbb{R}^{C\times d^h \times d^v \times F}$ are the weights of a 2D convolutional layer, C is the number of input planes, F is the number of output planes (feature maps) and $d^h \times d^v$ is the kernel size of each feature map (typically $d^h \equiv d^v \equiv d$). The rank-1 constraint, $\mathbf{f^i}$, can be rewritten as a linear combination of 1D filters:

$$\mathbf{f^i} = \sum_{k=1}^{K} \sigma_k^i \bar{v}_k^i \left(\bar{h}_k^i\right)^T, \tag{5}$$

Here, $b \in \mathbb{R}^F$ represents the bias term for each filter and $\mathbf{f^i} \in \mathbb{R}^{d^h \times d^v}$ indicates the ith kernel of a layer. σ_k^i is a scalar weight and K is a rank of $\mathbf{f^i}$. The length of $\bar{v}_k^i$ and $\left(\bar{h}_k^i\right)^T$ is d. The ith output of the decomposed layer a_i^1 can be expressed as a function of its input a_*^0 as follows:

$$a_i^1 = \varphi\left(b_i^h + \sum_{l=1}^{L} \bar{h}_{il}^T * \left[\varphi\left(b_l^v + \sum_{c=1}^{C} \bar{v}_{lc} * a_c^0\right)\right]\right), \tag{6}$$

Here, $\varphi(.)$ can be represented by the non-linearity of the 1D decomposed filters, which can be implemented with ReLU. As shown in Fig. 4, the 1-D kernel factorized layer's output feature map is fed into a CAM to generate the final representation of FAM.

3.2. Decoder network

As shown in Fig. 1 (the layers after the two branches), the decoder of SLSNet contains two consecutive groups of layers, in which each group comprises an upsampling layer, PAM, and two FAMs. These layers are followed by an upsampling layer to predict the segmentation masks with a size of 128 × 128. A threshold of 0.5 is applied to the predicted segmentation masks to produce the binary masks. It is worth noting that we use convolutional and deconvolutional filters with a kernel size of 3 × 3, and a stride of 2 and a padding of 1 in all layers of both encoder and decoder networks.

3.3. Discriminator network

Fig. 1(bottom) shows that D comprises four layers. The first three are convolutional layers with a kernel size of 4 × 4, a stride of 2, and a padding of 1. A PAM block is added in the second downsampling layer, while a CAM block is added in the third layer. In the final layer of the discriminator network, a sigmoid activation function is applied.

3.4. Model training

During the training of SLSNet, we followed the updated schema of the pix2pix model, in which adversarial back-propagation is used for alternately training the G and D networks. Firstly, by using the gradients computed from the loss function while fixing G, we train the D network once. In particular, the discriminator model is updated for a half batch of real samples, and then for a half batch of fake samples (the two halves together form one batch of weight updates). Next, we fix the D network and train the G network using the gradients computed from the same loss function passed from D to G. Assume that x is an input skin lesion image and the ground-truth of the segmented image is y. Let z be a random variable that can be introduced as a dropout in the layers of the decoder, which helps to avoid overfitting of the model and generalize the learning process. Thus, the outputs of the generator and the discriminator can be expressed as $G(x,z)$ and $D(x, G(x,z))$ respectively. The loss function of the generator G is composed of three loss functions: the binary cross-entropy loss, L_1-loss to reduce the outliers, and the Jaccard loss to increase the intersection between the segmented images and the ground-truth images:

$$\begin{aligned}\ell_{Gen}(G,D) &= \mathbb{E}_{x,y,z}(-\log(D(x,G(x,z)))) \\ &+ \lambda \mathbb{E}_{x,y,z}(\ell_{L_1}(y, G(x,z))) \\ &+ \alpha \mathbb{E}_{x,y,z}(\ell_{Jaccard\ loss}(y, G(x,z))),\end{aligned} \tag{7}$$

Here, λ and α are empirical weighting factors. In many cases, the adversarial loss term makes learning too slow, so SLSNet uses the L_1 *loss* to boost the learning process by properly formulating the gradient towards the expected segmented lesion boundaries. We also consider the optimization of the *Jaccard loss* for the lesion classes. If Gt is the hand-drawn ground-truth of the lesion region, and Pd its respective computer-generated segmentation mask, then the binary Jaccard loss is based on the Jaccard distance defined as follows (Yuan, 2017):

$$d_J(Gt,Pd) = 1 - J(Gt,Pd) = 1 - \frac{(Gt \cap Pd)}{|Gt| + |Pd| - |Gt \cap Pd|}. \tag{8}$$

A non-differentiable function $d_J(Gt,Pd)$ can be introduced for loss minimization. However, it is not easy to directly apply this function for back-propagation. To generate a binary mask from the continuous output of SLSNet and reduce the computation cost, we use the *Jaccard loss* function that can be defined as:

$$L_{d_J} = 1 - \frac{\sum_{i,j}(g_{ij}, p_{ij})}{\sum_{i,j} g_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j}(g_{ij}p_{ij})}, \tag{9}$$

Here, g_{ij} and p_{ij} are the pixel values at (i,j) in a ground-truth and predicted mask respectively. To balance the pixels of lesion regions and background, a weight map is used. However, this is not the case for the defined *Jaccard loss* because the Jaccard loss function is differentiable:

$$\begin{aligned}JL = \frac{\delta L_{d_J}}{\delta L_{p_{ij}}} &= -\frac{g_{ij}[\sum_{i,j} g_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j}(g_{ij}p_{ij})]}{[\sum_{i,j} g_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j}(g_{ij}p_{ij})]^2} \\ &+ \frac{(2p_{i,j} - g_{ij})[\sum_{i,j}(g_{ij}p_{ij})]}{[\sum_{i,j} g_{ij}^2 + \sum_{i,j} p_{ij}^2 - \sum_{i,j}(g_{ij}p_{ij})]^2}.\end{aligned} \tag{10}$$

During the training of SLSNet, the Jaccard loss can be efficiently integrated into the back-propagation. If the generator network is optimized correctly, the values of $D(x, G(x,z))$ approach 1.0, which means that the discriminator cannot differentiate the generated segmentation mask from the ground-truth. In this case, L_1 and Jaccard losses should approach 0.0, indicating that each generated mask matches the corresponding ground-truth mask both in overall pixel-to-pixel distances (L_1) and in tight convex surrogates (Jaccard loss) to all IoU. The loss function of the discriminator D can be expressed as follows:

$$\begin{aligned}\ell_{Dis}(G,D) &= \mathbb{E}_{x,y,z}(-\log(D(x,y))) \\ &+ \mathbb{E}_{x,y,z}(-\log(1 - D(x,G(x,z)))).\end{aligned} \tag{11}$$

Here, two terms are used to compute the binary cross-entropy (BCE) loss: the term $-\log(D(x,y))$ for ground-truth images, and $-\log(1 - D(x,G(x,z)))$ for the predicted image. The optimizer fits D by maximizing the loss values for the ground-truth images and minimizing the loss values for the predicted images. We assume that the classes expected for ground-truth and generated images are 1 and 0 respectively.

Table 2
Detailed descriptions of the skin lesion segmentation datasets ISBI 2017 and ISIC 2018.

Dataset	Training		Validation	Testing
	Before data augmentation	After data augmentation		
ISBI 2017	2000	16000	150	600
ISIC 2018	2594	20752	100	1000

4. Experimental results and discussion

4.1. Datasets

The efficacy of the proposed model, SLSNet, is assessed using two publicly available skin lesion datasets: the IEEE International Symposium on Biomedical Imaging (ISBI 2017) and Skin Lesion Analysis Towards Melanoma Detection, grand challenge datasets (ISIC 2018) (Codella et al., 2018). The ISBI 2017 dataset was divided into training, validation, and testing sets with 2000, 150, and 600 images. The ISIC 2018 dataset includes 2594 images with the corresponding ground-truth masks annotated by expert dermatologists. The validation and testing sets contain 100 and 1000 images, respectively, with no ground-truth (the evaluation can only be done on the ISIC 2018 validation leaderboard[2]). In our experiments, we used 80% of the ISIC 2018 training set to train the segmentation models and 20% for validation, as proposed in Al-Masni et al. (2018). We trained, validated, and tested SLSNet individually on the ISBI 2017 and ISIC 2018 datasets. Table 2 presents a detailed description of the two datasets.

4.2. Evaluation metrics

Five evaluation metrics are used to evaluate the performance of SLSNet. With the ISBI 2017 dataset, we use the Jaccard similarity coefficient (JSC), Dice similarity coefficient (DSC), accuracy (ACC), specificity (SPE), and sensitivity (SEN) (ISIC, 2018). For both the ground-truth y and the predicted image $\hat{y}$, the true positive (TP) rate can be defined as $TP = y \cap \hat{y}$, which is the area of the segmented region common to both $\hat{y}$ and y. The false positive (FP) rate can be defined as $FP = \overline{y} \cap \hat{y}$, which is the segmented area not belonging to y. The false-negative (FN) rate is defined as $FN = y \cap \overline{\hat{y}}$, which is the actual area missed in the predicated image. The true negative (TN) set can be defined as $TN = \overline{y} \cap \overline{\hat{y}}$, which is the set of image background common to both $\hat{y}$ and y. The mathematical expressions of the five metrics: ACC, DSC, JSC, SEN, and SPE are presented below.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$DSC = \frac{2.TP}{2.TP + FP + FN} \tag{13}$$

$$JSC = \frac{TP}{TP + FP + FN} \tag{14}$$

$$SEN = \frac{TP}{TP + FN} \tag{15}$$

$$SPE = \frac{TN}{TN + FP} \tag{16}$$

The predicted lesion masks of the ISIC 2018 challenge are assessed using a threshold JSC (JSC_{th}) (Codella, Rotemberg, Tschandl, Celebi, et al., 2019). The JSC of each test image is computed by comparing each pixel of the predicted image with its corresponding pixel in the ground-truth mask. The JSC_{th} can be formulated as follows:

$$JSC_{th} = \begin{cases} JSC & \text{if } JSC \gneq 0.65, \\ 0 & otherwise. \end{cases} \tag{17}$$

where the images with $JSC < 0.65$ will be given a score of 0.

4.3. Data augmentation and implementation

The two datasets have been augmented by flipping the images horizontally and vertically, applying gamma reconstruction ($\gamma = 0.5$, 1.0 and 1.5), and changing the contrast using adaptive histogram equalization (CLAHE) with different values for the original RGB images. We set the CLAHE threshold for the contrast limit between 1.00 and 2.00 in order to produce a variety of different contrast images. We increased the total number of training images to 16,000 and 20,752 after applying the data augmentation on ISBI 2017 and ISIC 2018 training datasets. In our experiment, we also tested the 'online augmentation' technique instead of extending the training set before the training process. The results were more or less similar.

The experiments were carried out on NVIDIA 1080Ti with 11 GB memory (training time was approximately 8 h). We implemented the proposed model on the PyTorch framework.[3] Adam optimizer (Kingma & Ba, 2014) was used with the parameters $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate and the batch size were set to 0.0002 and 8 respectively. The weighting factors of L_1-loss and Jaccard loss (λ and α) were 0.1 and 0.5, respectively. It should be noted that all layers of the generator and discriminator networks of the proposed model were trained from scratch.

4.4. Experimental results

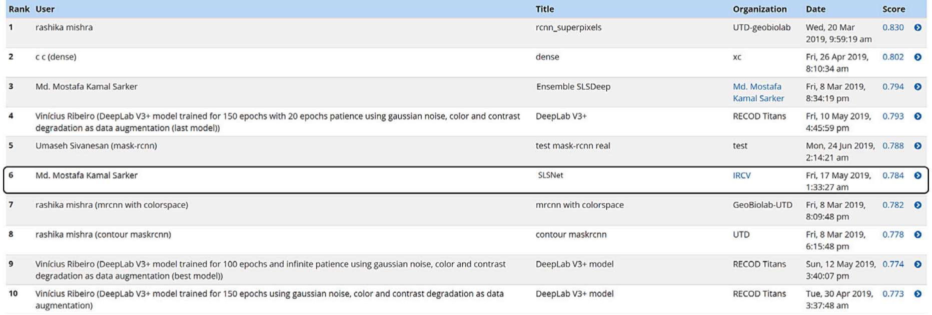
The size of the ISBI 2017 and ISIC 2018 images ranges from 542 × 718 to 2848 × 4288. These image sizes are too large for the purpose of training the proposed model. Thus, we resized the input images to $H \times W$ pixels to speed-up the training process, where H and W are the height and width of the images to be fed into the network. To select the best image size, we assessed the performance of the proposed model with three image sizes (64 × 64, 128 × 128, and 256 × 256). It should be noted that the best segmentation results are obtained with the input image size 128 × 128 (the ablation study is given below).

Tables 3 and 4, present the quantitative results of the proposed model on the ISBI 2017 test and ISIC 2018 validation sets. Table 3, uses the ISBI 2017 test dataset to compare the SLSNet with ten skin lesion segmentation methods: FCN (Long et al., 2015), U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), FrCN (Al-Masni et al., 2018), SLSDeep (Sarker et al., 2018), SegAN (Xue et al., 2018), YOLO+grabcut (Unver & Ayan, 2019), LCASA-Net (Ma et al., 2019), ENet (Paszke et al., 2016) and DAGAN (Lei et al., 2020). It should be noted that the test results of FCN, U-Net, SegNet, FrCN are taken from Al-Masni et al. (2018). The test results of LCASA-Net and ENet are taken from Ma et al. (2019) while the test results of DAGAN are taken from Lei et al. (2020). As can be seen in Table 3, SLSNet outperforms all the tested methods in terms of the ACC, DSC, JSC, and SPE metrics. SLSNet achieves ACC, DSC and JSC scores of 97.61%, 90.63% and 81.98%, which is 2.97%, 2.73% and 1.08% higher than the scores of the second-best method (i.e., LCASA-Net). Similarly, SLSNet yields DSC and SPE scores of 90.63% and 99.92%, which are 2.83% and 1.62% higher than the scores of the SLSDeep method. In turn, the YOLO+grabcut yields a SEN of 90.82%, which is 3.01% higher than SLSNet. SLSNet has fewer parameters (2.35 million parameters) than all other methods except LCASA-Net (Ma et al., 2019) and ENet (Paszke et al., 2016). Although LCASA-Net and ENet are lightweight models with 0.49 and 0.36 million parameters, respectively, they obtain DSC scores of 87.9%

[2] https://challenge.isic-archive.com/

Table 3
Comparing the performance of the proposed model with 10 state-of-the-art skin lesion segmentation methods on the ISBI 2017 dataset (test set) in terms of the accuracy and number of parameters (in millions) (Params (M)).

Methods	ACC	DSC	JSC	SEN	SPE	Params (M)
Encoder–decoder based						
FCN (Long et al., 2015)	92.72	83.83	72.17	79.98	96.66	134.3
U-Net (Ronneberger et al., 2015)	90.14	76.27	61.64	67.15	97.24	12.3
SegNet (Badrinarayanan, Kendall, & Cipolla, 2017)	91.76	82.09	69.63	80.05	95.37	11.50
FrCN (Al-Masni et al., 2018)	94.03	87.08	77.11	85.40	96.69	16.30
SLSDeep (Sarker et al., 2018)	93.60	87.80	78.2	81.60	98.30	46.65
YOLO+grabcut (Unver & Ayan, 2019)	93.39	84.26	74.81	**90.82**	92.68	–
LCASA-Net (Ma et al., 2019)	94.70	87.90	80.90	–	–	0.49
ENet (Paszke et al., 2016)	92.0	82.7	74.1	–	–	**0.36**
GAN based						
SegAN (Xue et al., 2018)	94.10	86.70	78.50	–	–	382.17
DAGAN (Lei et al., 2020)	93.50	85.90	77.1	83.50	97.60	–
Proposed SLSNet	**97.61**	**90.63**	**81.98**	87.81	**99.92**	2.35



Rank	User	Title	Organization	Date	Score
1	rashika mishra	rcnn_superpixels	UTD-geobiolab	Wed, 20 Mar 2019, 9:59:19 am	0.830
2	c c (dense)	dense	xc	Fri, 26 Apr 2019, 8:10:34 am	0.802
3	Md. Mostafa Kamal Sarker	Ensemble SLSDeep	Md. Mostafa Kamal Sarker	Fri, 8 Mar 2019, 8:34:19 pm	0.794
4	Vinícius Ribeiro (DeepLab V3+ model trained for 150 epochs with 20 epochs patience using gaussian noise, color and contrast degradation as data augmentation (last model))	DeepLab V3+	RECOD Titans	Fri, 10 May 2019, 4:45:59 pm	0.793
5	Umaseh Sivanesan (mask-rcnn)	test mask-rcnn real	test	Mon, 24 Jun 2019, 2:14:21 am	0.788
6	Md. Mostafa Kamal Sarker	SLSNet	IRCV	Fri, 17 May 2019, 1:33:27 am	0.784
7	rashika mishra (mrcnn with colorspace)	mrcnn with colorspace	GeoBiolab-UTD	Fri, 8 Mar 2019, 8:09:48 pm	0.782
8	rashika mishra (contour maskrcnn)	contour maskrcnn	UTD	Fri, 8 Mar 2019, 6:15:48 pm	0.778
9	Vinícius Ribeiro (DeepLab V3+ model trained for 100 epochs and infinite patience using gaussian noise, color and contrast degradation as data augmentation (best model))	DeepLab V3+ model	RECOD Titans	Sun, 12 May 2019, 3:40:07 pm	0.774
10	Vinícius Ribeiro (DeepLab V3+ model trained for 150 epochs using gaussian noise, color and contrast degradation as data augmentation)	DeepLab V3+ model	RECOD Titans	Tue, 30 Apr 2019, 3:37:48 am	0.773

Fig. 5. The rank of SLSNet on the ISIC 2018 leaderboard challenge (screenshot). SLSNet is highlighted.

Table 4
Evaluating the performance of SLSNet on the ISIC 2018 validation dataset in terms of the JSC_{th} and number of parameters (in millions) (Params (M)).

Methods	JSC_{th}	Params (M)
Encoder–decoder based		
FCN (Long et al., 2015)	74.70	134.30
U-Net (Ronneberger et al., 2015)	54.40	12.30
SegNet (Badrinarayanan et al., 2017)	69.50	11.50
FrCN (Al-Masni et al., 2018)	74.60	16.30
Rcnn-superpixels (Mishra & Daescu, 2019)	**83.00**	–
Mask R-CNN (Sivanesan, Braga, Sonnadara, & Dhindsa, 2019)	78.80	–
GAN based		
GAN-FCN (Bi, Feng, & Kim, 2018)	77.80	10.61
Proposed SLSNet	78.40	**2.35**

and 82.7% respectively. Note that these scores are 2.73% and 7.93% lower than the scores of SLSNet.

Table 4 compares SLSNet with the FCN, U-Net, SegNet, FrCN, GAN-FCN, Rcnn-superpixels and Mask R-CNN models using the ISIC 2018 validation dataset. Note that the ISIC 2018 dataset includes 100 images for validation and 1000 images for testing, without ground-truth. The evaluation can only be done on the ISIC 2018 validation leaderboard. The validation results of FCN, U-Net, SegNet, and FrCN are taken from Al-masni et al. (2018). SLSNet outperforms the GAN-FCN and FrCN in terms of the JSC_{th} score with improvements of 0.6% and 3.8%, respectively. As can be seen in Table 4, SLSNet has much fewer parameters than all the other models. In particular, the SLSNet model has 2.35 million parameters while the GAN-FCN model (the closest one) has 10.61 million. The SegAN model, a GAN-based model, has 382.17 million parameters (the heaviest model). The number of parameters SLSNet has 57, 5, 4, 6, and 19 times lower than the FCN, U-Net, SegNet, FrCN, and SLSDeep models, respectively, thanks to the use of 1-D kernel, PAM, and CAM, which significantly reduces the number of parameters.

Fig. 5 shows the rank of SLSNet with the ISIC 2018 challenge validation set. The proposed model is highlighted with a black box and entitled SLSNet. It was ranked in the 6th position at the time of submission (17 May 2019, authored by the IRCV group). The methods preceding SLSNet on the leaderboard (rank 1 to 5) use residual networks (i.e., ResNet) (He, Zhang, Ren, & Sun, 2016). It should be noted that ResNet50 has 23 million parameters, considerably more than the number of parameters that SLSNet has. In turn, Fig. 5 and Table 4 demonstrate that the Mask R-CNN (Sivanesan et al., 2019) and Rcnn-superpixels (Mishra & Daescu, 2019) models achieve a JSC_{th} score a little higher than SLSNet. However, the authors of Mask R-CNN (Sivanesan et al., 2019) and Rcnn-superpixels (Mishra & Daescu, 2019) did not mention the number of parameters each model has. As shown, both Mask R-CNN (Sivanesan et al., 2019) and Rcnn-superpixels (Mishra & Daescu, 2019) use the ResNet 101 backbone. Since ResNet 101 has 44.5 million parameters, the Mask R-CNN and Rcnn-superpixel models are much heavier than SLSNet.

Fig. 6 presents the qualitative segmentation results of SLSNet with some examples from the ISBI 2017 test dataset. As can be seen in Fig. 6(left), the regions of skin lesions and healthy skin have similar colors. Some skin lesions are also tiny and have fuzzy boundaries. However, SLSNet accurately segments the boundary of skin lesions with an accuracy of approximately 95%. The examples in Fig. 6(right) have tiny skin regions when compared to the size of the lesion regions. Also, the lesion regions fill most of the image and intersect three margins of the images. SLSNet yields inaccurate segmentation results in these cases because it is difficult to segment the boundaries of skin lesions accurately when there is no proper boundary between the lesion and healthy skin tissue. In such cases, it is hard for any segmentation model to precisely delineate the shape of the lesion region to get a proper segmentation.

[3] https://pytorch.org/

Table 5
Comparing the inference times of SLSNet, ENet (two lightweight architectures) and GAN-FCN at different image resolutions.

Model	64 × 64			128 × 128			256 × 256		
	ms	fps	MACs(G)	ms	fps	MACs(G)	ms	fps	MACs(G)
GAN-FCN (Bi et al., 2018)	9	120.64	0.37	14	87.62	1.51	21	57.15	6.06
ENet (Paszke et al., 2016)	**2**	**273.65**	**0.03**	**3**	**136.84**	**0.13**	**7**	**91.29**	**0.50**
Proposed SLSNet	5	168.71	0.23	8	110.3	0.90	14	78.63	3.16

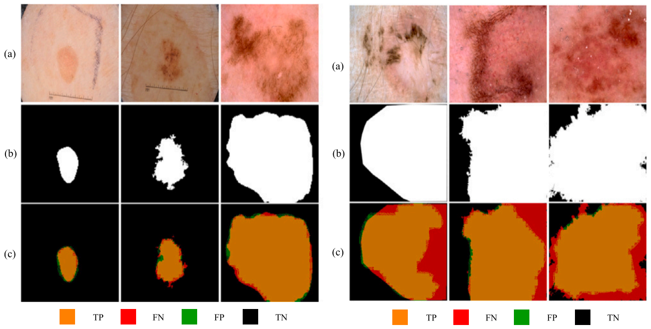


Fig. 6. Segmentation results of SLSNet: (a) input image (b) ground-truth (c) left: accurately segmented lesions (c) right: incorrectly segmented lesions. TP, FN, FP and TN stand for True Positive, False Negative, False Positive and True Negative, respectively.

Table 5 presents the inference times of SLSNet and the lightweight model GAN-FCN (Bi et al., 2018) with different input image sizes (64 × 64, 128 × 128 and 256 × 256). With an image size of 128 × 128, the inference time of SLSNet is 8 ms (around 110 FPS). As can be seen, FPS is higher for SLSNet than for GAN-FCN but lower than for ENet with the same input image size of 128 × 128. To compare the computational complexity of GAN-FCN, ENet, and SLSNet, we computed the multiply-accumulate operation (MACs) in billions of operations, (MACs(G)). As shown in Table 5, SLSNet, ENet and GAN-FCN have 0.90, 0.13 and 1.51 MACs(G), respectively, with an input image size of 128 × 128. It should be noted that the segmentation performance of ENet is poorer than that of SLSNet. Indeed, SLSNet can provide fast and accurate skin lesion segmentation results. On the basis of the computed inference time, it is obvious that SLSNet can be run on a single Mobile GPU while assuring accurate skin lesion segmentation in real time.

4.5. Comparing different variations of slsnet

Here, we assessed the effect of PAM and CAM block on the baseline GAN model with and without the multiscale block. Firstly, we assessed the baseline GAN segmentation model (BL GAN). The *G* network of BL GAN has sequentially stacked factorized kernels in all convolution and deconvolution layers. As shown in Table 6, BL GAN has DSC and JSC scores of 83.61% and 72.93%, respectively. Secondly, we added the PAM block to the BL GAN model (BL + PAM). Specifically, in this BL + PAM model, we added a PAM module after each downsampling and upsampling layer in both the encoder and decoder parts. We also added a PAM module after the first downsampling layer in the discriminator network. The BL + PAM model has DSC and JSC scores of 86.01% and 75.96%, respectively. Thirdly, we added a CAM block to the BL GAN model (BL + CAM). In the BL + CAM model, we added a CAM module after each factorized layer in the *G* network. We also added a CAM module after the second downsampling layer in the discriminator network. The BL + CAM model gives DSC and JSC scores of 87.23% and 76.65%, respectively. As can be seen in Table 6, the scores of BL + CAM are better than BL + PAM, thanks to the addition of the CAM mechanism to the BL GAN model, which provides efficient feature discriminability between skin lesion regions and normal skin regions in skin images.

Furthermore, we added the PAM and CAM blocks to the BL GAN model without multiscale (BL + CAM + PAM w/o MS). This model has DSC and JSC scores of 88.87% and 78.76%, respectively. As shown in the last row of Table 6, the addition of the multiscale, PAM, and CAM blocks to the BL GAN increases the DSC and JSC scores to 90.63% and 81.98%, respectively. As can be seen in Table 6, SLSNet outperforms the other BL GAN variations on all evaluation metrics. Specifically, with an input image size of 128 × 128, SLSNet and BL GAN achieve 0.90 and 0.85 MACs(G), respectively. However, the number of operations of SLSNet is 0.05 higher than the baseline model BL GAN, and SLSNet increases its JSC score and Dice score by 9% and 7%, respectively.

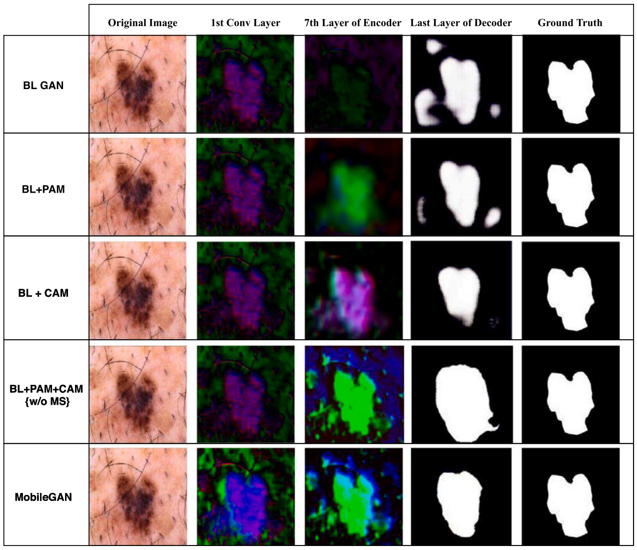


Fig. 7. Visualization of the intermediate layers of different configurations of SLSNet. BL GAN, PAM, CAM and w/o MS refer to the Baseline Generative Adversarial Network, Position Attention Module, Channel Attention Module and without Multi-Scale, respectively.

Fig. 7 shows the intermediate layers of the different configurations of SLSNet. As can be seen, the additions of CAM and PAM help generate skin lesion relevant features in the 7 layer, which means that the addition of CAM and PAM helps the network encode features that discriminate between the boundaries of melanoma and non-melanoma regions. Additionally, the resulting binary masks are refined in the decoder layers, by reducing the artifacts on the melanoma boundaries and in the background. Besides, the insertion of the multiscale block at the beginning of the proposed model is a considerable improvement on the variations in the single-scale model.

Indeed, most of the related deep learning-based segmentation models downsample the input images to avoid the high computation requirements of the deep models. For example, FrCN used an input image size of 192 × 256, and SLSDeep an input image size of 384 × 384. Table 7 presents the results of the proposed model with different resolutions of input skin images (64 × 64, 128 × 128, and 256 × 256). With an input image size of 64 × 64, the last layer of SLSNet generates an 8 × 8 feature map that yields very coarse level information, in which most of the important details are lost and segmentation accuracy is low. With an input image size of 256 × 256, the last layer in SLSNet generates a feature map of 32 × 32, which also extracts skin lesion-irrelevant features (i.e., artifacts) that reduce the overall accuracy. In turn, the input image size 128 × 128 generates a 16 × 16 feature map

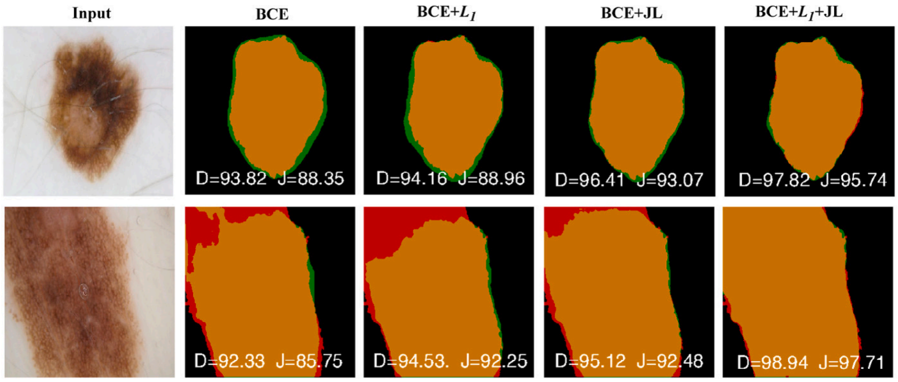


Fig. 8. The effect of different loss functions on the performance of SLSNet with the test set of ISBI 2017. BCE, JL, D, and J stand for Binary Cross-Entropy, Jaccard Loss, Dice Similarity Coefficient, and Jaccard Similarity Coefficient, respectively.

Table 6
The performance of different configurations of SLSNet with the ISBI 2017 test dataset (highest values are in bold).

Methods	ACC	DSC	JSC	SEN	SPE	MACs(G)
BL GAN	95.63	83.61	72.93	79.42	96.91	0.85
BL + PAM	96.72	86.01	75.96	82.48	98.40	0.88
BL + CAM	96.90	87.23	76.65	83.31	99.10	0.85
BL + PAM + CAM w/o MS	97.25	88.87	78.76	85.49	99.59	0.88
SLSNet	**97.61**	**90.63**	**81.98**	**87.81**	**99.92**	**0.90**

Table 7
The performance of SLSNet with different input image resolutions of the ISBI 2017 test dataset.

Input size	ACC	DSC	JSC	SEN	SPE	MACs(G)
64 × 64	94.44	87.59	77.76	84.16	95.36	0.23
128 × 128	**97.61**	**90.63**	**81.98**	**87.81**	**99.92**	0.90
256 × 256	96.72	89.72	79.49	86.36	98.21	3.61

Table 8
Effect of different loss functions on the performance of SLSNet with the ISBI 2017 test dataset.

Loss	ACC	DSC	JSC	SEN	SPE
SLSNet + BCE	95.32	85.11	74.48	83.71	98.02
SLSNet + BCE + L_1	96.90	87.26	76.80	85.05	99.30
SLSNet + BCE + Jaccard Loss	96.97	89.56	79.88	86.90	99.49
SLSNet + BCE + L_1 + Jaccard Loss	**97.61**	**90.63**	**81.98**	**87.81**	**99.92**

at the last layer that retains skin lesion-relevant features and discards the irrelevant ones, thus yielding the best segmentation. Since the main goal of this study is to achieve a lightweight skin lesion segmentation model with high accuracy, we used the input image size 128 × 128 to train SLSNet. It should be noted that the size of the input images has a potential impact on MACs operations. The image size of 128 × 128 yields 0.90 MACs(G), which is 0.67 MACs(G) higher than for image size of 64 × 64, and 2.71 MACs(G) lower than for image size 256 × 256.

Table 8 presents the effect of different loss function variations on SLSNet performance. SLSNet + BCE, SLSNet + BC E + L_1, SLSNet + BCE +Jaccard Loss and SLSNet + BCE + L_1 + Jaccard Loss obtain incremental JSC scores of 74.48%, 76.80%, 79.88% and 81.98%, respectively. The proposed loss function significantly improves the five evaluation metrics. The use of the L_1-loss function reduces the sensitivity to outliers. In turn, the use of Jaccard loss enables SLSNet to detect subtle abnormalities that the cross-entropy loss could not detect. It should be noted that the combination of the BCE + L_1 + Jaccard Loss with SLSNet considerably decreases the number of false positives in the resulting segmentation masks. Fig. 8 presents two difficult samples from the ISBI 2017 dataset with the DSC and JSC scores for each mask. D and J stand for DSC and JSC, respectively. The BCE + L_1 + Jaccard Loss obtains a JSC score of 95.74% and 97.71% with the top and bottom examples in Fig. 8, respectively. This analysis shows that the proposed loss function (BCE + L_1 + Jaccard Loss) yields the best improvement in the segmentation results.

4.6. Limitations

In some cases, when the lesion regions intersect the margins of the images, the segmentation results of SLSNet are inaccurate. Indeed, it is a real challenge to accurately segment the boundaries of skin lesions when there is no proper boundary between the lesion and healthy skin tissue. In these cases, a sequence of images is needed to capture the whole melanoma area, which does not appear in a single image. Moreover, it is essential to consider the temporal coherence of a sequence of images. The proposed method does not consider the temporal cues to increase accuracy or efficiency. Neither does it tackle the problem of coherence. Besides, we reduced the number of parameters and operations of the proposed method by rescaling/resampling the original skin images fed into the network and decreasing the image size, so they may suffer from aliasing or blurriness and artifacts on the object boundaries of some segmented images. All these limitations will be considered and investigated in future studies.

5. Conclusions

This article has proposed a lightweight and efficient GAN-based model for skin lesion segmentation, called SLSNet. SLSNet was built by adapting a GAN model that consists of a 1-D kernel factorized network, multiscale aggregation, and position and channel attention mechanisms. SLSNet was been assessed on the ISBI 2017 test and ISIC 2018 validation datasets. With the ISBI 2017 test dataset, SLSNet yielded precise segmentation results with accuracy, sensitivity, specificity, Dice coefficient, and Jaccard index of 97.61%, 87.81%, 99.92%, 90.63%, and 81.98%, respectively. In turn, SLSNet achieved a threshold JSC score of 78.4% with the ISIC 2018 validation dataset. SLSNet has approximately 2.35 million parameters and its segmentation accuracy is comparable to the state-of-the-art methods. In future work, we will implement a mobile application based on the SLSNet model to segment skin lesions in images captured by low-resolution cameras.We will also assess the performance of SLSNet with different embedded systems, such as NVIDIA Jetson GPUs, FPGAs, and Mobile SoCs, on four different parameters: upgradeability, deployment, efficiency, and performance.

CRediT authorship contribution statement

Md. Mostafa Kamal Sarker: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Visualization, Investigation, Validation, Data curation. **Hatem A. Rashwan:** Conceptualization, Methodology, Writing - original draft, Writing - review

& editing, Visualization, Investigation, Supervision. **Farhan Akram:** Writing - review & editing. **Vivek Kumar Singh:** Writing - review & editing. **Syeda Furruka Banu:** Data curation, Writing - review & editing. **Forhad U.H. Chowdhury:** Data curation, Writing - review & editing. **Kabir Ahmed Choudhury:** Writing - review & editing. **Sylvie Chambon:** Writing - review & editing. **Petia Radeva:** Conceptualization, Writing - original draft, Writing - review & editing, Supervision. **Domenec Puig:** Supervision, Project administration, Funding acquisition, Writing - review & editing. **Mohamed Abdel-Nasser:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is funded by the program "Martí i Franquès" under the agreement between Universitat Rovira Virgili and Fundacio Catalunya La Pedrera, projects RTI2018-095232-B-C2, SGR 1742, CERCA, Nestore Horizon2020 SC1-PM-15-2017 (n 769643), EIT Health Validithi project. The authors acknowledge the support of the NVIDIA Corporation who donated several Titan Xp GPU used for this research. M. Abdel-Nasser, Hatem A. Rashwan and D. Puig are partially supported by the Spanish Government under project PID2019-105789RB-I00.

References

Adegun, A., & Viriri, S. (2019). Deep learning model for skin lesion segmentation: Fully convolutional network. In *International conference on image analysis and recognition* (pp. 232–242). Springer.

Agarwal, A., Issac, A., Dutta, M. K., Riha, K., & Uher, V. (2017). Automated skin lesion segmentation using K-means clustering from digital dermoscopic images. In *2017 40th international conference on telecommunications and signal processing* (pp. 743–748). IEEE.

Al-Masni, M. A., Al-antari, M. A., Choi, M. -T., Han, S. -M., & Kim, T. -S. (2018). Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. *Computer Methods and Programs in Biomedicine*, *162*, 221–231.

Al-masni, M., Al-antari, M., Rivera, P., Valarezo, E., Gi, G., Kim, T., et al. (2018). Automatic skin lesion boundary segmentation using deep learning convolutional networks with weighted cross entropy. In *ISIC2018: Skin image analysis workshop and challenge*.

Apalla, Z., Nashan, D., Weller, R. B., & Castellsague, X. (2017). Skin cancer: Epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatology and Therapy*, *7*, 5–19.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*, 2481–2495.

Bethge, J., Bartz, C., Yang, H., Chen, Y., & Meinel, C. (2020). MeliusNet: Can binary neural networks achieve MobileNet-level accuracy? arXiv preprint arXiv:2001.05936.

Bi, L., Feng, D., & Kim, J. (2018). Improving automatic skin lesion segmentation using adversarial learning based data augmentation. arXiv:1807.08392.

Bi, L., Kim, J., Ahn, E., Kumar, A., Feng, D., & Fulham, M. (2019). Step-wise integration of deep class-specific learning for dermoscopic image segmentation. *Pattern Recognition*, *85*, 78–89.

Bisla, D., Choromanska, A., Stein, J. A., Polsky, D., & Berman, R. (2019). Skin lesion segmentation and classification with deep learning system. arXiv preprint arXiv:1902.06061.

Bissoto, A., Perez, F., Ribeiro, V., Fornaciali, M., Avila, S., & Valle, E. (2018). Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD titans at ISIC challenge 2018. arXiv preprint arXiv:1808.08480.

Celebi, M. E., Wen, Q., Iyatomi, H., Shimizu, K., Zhou, H., & Schaefer, G. (2015). A state-of-the-art survey on lesion border detection in dermoscopy images. *Dermoscopy Image Analysis*, 97–129.

Chen, L. -C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3640–3649).

Cheng, Y., Wang, D., Zhou, P., & Zhang, T. (2017). A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Kittler, H., et al. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging, hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 168–172). IEEE.

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., et al. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv:1902.03368.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115.

Fu, J., Liu, J., Tian, H., Fang, Z., & Lu, H. (2018). Dual attention network for scene segmentation. arXiv preprint arXiv:1809.02983.

Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, *7*, 87–93.

Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *International conference on learning representations*.

Hartanto, C. A., & Wibowo, A. (2020). Development of mobile skin cancer detection using faster r-CNN and mobilenet v2 model. In *2020 7th international conference on information technology, computer, and electrical engineering* (pp. 58–63). IEEE.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

ISIC (2018). ISIC 2018: Skin lesion analysis towards melanoma detection. https://challenge2018.isic-archive.com/live-leaderboards/.

Isola, P., Zhu, J. -Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Izadi, S., Mirikharaji, Z., Kawahara, J., & Hamarneh, G. (2018). Generative adversarial networks to segment skin lesions. In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 881–884). IEEE.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lateef, F., & Ruichek, Y. (2019). Survey on semantic segmentation using deep learning techniques. *Neurocomputing*, *338*, 321–348.

Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., et al. (2020). Skin lesion segmentation via generative adversarial networks with dual discriminators. *Medical Image Analysis*, *64*, Article 101716. http://dx.doi.org/10.1016/j.media.2020.101716, URL: http://www.sciencedirect.com/science/article/pii/S1361841520300803.

Li, X., Yu, L., Chen, H., Fu, C. -W., & Heng, P. -A. (2018). Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. arXiv preprint arXiv:1808.03887.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Ma, D., Wu, H., Sun, J., Yu, C., & Liu, L. (2019). A light-weight context-aware self-attention model for skin lesion segmentation. In *Pacific rim international conference on artificial intelligence* (pp. 501–505). Springer.

Mahmoud, H., Abdel-Nasser, M., & Omer, O. A. (2018). Computer aided diagnosis system for skin lesions detection using texture analysis methods. In *2018 international conference on innovative trends in computer engineering* (pp. 140–144). IEEE.

Mishra, R., & Daescu, O. (2019). AlgoDerm: An end-to-end mobile application for skin lesion analysis and tracking. In *Proc. int. conf. health informat. med. syst* (pp. 3–9).

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204–2212).

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Paszke, A., Chaurasia, A., Kim, S., & Culurciello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

Polino, A., Pascanu, R., & Alistarh, D. (2018). Model compression via distillation and quantization. In *International conference on learning representations*.

Rahman, M., Alpaslan, N., & Bhattacharya, P. (2016). Developing a retrieval based diagnostic aid for automated melanoma recognition of dermoscopic images. In *Applied imagery pattern recognition workshop* (pp. 1–7). IEEE.

Rastegari, M., Ordonez, V., Redmon, J., & Farhadi, A. (2016). Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision* (pp. 525–542). Springer.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*, 2352–2449.

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2018). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, *19*, 263–272.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Sae-Lim, W., Wettayaprasit, W., & Aiyarak, P. (2019). Convolutional neural networks using mobilenet for skin lesion classification. In *2019 16th international joint conference on computer science and software engineering* (pp. 242–247). IEEE.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. -C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510–4520).

Sarker, M. M. K., Rashwan, H. A., Akram, F., et al. (2018). SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *International conference on MICCAI* (pp. 21–29). Springer.

Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., et al. (2018). Attention-gated networks for improving ultrasound scan plane detection. arXiv preprint arXiv:1804.05338.

Silveira, M., Nascimento, J. C., Marques, J. S., Marçal, A. R., Mendonça, T., Yamauchi, S., et al. (2009). Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. *IEEE Journal of Selected Topics in Signal Processing, 3*, 35–45.

Sivanesan, U., Braga, L. H., Sonnadara, R. R., & Dhindsa, K. (2019). Unsupervised medical image segmentation with adversarial networks: From edge diagrams to segmentation maps. arXiv preprint arXiv:1911.05140.

Tung, F., & Mori, G. (2019). Similarity-preserving knowledge distillation. In *Proceedings of the IEEE international conference on computer vision* (pp. 1365–1374).

Unver, H. M., & Ayan, E. (2019). Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm. *Diagnostics, 9*, 72.

Vesal, S., Ravikumar, N., & Maier, A. (2018). SkinNet: A deep learning framework for skin lesion segmentation. arXiv preprint arXiv:1806.09522.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., et al. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156–3164).

Xue, Y., Xu, T., & Huang, X. (2018). Adversarial learning with multi-scale loss for skin lesion segmentation. In *2018 IEEE 15th international symposium on biomedical imaging* (pp. 859–863). IEEE.

Yu, L., Chen, H., Dou, Q., Qin, J., & Heng, P. -A. (2017). Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging, 36*, 994–1004.

Yuan, Y. (2017). Automatic skin lesion segmentation with fully convolutional-deconvolutional networks. arXiv preprint arXiv:1703.05165.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018). Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318.

Zhao, Z. -Q., Zheng, P., Xu, S. -t, & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks Learning Systems*.

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., & Zou, Y. (2016). Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. ArXiv Preprint arXiv:1606.06160.