IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 24, 2019, accepted September 14, 2019, date of publication September 25, 2019, date of current version October 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2943628

DSM: A Deep Supervised Multi-Scale Network Learning for Skin Cancer Segmentation

GUOKAI ZHANG[1], XIAOANG SHEN[1], SIRUI CHEN[2], LIPENG LIANG[1], YE LUO[1], JIE YU[3], AND JIANWEI LU[1,4]

[1]School of Software Engineering, Tongji University, Shanghai 201804, China
[2]College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
[3]Qingdao Central Hospital, Qingdao 266042, China
[4]Institute of Translational Medicine, Tongji University, Shanghai 200092, China

Corresponding authors: Ye Luo (yeluo@tongji.edu.cn), Jie Yu (timjie@live.cn), and Jianwei Lu (jwlu33@tongji.edu.cn)

This work was supported in part by the General Program of National Natural Science Foundation of China (NSFC) under Grant 61806147, in part by the Shanghai Natural Science Foundation of China under Grant 18ZR1441200, in part by the Fundamental Research Funds for the Central Universities under Grant 22120180012, and in part by the NSFC under Grant 81571347 and Grant 61572362.

ABSTRACT The automatic segmentation of the skin lesion on dermoscopy images is an important step for diagnosing the melanoma. However, the skin lesion segmentation is still a challenging task due to the blur lesion border, low contrast between the skin cancer region and normal tissue background, and various sizes of cancer regions. In this paper, we propose a deep supervised multi-scale network (DSM-Network), which achieves satisfied skin cancer segmentation result by utilizing the side-output layers of the network to aggregate information from shallow&deep layers, and designing a multi-scale connection block to handle a variety of cancer sizes' changes. Moreover, a post-processing of the contour refinement strategy is adopted by a conditional random field (CRF) model to further improve the segmentation results. Extensive experiments on two public datasets: ISBI 2017 and PH2 have demonstrated that our designed DSM-Network has gained competitive performance compared with other state-of-the-art methods.

INDEX TERMS Skin cancer, dermoscopy image, deep supervised learning, multi-scale feature, conditional random field.

I. INTRODUCTION

Skin cancer has become one of the most common malignant tumors in the world, and the death rate of melanoma (i.e., a kind of skin cancer) has increased to 75% [1]. The early diagnosis of melanoma can significantly improve patients' survival rate. Currently, the most common way to detect this cancer is by using dermoscopy images. Dermoscopy is a popular tool in vivo non-invasive imaging that employs polarized light to obtain magnified and illuminated images of localized areas of the skin, and it can help improve diagnostic performance and reduce skin cancer mortality. However, most of the current dermoscopy images are manually analyzed by dermatologists, which is time-consuming, expensive, and laborious. Meanwhile, the diagnosis result is easily biased by each individual dermatologist. Computer-aided system (CAD) can effectively solve these problems. It not only improves the detection efficiency but also helps make diagnosis more objective. Therefore, the development of a computer-aided detection system for melanoma segmentation is imminent and essential.

At present, according to the types of extracted skin cancer features, there are mainly two categories of methods (hand-crafted, deep learning) proposed in skin cancer segmentation. One is based on manually defined traditional features such as color, shape, size, texture, and so on. For example, Abbas *et al.* [2] proposed a method using a dynamic programming technique to find the optimal boundary of the lesion in dermoscopy images. Celebi *et al.* [3] presented an automated method for skin lesion border detection by using ensembles of four different thresholding methods. In order to improve the performance of the classic gradient vector flow snake, Zhou *et al.* [4] integrated a mass density function into the optimization objective functional, which can be solved with the support of mean shift estimation. However, the optimization process involves a large amount of computation to converge. Celebi *et al.* [5] performed an automatic border detection to segment the lesion from the background skin, and

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Song.

then extracted a series of color, texture, and shape features based on the extracted lesion region [6]. Stanley *et al.* [7] used the color histogram analysis over a training set of images to determine the colors characteristics of melanoma. Messadi *et al.* [8] applied a sequence of transformations to the image to measure a set of attributes (A: asymmetry, B: border, C: color and D: diameter) which contain sufficient information to differentiate melanoma from benign lesions.

Unlike manually defined traditional features, deep learning can automatically extract features by learning from a hierarchical network structure. Besides the great success of deep learning in natural image recognition tasks such as image classification [9], convolution neural networks (CNNs) have also shown promising performance in various medical image computing problems. For skin segmentation task, Sadri *et al.* [10] introduced a fixed-grid wavelet network, in which orthogonal least squares were used to calculate the network weights and then to optimize the network structure. A supervised method in [12] was proposed, in which a self-generating neural network is combined with the genetic algorithms for skin lesions segmentation. Yuan *et al.* [13] developed an end-to-end deep convolution neural network with a jaccard distance-based loss for skin lesion segmentation without prior knowledge and sample re-weighting. Li *et al.* [14] presented a new dense deconvolutional network based on residual learning to segment lesions. Mirikharaji *et al.* [15] proposed to encode the star shape prior to the loss of deep convolutional neural network to guarantee a global structure in segmentation results. Sarker *et al.* [16] applied a robust deep encoder-decoder network learning framework to segment the boundaries of lesion regions accurately. To simultaneously produce segmentation and coarse classification result, two fully convolutional residual networks were proposed in [17]. A deep ResNet was also utilized in [18] to enhance robust visual features learning, and they used 50 layers for skin lesions segmentation to obtain good performance. Moreover, an enhanced Convolutional-Deconvolutional Network [11] was used for automatic segmentation of skin lesions. However, the proposed method in [11] can't realize deep supervision, which leads to the loss of lots of detailed information thus can't further improve the performance of segmentation on skin lesion. Codella *et al.* [19] proposed a system that combines recent developments in deep learning with established machine learning approaches. Specifically, it creates ensembles of methods that are capable of segmenting skin lesions and analyzing the detected area and surrounding tissues for melanoma detection.

Although those methods have achieved great success, there still remain several challenges to the skin cancer segmentation task. Firstly, dermoscopy image may include hair, blood vessels, and other factors that interfere with segmentation. Moreover, the low contrast between the lesion area and the surrounding skin causes blurry boundary, which makes it difficult to segment the lesion accurately. At last, melanoma usually has different sizes, shapes, and colors depending on different skin condition, which could be a hamper to achieve high segmentation accuracy.

In order to accurately segment the skin cancer, we propose a deep supervised multi-scale network (DSM-Network) to extract strong and robust features of skin cancers. Before extracting features by the proposed DSM-Network, we first preprocess the input image to remove the possible hairs influence on the lesion region, then two image contrast enhancement techniques are utilized to generate two additional images as the inputs to the network. After obtaining the primary prediction mask of the input image by DSM-Network, a contour refinement by CRF is applied to further enhance the segmentation result. And the final prediction mask is the averaged one among the results of the original image and two contrast-enhanced versions of the original image. Although our DSM-Network shares similar network architecture as U-Net, we distinguish ourselves from the existing methods from the following three aspects: (1) we add extra residual block after each convolution layer to enhance the feature learning ability of the network; (2) we design a multi-scale connection (MSC) block at each skip connection layer of the network to handle variety of cancer size changes; (3) we aggregate complementary information at different layer by side-output layers considering that the deeper side outputs encode high-level semantic knowledge while the shallower side outputs capture rich spatial information which is capable of successfully highlighting the boundaries of the cancers.

The rest of the paper is organized as follows. In Section II, we introduce our proposed approach of DSM-Network. In Section III, we conduct experiments on two public datasets: ISBI 2017 and PH2, and then make thorough comparisons with state-of-the-art methods. In the last Section, we present the conclusion and discussions of the future work.

II. THE PROPOSED APPROACH

The overview of the proposed DSM-Network architecture is illustrated in Fig. 1. The backbone of our DSM-Network is U-Net [20] which has been successfully used in many medical image segmentation tasks. Like U-Net based methods, our method also consists of the encoder and the decoder stages, respectively. The encoder stage is composed of successive convolution and pooling layers to extract the image features, while the decoder stage is to upsample the feature maps at different layers to the original image size, such that they can be concatenated for the final segmentation task.

The pipeline of the proposed method is as follows. Before sending one input image to our DSM-Network, we first apply a preprocessing operation to the input image to remove the possible hairs influence on the lesion region, and then two image contrast enhancement methods are performed individually. Both the enhanced images and the original images are inputs and sent to the network with the expectation that more image data could further improve the model performance. Moreover, in order to improve the feature learning ability of the model, we utilize the residual block [28] after

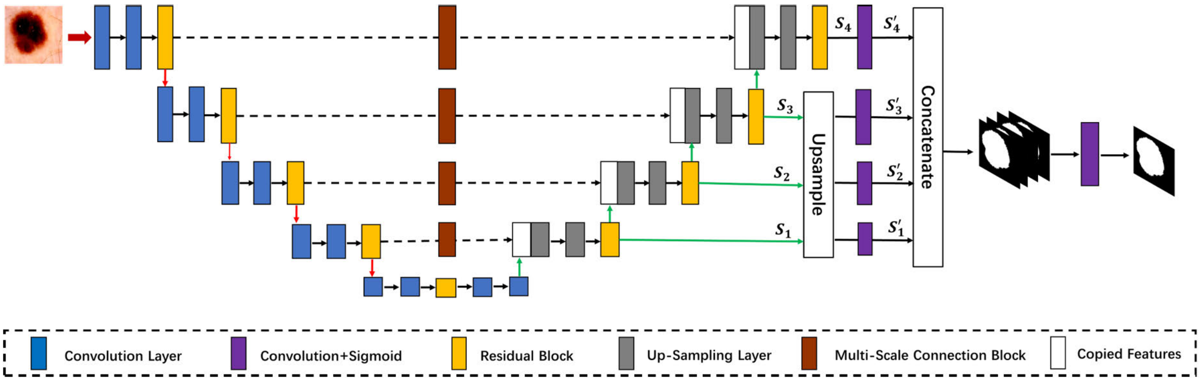


FIGURE 1. The main architecture of our proposed DSM-Network, which achieves satisfied skin cancer segmentation result by utilizing the side-output layers of the network to aggregate information from shallow&deep layers (i.e., S'_1, S'_2, S'_3 and S'_4), and designing multi-scale connection blocks to handle variety of cancer sizes' changes. The final predicted mask is obtained by sending the concatenated segmentation masks from different layers to a convolution+sigmoid layer.

each convolution operation as one basic unit of the network (i.e., the orange blocks in Fig. 1). At the different layer of the network, in order to handle a variety of cancer sizes' changes, a multi-scale connection block is designed to concatenate the discriminative features from different scales. The detailed structure of the MSC block can be referred to Fig. 3. Furthermore, we use a deep supervision learning with side-output layers to aggregate complementary information from various layers (i.e., layers connected to the upsampling and the concatenation blocks in Fig. 1). The final output of the decoder stage is then fed into a 1 × 1 convolution layer with the sigmoid activation to produce the final predicted feature map (as the purple block shows in Fig. 1.). After that, for the predicted mask of each input image (we use three images as inputs: one original image, and two enhanced images), a contour enhancement operation by CRF is performed, and the final predicted mask is obtained by fusing three refined masks equally.

A. IMAGE PREPROCESSING AND CONTRAST ENHANCEMENT

The dense hairs usually cover the skin cancer regions of the images (Fig. 2 (a)), which could hamper the model to obtain accurate segmentation results. Thus, before training the network, we first apply a morphological transformation to remove the effect of the hairs. The operation of morphological transformation is operated by closing operation [29] on three image channels, respectively. It performs with a kernel size of $k \times k$ to close the hair region pixels with the surrounding tissue pixels. Empirically, we set $k = 11$ in all our experiments. The detailed visualization results of the closing operation are shown in Fig. 2 (b). From this column, we can see that the hairs near the lesion regions are clearly removed.

Moreover, in order to increase the contrast between normal tissue and cancer region, after the morphological operation, we perform the unsharp masking [30] and the intensity rescaling [31] to the image, and two more images are generated then as inputs sent into our DSM-Network. The unsharp masking technique comes from a publishing industry process in which an image is sharpened by subtracting a blurred (unsharp) version of the image from itself. The intensity rescaling is adopted from the Scikit-image [31] python package, and it aims to enhance the local contrast of the image. The results by the unsharp masking and the intensity rescaling are shown in Fig. 2 (c) and Fig. 2 (d), respectively.

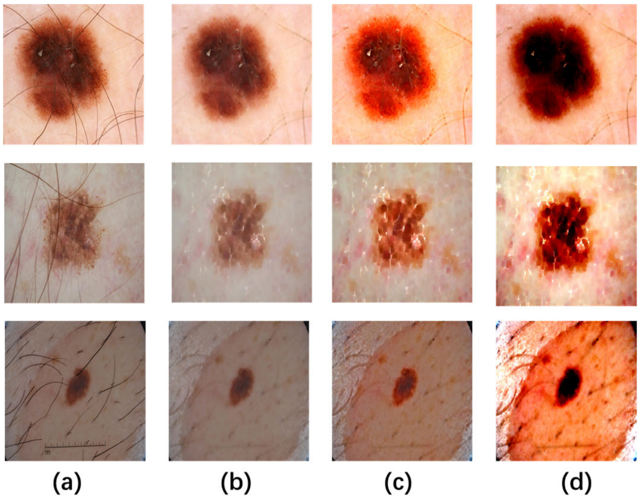


FIGURE 2. Example images of the skin cancer regions before and after image pre-processing and enhancement. (a) original images with hairs; (b) hair removing results after the closing operation on three channels of the original image; (c) image contrast enhancement with unsharp masking after (b); and (d) another image contrast enhancement with the rescale intensity after (b).

B. DEEP SUPERVISED LEARNING

In our network architecture, the deep supervised learning is achieved by adding the side-output layers, which generate the output segmentation map from the early layers. With the deep supervised learning, it could help the network training and alleviate the gradient vanishing. Furthermore, the features of different layers could also contain much more different level

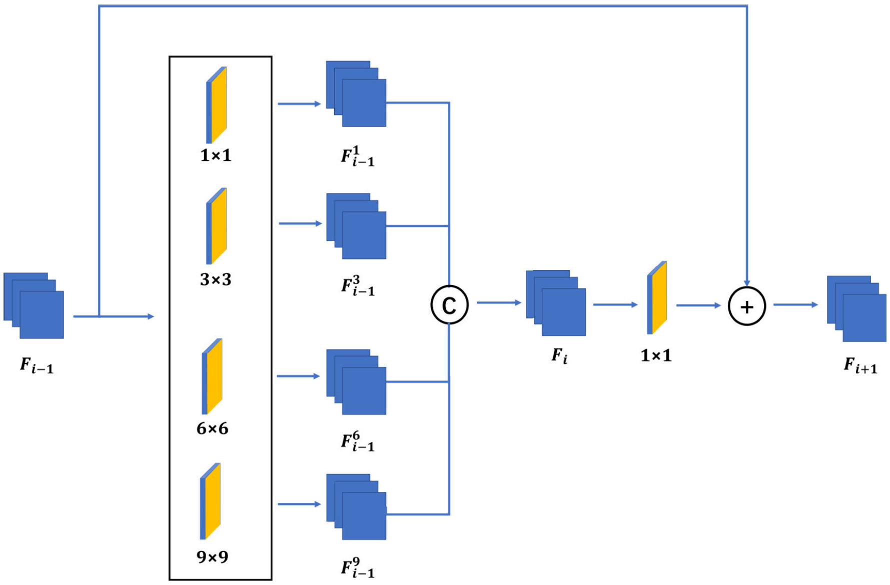


FIGURE 3. The detailed structure of our multi-scale skip connection block. Given an input feature F_{i-1}, four sets of convolutions with size of 1 × 1, 3 × 3, 6 × 6, and 9 × 9 are performed on it and then we concatenate them to be F_i. After using a 1 × 1 convolution to reduce the dimension of F_i, the output feature map F_{i+1} is acquired by employing a residual learning between F_{i-1} and F_i.

features which help the network achieve higher performance. The detailed deep supervised learning structure is illustrated in Fig. 1. Let denote the output feature map of each side-output layer as S_i where $i \in \{1, 2, 3, 4\}$. For S_1, S_2, S_3, we first upsample each feature map to the original image size. For S_4, we keep it untouched since it has the same size as the original image. Then a 1 × 1 convolution layer with a sigmoid activation is applied to S_i to generate the predicted activation map S'_i. After that, these predicted activation maps are concatenated to gain the stacked predicted activation map S_t:

$$S_t = [S'_1, S'_2, S'_3, S'_4]. \tag{1}$$

Finally, we apply an extra 1 × 1 convolution layer with sigmoid activation to S_t to fuse all the predicted activation maps and gain the final predicted result. Based on our deep supervision learning, the influence of the gradient vanishing could be alleviated, and the network could also extract more fine detail representations to further boost the performance of the model.

C. MULTI-SCALE CONNECTION BLOCK

The convolution operation learns the local features from the input image through the filters automatically. Different sizes of filter kernels could provide diverse multi-scale features. For the small size kernel at lower layers, it tends to learn the detailed low-level information of the images. Meanwhile, for the large size kernel at high layers, it could extract the high-level or semantic and usually big representations. At last, the skip connection of U-Net is to alleviate the gradient vanishing and provide more spatial information of the previous layers to the deep layers. Inspired by those findings, we design an MSC block to enhance the feature learning ability of the segmentation network. The detailed structure of the MSC block is illustrated in Fig. 3. Consider F_{i-1} as the input feature from the encoder stage layer. Four sets of convolutions with size of 1 × 1, 3 × 3, 6 × 6, and 9 × 9 are performed, respectively. For 1 × 1 and 3 × 3 convolutions, they aim to extract the features with a small receptive field such that tiny information can be captured. For 6 × 6 and 9 × 9 convolutions, they can learn representations with large receptive fields which are suitable for the large representations learning. Here, denote the feature maps after the four convolutions as $F^1_{i-1}, F^3_{i-1}, F^6_{i-1}$, and F^9_{i-1}, respectively. After that, we concatenate those four scale features to gain the feature map F_i, and then a 1 × 1 convolution is applied to reduce the dimension of F_i. Finally, the output feature map F_{i+1} is acquired by employing a residual learning between F_{i-1} and F_i. The designed MSC block is placed in the skip connection as one intermediate layer so that more multi-scale features of the encoder stage's layers could be learned.

D. SIDE-OUTPUT WEIGHTED LOSS

In our designed model, we use the combination of the binary cross-entropy and dice loss to train the network. The binary cross-entropy loss can be written as:

$$L_{bce} = -\sum_{i=1}^{N}((y_i ln(p_i)) + (1 - y_i)ln(1 - p_i)), \tag{2}$$

and the dice loss function can be formulated as :

$$L_{di} = \frac{2 * \sum_i^N y_i p_i}{\sum_i^N y_i + \sum_i^N p_i}, \tag{3}$$

where N is the pixel number, y_i is the ground truth label (i.e. 0 or 1) and p_i is the predicted probability. Thus, the combination loss L_c can be written as:

$$L_c = L_{bce} + (1 - L_{di}). \tag{4}$$

Then, the side-output layer loss L_s is given as:

$$L_s = \sum_{m=1}^{M} \beta L_c, \tag{5}$$

where β is the fusion-weight of each side-output layer. We make it as 0.1 empirically, and the values of M and N are 4 identically. Finally, the total loss L_t of the network is calculated by:

$$L_t = \sum_{m=1}^{M} \beta L_c + (1 - N * \beta) L_f, \tag{6}$$

where L_f is the loss from the final fused activation map, and it can be similarly formulated as:

$$L_f = L_{bce} + (1 - L_{di}). \tag{7}$$

E. CONTOUR REFINEMENT WITH CRF

Although our proposed model shows competitive capability in generating high-quality probability maps, the final segmentation can sometimes be ambiguous and rough on edges of the regions, especially when the dense hair on the skin covers the lesion contour. Besides, since the deep convolutional networks usually have a very large receptive field, the produced probability maps are too coarse for pixel-level skin lesion segmentation. Hence, in our model, CRF is used for refining the imprecise lesion contour through building non-local pixel relations, yielding accurate semantic segmentation results. The pipeline of the contour refinement by CRF is shown in Fig. 4. For each predicted mask, CRF is performed to refine the coarse boundary and imperfect segmentation result. Each predicted mask by DSM-Network for the original image and two enhanced images (unsharp masking image and the intensity re-scaled image) are refined by CRF respectively, and then the three refined results are averaged to obtain the final segmentation mask.

Here, we use $x \in \mathcal{X}$ denoting the image to be refined while $y \in \mathcal{Y}$ is corresponding labeled results that include the label configuration of each pixel of x in the CRF graph. The objective is to solve the following energy function:

$$E(y, x) = \sum_{i} E_i(y_i, x_i) + \sum_{ij} E_{ij}(y_i, y_j, x_i, x_j). \tag{8}$$

Here, the first term is the unary potential, and it calculates for each individual pixel. Thus, the $E_i(y_i, x_i)$ can be defined as:

$$E_i(y_i, x_i) = -logp(y_i, x_i), \tag{9}$$

where $p(y_i, x_i)$ denotes the prediction probability at pixel x_i. The second term of Eq. 8 is a pairwise potential. It considers the compatibility between each pair of adjacent pixels. The $E_{ij}(y_i, y_j, x_i, x_j)$ can be formulated as:

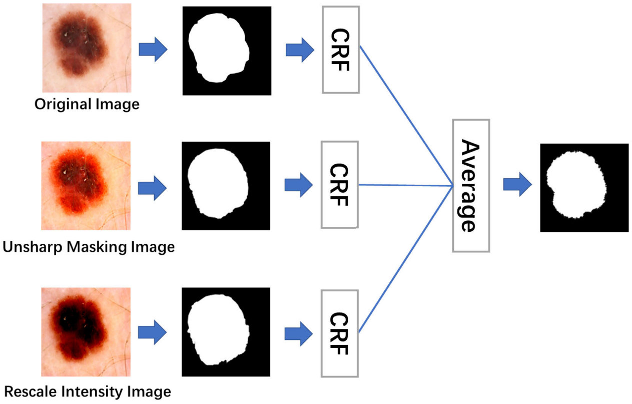


FIGURE 4. The pipeline of the contour refinement by CRF and the procedure to obtain the final segmentation mask. Each predicted mask by DSM-Network for the original image and two enhanced images are refined by CRF respectively, and then the three refined results are averaged to obtain the final segmentation mask.

$$E_{ij}(y_i, y_j, x_i, x_j) = \phi(y_i, y_j) f(y_i, y_j). \tag{10}$$

Specifically, the second term of the above equation can be further defined as: $\phi(y_i, y_j)$ is 1 if $y_i \neq y_j$, and 0 otherwise. And $f(y_i, y_j)$ can be defined as:

$$f(y_i, y_j) = exp\{-\frac{|x_i - x_j|^2}{2\sigma_\alpha^2}\} + \lambda exp\{-\frac{|p_i - p_j|^2}{2\sigma_\beta^2}\}, \tag{11}$$

where σ_α and σ_β are the hyper parameters to adjust the Gaussian kernels, and λ is used to weight between the RGB color space and the position space. In particular, in Eq. 8, the first term constrains that the connected pixels with similar color appearance tend to be assigned with the same category label, and the second term constrains spatial proximity while enforcing smoothness. The optimal solution of $E(y, x)$ can be obtained by using efficient mean field approximation inference [34].

III. EXPERIMENTS

A. DATASETS

We evaluate the proposed network on two benchmark datasets:

ISBI 2017 dataset To validate the performance of our designed model, we conduct experiments on ISBI 2017 dataset [32], in which images were captured from different clinical centers. Each image is paired with the expert manual tracing of the skin lesion boundaries for segmentation task. During the validation process, 2000 data samples are used for training, 150, and 600 samples for validation and testing, respectively.

PH2 dataset PH2 public dataset [33] contains 200 dermoscopy images with the resolution of 768 × 560 pixels, including 160 nevus, and 40 melanomas. All the dermoscopy images are gained from the Pedro Hispano

TABLE 1. The effectiveness of residual learning (RL) on two employed datasets.

Dataset	ISBI 2017				PH2			
Method	AC (%)	SE (%)	JA (%)	DI (%)	AC (%)	SE (%)	JA (%)	DI (%)
No RL + U-Net	90.1	81.3	75.6	83.2	91.3	86.6	87.1	91.0
RL + U-Net	92.1	82.4	76.2	84.5	91.7	87.2	87.5	91.2
No RL + DSM-Network	93.7	84.6	77.2	86.7	92.3	87.6	88.4	91.5
RL + DSM-Network	**94.3**	**85.9**	**78.5**	**87.5**	**93.1**	**88.9**	**89.1**	**92.0**

Hospital, the research group of the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal. The lesion segmentation boundary mask was annotated by professional experts.

B. IMPLEMENTATION DETAILS

The DSM-Network is achieved based on the Tensorflow deep learning framework with an NVIDIA GTX 1080 graphic processing unit (GPU). The initial learning rate of the model is 0.001, and we reduce it by a factor 0.1 dynamically. The optimization of our model is Adam optimizer. During the training, we input the image with a size of 256 × 256. The random flipping, rotation, whitening, and two extra generated images are used as the data augmentation approach to further improve the performance of the model.

C. EVALUATION METRICS

In this paper, we use accuracy (AC), sensitivity (SE), jaccard index (JA), and dice coefficient (DI) as the basic metrics to evaluate our designed model. Consider TP as the true positive, FP as the false positive, TN as the true negative, FN as the false negative. The evaluation metrics are defined as follows:

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

$$SE = \frac{TP}{TP + FN}, \tag{13}$$

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

$$DI = \frac{2 * TP}{2 * TP + FP + FN}. \tag{15}$$

D. THE EFFECTIVENESS OF RESIDUAL BLOCK

Since the deeper network could extract more high-level and abstract features from the original image. Nevertheless, directly adding the layers may result in the gradient vanishing problem, and make the network difficult to train. Thus, in our model, we add the residual block (RL) [28] to resolve the gradient vanishing problem effectively. In this section, we conduct experiments to verify the effectiveness of this module. Especially, we also use U-Net as the baseline model for better comparison.

The detailed comparison result is illustrated in Table 1. The comparison result demonstrates that with the residual block, the model could gain better segmentation performance. Specifically, on ISBI 2017 dataset, the best result is achieved with 94.3% AC, 85.9% SE, 78.5% JA and 87.5% DI, respectively. While on PH2 dataset, the performance is 93.1% AC, 88.9% SE, 89.1% JA and 92.0% DI, separately. The reason may lie on that with the residual block the proposed model can deepen the network thus is able to extract more high-level and abstract features for accurate segmentation result.

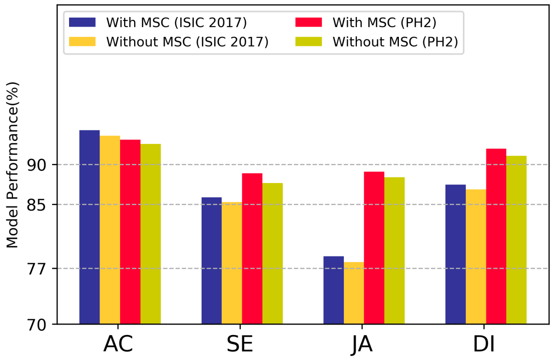


FIGURE 5. Performance comparisons of multi-scale connection block on ISBI 2017 and PH2 datasets, respectively.

E. THE EFFECTIVENESS OF MULTI-SCALE CONNECTION BLOCK

The skip connection is to alleviate the gradient vanishing and provide more spatial information of the previous layers. Inspired by that, we design a multi-scale connection block to enhance the model learning more scale-relevant features. Different comparison results on those two datasets are illustrated in Fig. 5. From the result, we can see that with the MSC module, the segmentation performance of our proposed DSM-Network on those two datasets can be further improved. On ISBI 2017 dataset, AC and JA are improved by 0.7%, and SE and DI are improved by 0.6%, respectively. On PH2 dataset, after adding MSC block, the performance of our model is enhanced AC, SE, JA, and DI by 0.5%, 1.2%, 0.7%, and 0.9%, respectively. That further validates that discriminative multi-scale features adding from the previous layers could be an efficient way to boost the segmentation performance of the proposed model.

F. THE EFFECTIVENESS OF SIDE-OUTPUT LAYERS

In our designed model, the deep supervised learning is achieved by adding the side-output layers, which generate

TABLE 2. Comparisons with state-of-the-arts on ISBI 2017 dataset.

Method	AC (%)	SE (%)	JA (%)	DI (%)
U-Net [20]	90.1	67.2	61.6	76.3
SegNet [21]	91.8	80.1	69.6	82.1
CDNN [11]	93.4	82.5	76.5	84.9
FrCN [22]	94.0	85.4	77.1	87.0
DDN [14]	93.9	82.5	76.5	86.6
SLS [16]	93.6	81.6	78.2	**87.8**
DSM-Network	**94.3**	**85.9**	**78.5**	87.5

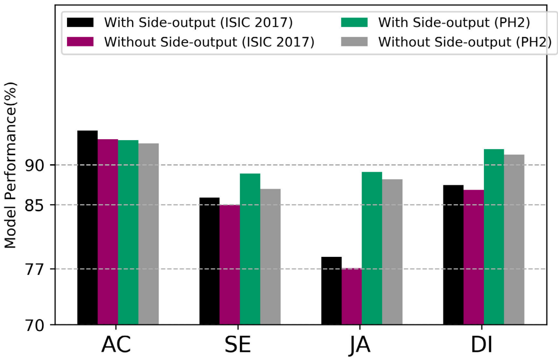


FIGURE 6. The results of our DSM-Network with or without the side-output layers on ISBI 2017 and PH2 datasets, respectively.

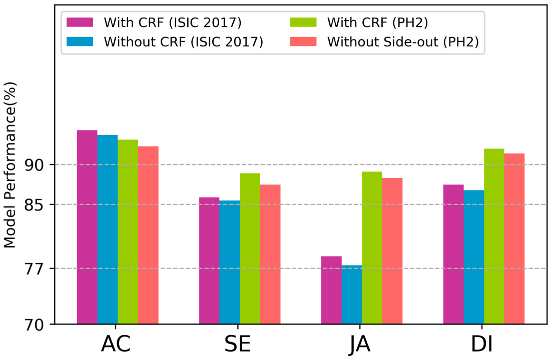


FIGURE 7. The comparison results of our DSM-Network with or without the CRF enhancement on ISBI 2017 and PH2 datasets, respectively.

the output segmentation map from the early layers. In this section, we compare the results with or without the deep side-output layers to explore the effectiveness of this design. The detailed comparison result is shown in Fig. 6. It demonstrates that adding the side-output layers could efficiently improve the overall performance of the model, especially in SE and JA. For ISBI 2017 dataset, the AC, SE, JA, and DI are improved by 1.1%, 0.9%, 1.4% and 0.6%, respectively. For PH2 dataset, the performance by adding the side-output layers boosts AC, SE, JA and DI by 0.4%, 1.9%, 0.9% and 0.7%, respectively. All these illustrate that adding the side-output layers could help the network learn more level-wise features, which are complementary information to further boost the segmentation performance.

G. THE EFFECTIVENESS OF CRF REFINEMENT

In our model, we use CRF to refine the imprecise lesion contour through building non-local pixel relations, yielding accurate semantic segmentation results. We first use three predicted probability maps (original image, unsharp masking image, re-scaled intensity image) with CRF operation and then ensemble them with an average mode. The final comparison result of the CRF enhancement operation is presented in Fig. 7. On ISBI 2017 dataset, AC and JA are improved by 0.6% and 0.4%, and SE and DI are improved by 1.1% and 0.7%, respectively. On PH2 dataset, with the CRF enhancement operation, AC, SE, JA, and DI are improved by 0.8%, 1.4%, 0.8%, and 0.6%, separately. The improvement of the segmentation performance further validates the effectiveness of CRF as post-processing of our method.

H. COMPARISONS WITH STATE-OF-THE-ART METHODS

In this section, we compare our model with state-of-the-art methods on the employed two datasets. The detailed results are illustrated in Table 2.

On ISBI 2017 dataset, the traditional segmentation model U-Net [20] and SegNet [21] are trained as baseline models for comparison. Then, a enhanced convolutional-deconvolutional network [11], a full resolution convolutional networks (FrCN) [22], a dense deconvolutional network (DDN) [14], and a robust Skin lesion segmentation(SLS) model [16] are compared separately. The comparison result shows that our model could achieve competitive results in AC, SE, and JA. And our DI achieves a comparative result which is the second-best among all the compared methods.

On PH2 dataset, we compared the performance of our proposed method with state-of-the-art methods and the results are listed in Table 3. The traditional segmentation method Adaptive thresholding (AT) [23] and a simple linear iterative clustering (SLIC) [24] method are trained as baseline models for comparisons. Then, other methods compared are Multi-Scale Segmentation (MSS) [25], Level Set Active

TABLE 3. Comparisons with state-of-the-arts on PH2 dataset. "-" denotes no corresponding result provided by the method on this dataset.

Method	AC (%)	SE (%)	JA (%)	DI (%)
AT [23]	-	-	72.4	80.4
SLIC [24]	90.4	**91.0**	-	-
MSS [25]	-	-	76.0	86.1
LSAC [26]	-	-	76.3	83.5
DermoNet [27]	-	-	85.3	91.5
Deep FCN [13]	-	-	81.5	91.5
DSM-Network	**93.1**	88.9	**89.1**	**92.0**

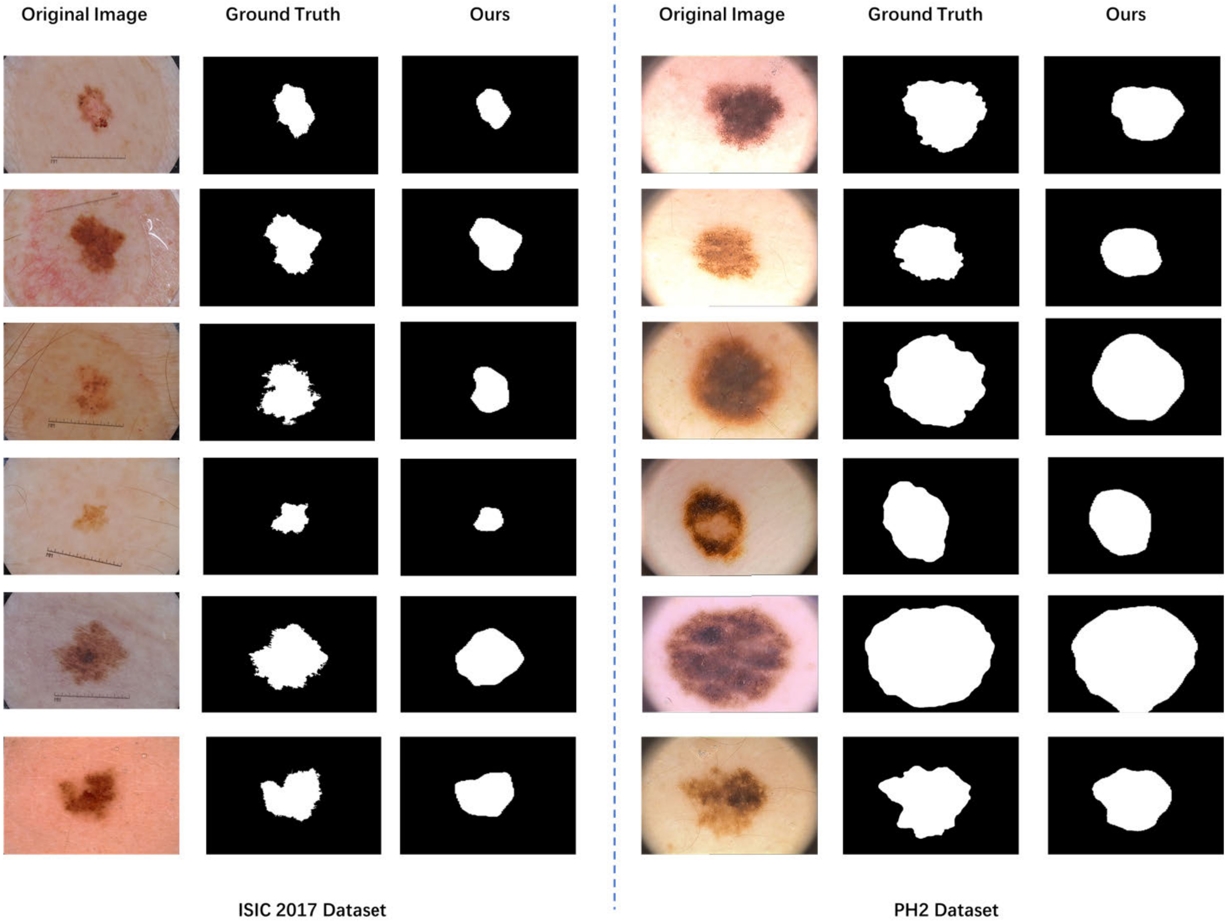


FIGURE 8. Sampled segmentation results of our method compared with human labeled ground truth.

Contours (LSAC) [26], DermoNet [27] and Deep FCN [13]. It can be seen from the results that our method performs the best on dataset PH2 and is better than the existing segmentation methods in the literature.

I. QUALITATIVE ANALYSIS

In this section, we conduct the qualitative analysis of our DSM-Network, the detailed visualization results are illustrated in Fig. 8. The result shows that the overall segmentation performance on PH2 dataset is better than that on ISBI 2017 dataset. The reason may be that images from ISBI 2017 dataset usually contain more complex and subtle information, which is difficult for the network to learn. Meanwhile, the segmentation performance of the contextual edge information is not as good as expected. That's due to the missing features by the successive pooling layers and the small difference between the cancer region and the normal tissue region.

IV. CONCLUSION

In this paper, we propose a novel DSM-Network for skin segmentation, which uses the deep supervision feature learning with the side-output layers to learn different level features, and a multi-scale connection block is designed to improve the ability of extracting scale-relevant features. Furthermore, a post-processing of contour enhancement strategy is adopted by a CRF to obtain better segmentation performance. The extensive comparison results on two public datasets demonstrate that our model could achieve state-of-the-art performance. In future work, we will try to re-design the network with the post-processing by an end-to-end training mode.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA Cancer J. Clin.*, vol. 66, no. 1, pp. 7–30, Jan. 2016.

[2] Q. Abbas, M. E. Celebi, I. F. García, and M. Rashid, "Lesion border detection in dermoscopy images using dynamic programming," *Skin Res. Technol.*, vol. 17, no. 1, pp. 91–100, Feb. 2011.

[3] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.*, vol. 19, no. 1, pp. e252–e258, Feb. 2013.

[4] H. Zhou, X. Li, G. Schaefer, M. E. Celebi, and P. Miller, "Mean shift based gradient vector flow for image segmentation," *Comput. Vis. Image Understand.*, vol. 117, no. 9, pp. 1004–1016, 2013.

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

[6] G. Schaefer, B. Krawczyk, C. Me, and H. Iyatomi, "An ensemble classification approach for melanoma diagnosis," *Memetic Comput.*, vol. 6, no. 4, pp. 233–240, Dec. 2014.

[7] R. J. Stanley, W. V. Stoecker, and R. H. Moss, "A relative color approach to color discrimination for malignant melanoma detection in dermoscopy images," *Skin Res. Technol.*, vol. 13, no. 1, pp. 62–72, Feb. 2007.

[8] M. Messadi, A. Bessaid, and A. Taleb-Ahmed, "Extraction of specific parameters for skin Tumour classification," *J. Med. Eng. Technol.*, vol. 33, no. 4, pp. 288–295, Jul. 2009.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[10] A. R. Sadri, M. Zekri, S. Sadri, N. Gheissari, M. Mokhtari, and F. Kolahdouzan, "Segmentation of dermoscopy images using wavelet networks," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1134–1141, Apr. 2013.

[11] Y. Yuan and Y.-C. Lo, "Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 519–526, Mar. 2019.

[12] F. Xie and A. C. Bovik, "Automatic segmentation of dermoscopy images using self-generating neural networks seeded by genetic algorithm," *Pattern Recognit.*, vol. 46, no. 3, pp. 1012–1019, Mar. 2013.

[13] Y. Yuan, M. Chao, and Y.-C. Lo, "Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876–1886, Sep. 2017.

[14] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, S. Chen, T. Wang, and B. Lei, "Dense deconvolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 527–537, Mar. 2019.

[15] Z. Mirikharaji and G. Hamarneh, "Star shape prior in fully convolutional networks for skin lesion segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2018.

[16] M. K. Sarker, H. A. Rashwan, F. Akram, S. F. Banu, A. Saleh, V. K. Singh, F. U. H. Chowdhury, S. Abdulwahab, S. Romani, P. Radeva, and D. Puig, "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2018.

[17] Y. Li and L. Shen, "Skin lesion analysis towards melanoma detection using deep learning network," *Sensors*, vol. 18, no. 2, p. 556, Feb. 2018.

[18] L. Bi, J. Kim, E. Ahn, and D. Feng, "Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks," Mar. 2017, *arXiv:1703.04197*. [Online]. Available: https://arxiv.org/abs/1703.04197

[19] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba, A. C. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Develop.*, vol. 61, nos. 4–5, pp/ 5:1–5:15, Jul./Sep. 2017.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[22] M. A. Al-masni, M. A. Al-antari, M.-T. Choi, S.-M. Han, and T.-S. Kim, "Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks," *Comput. Methods Programs Biomed.*, vol. 162, pp. 221–231, Aug. 2018.

[23] M. Silveira, J. C. Nascimento, J. S. Marques, A. R. S. Marcal, T. Mendonca, S. Yamauchi, J. Maeda, and J. Rozeira, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 1, pp. 35–45, Feb. 2009.

[24] D. Patiño, J. Avendaño, and J. W. Branch, "Automatic skin lesion segmentation on dermoscopic images by the means of superpixel merging," in *Medical Image Computing and Computer Assisted Intervention*. Cham, Switzerland: Springer, 2018.

[25] Bozorgtabar, Behzad, Mani Abedini, and Rahil Garnavi, "Sparse coding based skin lesion segmentation using dynamic rule-based refinement," in *Machine Learning in Medical Imaging*. Cham, Switzerland: Springer, 2016.

[26] C. Li, C.-Y. Kao, J. C. Gore, and Z. Ding, "Minimization of region-scalable fitting energy for image segmentation," *IEEE Trans. Image Process.*, vol. 17, no. 10, pp. 1940–1949, Oct. 2008.

[27] S. B. Salimi, S. Bozorgtabar, P. Schmid-Saugeon, H. K. Ekenel, M. S. Rad, and J. P. Thiran, "DermoNet: Densely linked convolutional neural network for efficient skin lesion segmentation," Ecole Polytechnique Federale Lausanne, Lausanne, Switzerland, Tech. Rep., 2018.

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

[29] J. Gil and R. Kimmel, "Efficient dilation, erosion, opening, and closing algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 12, pp. 1606–1617, Dec. 2002.

[30] Bilcu, Radu Ciprian, and Markku Vehvilainen, "Constrained unsharp masking for image enhancement," in *Image and Signal Processing*. Berlin, Germany: Springer, 2008.

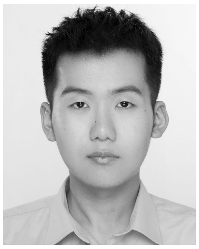
[31] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu, "Scikit-image: Image processing in Python," *Peer J.*, vol. 2, p. e453, 2014.

[32] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "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)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 168–172.

[33] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH^2- A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5437–5440.

[34] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFS with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

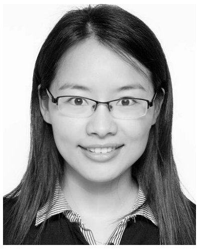
GUOKAI ZHANG is currently pursuing the Ph.D. degree with the College of Software Engineering, Tongji University, Shanghai, China. His current research interests include deep learning, object detection, and medical image analysis.



XIAOANG SHEN is currently pursuing the B.Sc. degree with Tongji University, Shanghai, China. His research interests include machine learning, image processing, reconstruction, and 3-D visualization.

SIRUI CHEN is currently pursuing the B.Sc. degree with Tongji University, Shanghai, China. Her research interests include machine learning, image processing, and reconstruction.

LIPENG LIANG is currently pursuing the master's degree with the School of Software Engineering, Tongji University. Her majors are software engineering and project management. She is interested in computer vision, medical image processing, software developing, and the Internet of Things.



YE LUO received the M.S.E. degree in signal and information processing from Anhui University, Hefei, China, in 2008, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2014. She is currently an Assistant Professor with Tongji University. Her research interests include computer vision, machine learning, content/perceptual-based video analytics, and medical image processing.

JIE YU received the master's degree from Qingdao University, China. He is currently a Full Computer Technology Engineer with Qingdao Central Hospital, Qingdao, China. His research interests include bioinformatics and big data analytics.

JIANWEI LU received the Ph.D. from the Department of Computer Science, University of Southern California.

He is currently a Professor with the School of Software Engineering Advanced, Institute of Translational Medicine, Tongji University. His research interest includes vision science.