IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 28, 2020, accepted August 11, 2020, date of publication August 24, 2020, date of current version September 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019183

Accurate Pixel-Wise Skin Segmentation Using Shallow Fully Convolutional Neural Network

KOMAL MINHAS[1], TARIQ M. KHAN[2], (Member, IEEE), MUHAMMAD ARSALAN[3], (Member, IEEE), SYED SAUD NAQVI[2], MANSOOR AHMED[1], HAROON AHMED KHAN[2], (Member, IEEE), MUHAMMAD ADNAN HAIDER[3], AND ABDUL HASEEB[4]

[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan
[2]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad 45550, Pakistan
[3]Division of Electronics and Electrical Engineering, Dongguk University, Seoul 100-715, South Korea
[4]Department of Electrical Engineering, Institute of Space Technology (IST), Islamabad 44000, Pakistan

Corresponding authors: Komal Minhas (komalminhas7@gmail.com) and Tariq M. Khan (tariq045@gmail.com)

ABSTRACT Skin segmentation plays an important role in human activity recognition, video surveillance, hand gesture identification, face detection, human tracking and robotic surgery. The accurate segmentation of the skin is necessary to recognize the human activity. Segmentation of skin is easy to realize in ideal situations because of similar backgrounds. But it becomes complicated because of presence of skin-like pixels, background illuminations, and certain changes in environment. These problems are addressed by incorporating preprocessing stages in current studies, but this raises the total cost of the system. However, there are some limitations associated with these methods in terms of accuracy and processing speed. In this work, we propose a skin semantic segmentation network (SSS-Net) that is able to capture the multi-scale contextual information and refines the segmentation results especially along object boundaries. Moreover our network helps to reduce the cost of the preprocessing as well. We have performed experiments on the five open datasets of human activity recognition for the segmentation of skin. Experimental results show SSS-Net outperforms the state-of-the-art methods in skin segmentation in terms of accuracies.

INDEX TERMS Skin segmentation, semantic segmentation, low-level semantic information, deepLabv3+.

I. INTRODUCTION

Skin segmentation aims to detect the region of a human skin in an image. It is one of the important tasks which works as a step for pre-processing in various systems and applications, such as hand gesture analysis, face recognition, face tracking and detection, content based image retrieval, etc. [1]. Skin detection is a process of identifying the pixels of a given image that correspond to human skin. Skin detection is also very helpful to humans while performing complex tasks through human computer interaction. As in case of hand gesture recognition, it provides help in recognizing certain actions [2]. In recent years, with the advancements in deep neural networks, the networks used for other detection tasks have been adapted as skin detection methods as well [3], [4].

Recognition of human activity is an important area and has received a great deal of attention due to the growing demands of many applications. These include, but are not limited to identification of individual activity, interaction between multiple persons and analysis of crowd behavior. Recognition of human posture in single person activity helps detect the nature of the activity. Nonetheless, these task are inherently challenging since human poses vary enormously. These problems get compounded when the activity of multiple subjects is involved. This is an area of active interest, for instance crowd monitoring to detect antisocial behaviour is being tested and deployed extensively [5]. Activity monitoring is also being used in sports, for instance recognition of the actions of players during a tennis game [6]. Human activity recognition depends quite critically on accurate skin segmentation [7]. This challenge is compounded by the structural variations within a single human's limbs and body parts, making consistent skin segmentation difficult. The problem becomes significantly complex with multiple subjects in a frame. Therefore traditional machine learning algorithms fail to detect multiple features at one time

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

where skin segmentation is being used. The Skin Semantic Segmentation Network (SSS-Net) presented in this research deals with the limitations of skin segmentation innovatively by capturing multi-scale contextual information and refining the segmentation results especially along object boundaries.

In this paper, SSS-Net is used for skin segmentation tasks for the semantic labeling of pixels in a pixel-wise classification framework. The contributions of this work are:

1) The task of skin segmentation is modeled as a semantic pixel-wise segmentation problem. For this reason, a SSS-Net with reduced tunable hyper parameters is considered. We believe this work will help bridge the gap between skin segmentation and semantic segmentation;
2) Low-level semantic information is preserved and the preservation of edge information results in robust detection of skin information;
3) The proposed method is robust to skin detection;
4) A much smaller (in terms of tunable parameters) deep neural network is proposed for skin segmentation that does not require additional pre-processing steps;
5) Low computational time overhead during inference in both train and test stages.

This paper is organized as follows; In section 2, relevant literature is discussed. The methodology is reviewed in section 3. Section 4 presents the experimental results. This is followed by a discussion of the proposed network in section 5. Section 6 concludes the findings of this research.

II. RELATED WORK

Skin detection is being used extensively for a variety of applications in image processing and visual computing. Many studies based on skin detection use a variety of different methods. These methods can be divided into different categories, i.e. thresholding, traditional handcrafted features and deep neural network. To separate the skin and no-skin areas, different image channels are used by these procedures. In [8] skin and non-skin areas are detected by using two detectors that are based on color channels and thresholding. Thresholding and these channels are dynamically selected and are based on agreement maximization framework. Thresholding concentrate on selecting a certain region in color spaces, thus if a pixel belongs to that region it will be treated as skin. But there are several challenges involved in detecting skin and non-skin pixels. This is primarily due to the similarity of background objects with the color of the skin due to various reasons, making skin detection a very challenging and difficult task. Reference [1] proposed a method where an eye detector has been shown to improve the accuracy of skin detection regardless of variations in illumination and ethnicity. In [9], a method is introduced for handling skin like pixels in the background. Proposed method significantly helps in reducing the error in the detection thus reducing the false detection of skin color. Interested readers are referred to [24] and [25] for more information for the selection and weighting. For the detection of skin color, multi-color spaces have been introduced for the skin color model. For instance, [11] performed dynamic skin detection using multi-color space instead of using the single color space. The proposed method improves the precision rate as well as reducing the error in skin color detection. As skin detection is an important step at the time of pre-processing of images, [12] proposed a method that used a clustering technique which makes clusters of similar pixels in the image. The proposed method is able to produce good results with effective skin detection of human images irrespective of the ethnicity. Moreover the proposed method performs well with the illumination changes as well. Reference [26] proposed a network for improving the segmentation results specially in terms of accuracy for the large scale objects. The network uses several scales that enable it to achieve the detailed information with increased sensitivity. Reference [27] proposed an algorithm for the object detection that is effective in detecting the small areas as well as the occluded ones using different scales.

Skin detection plays a very important role in various medical application of visual computing systems such as those used for the detection of certain diseases related to skin. Reference [13] proposed an approach for the detection of skin regions in human images using the specific color space. The proposed approach provides promising results related to detection and shows good detection rate. Reference [14] proposed a method for the classification of human skin pixels under the varying illumination conditions and shows good results. Reference [16] presents the comparative study of the two color spaces for the detection of human skin color and selected the specific threshold for detecting skin color to evaluate these color spaces. The overlapping of skin and non-skin pixels is one of the constraints in detection of human skin. To improve the accuracy in the skin detection process, [18] proposed a method based on color space that includes the texture features of human skin. In the field of biometric security, palm-prints are being used extensively over other methods that depend on accurate skin detection. Reference [28] presented a method for the segmentation of palm print to achieve accurate and improved detection compared to existing methods. Reference [29] solved a problem related to technical issue involved in the non-contact palm-print system by developing a system on personal computer. Reference [30] also worked for developing a system for the pre-processing of palm-print in the contact-less scenario.

In order to handle the problem due to changes in illumination causing similarity of background color to the skin color, [19] proposed a method that used combination of two techniques which improves skin detection performance. To improve the skin detection a method is proposed by [31], which uses a neural network for the detection of skin and body. However current methods that are based on machine learning or traditional neural networks have some limitation regarding performance under certain illumination conditions. To overcome this problem deep learning based methods have been introduced. Using skin segmentation, the tasks like hand detection is performed which may be used in interpreting

TABLE 1. SSS-Net comparison with existing methods of skin segmentation.

Author	Method	Strength	Weakness
SanMiguel et al. [8]	Color channel threshold	A convenient way to estimate pixels in the image	Changes in illumination could lead to false results
Rahmat et al. [1]	WSPM	Method is helpful in reducing the error in detection	Errors related to skin tone can be caused by threshold
Chakraborty et al. [9]	Color information model	Can perform good in indoor and outdoor environments	Result will be affected by the involvement of different person skin pixels
Zhang et al. [10]	Multi-color-based threshold	Multi-color space can be included with the dynamic detection of skin	Thresholding can cause a problem with skin like background pixels
Osman et al. [11]	K-means clustering	Non skin pixels detection is reduced by removing the background pixels	Step needed for the refined detection results.
Buza et al. [12]	Explicit skin model	Uses union of color space helps reducing error	Changes in illumination could lead to false results
Hassan et al. [13]	Gaussian skin color model	Temperature of light source is used to produce skin sample	Needed to refine the detection
Ueng et al. [14]	Binarization with morphology and Sobel edge detection	A Simple image processing technique	Results for edge detection are not satisfactory for the cluttered background
Bhame et al. [15]	Histogram-based technique	A simple histogram technique is used for threshold setting	Thresholds needs to be adjusted
Shaik et al. [16]	Combined skin color and texture properties	Proposed method improves accuracy	Not effective for skin detection in multiple people interaction
Huang et al. [17]	MLP and K-means clustering with statistical features	More reliable approach than handcrafted schemes for feature extraction	Time consuming as it is based on patch based scheme
Al-Mohair et al. [18]	Multi-agent learning, Bayesian method, and neural network		Data is dependent on particular factor
Zaidan et al. [19]	Two-stage deep learning approach	Very positive results has been seen by hand detection method	Time consuming as it is based on patch based scheme.
Roy et al. [20]	Network-in-network approach	High accuracy is observed by the modified inception module	Connections and the network itself is quite complex.
Kim et al. [21]	Domain adapted CNN	Proposed method performs good	Deep Neural network is used
Dourado et al. [22]	SegNet based method	Can improve results using the trained morphology	Weight initialization is required
Lumini et al. [23]	OR-Skip-Net	Method is able to perform without loss of information and less training time	Network is very demanding in terms of training.
Arsalan et al. [7]			

the sign language. Reference [20] developed a technique for detecting the hand in the human images with the cluttered environment and they performed this task by using deep learning approaches. Reference [21] introduced the deep learning network with reduced number of parameters for the skin detection, and has produced good results as compared to state of the art networks. References [7], [22], [23] have introduced certain schemes based on deep learning to handle such problems that leads to better skin detection and segmentation while reducing the error rates. In this work, we propose skin semantic segmentation network (SSS-Net) for skin segmentation that eliminates the pre-processing steps and uses a reduced number of parameters compared to existing solutions. SSS-Net is able to capture the multi-scale contextual information and provides results with sharper object boundaries.

III. NETWORK ARCHITECTURE

A. DESIGNING AND LEARNING

The challenges of skin segmentation are cluttered background, objects at multiple scales and small and deformable objects. In the work presented, we have treated skin segmentation as a semantic segmentation problem due to shared challenges in both. Therefore, for skin segmentation, we adapted the well-known DeepLabv3+ architecture which is state-of-the-art in semantic segmentation. The inherent nature of the DeepLabv3+ architecture is tailored to scenes with cluttered background. The image is first subjected to residual learning to tackle few challenges of skin segmentation including color similarity of foreground and background, skin reflectance variations due to illumination conditions. We start by shredding off residual blocks by keeping only four residual blocks in the proposed SSS-Net. The underlying reasons for reduced residual blocks are two-fold: First, to preserve the details of small that are otherwise lost in the repetitive convolution process. Reducing the number of layers to preserve object and semantic information for semantic segmentation is supported by previous works [37]. Second, to reduce the number of parameters and the computational load. In order to preserve feature information, downscaling is not introduced in the residual learning process. Contextual information is of utmost importance in skin segmentation due to the deformable nature and small extent of skin regions. Therefore, we employ atrous spatial pyramid pooling (ASPP) to capture image context at multiple scales. Due to its intuitive local feature processing and subsequent fusion, ASPP has proven to be robust in detecting objects at multiple scales as well as efficient in mitigating background clutter in object recognition scenarios [38]. Spatial pyramids have successfully been employed for dense prediction tasks [39], [40] owing to the multi-scale contextual information contained in them. We note that the ASPP if not carefully designed, can miss small skin regions. Therefore, we experimented with different dilation rates such that our filters simultaneously cater small and large skin regions alike. We found in our experiments that limiting the number of residual blocks to four preserves vital semantic

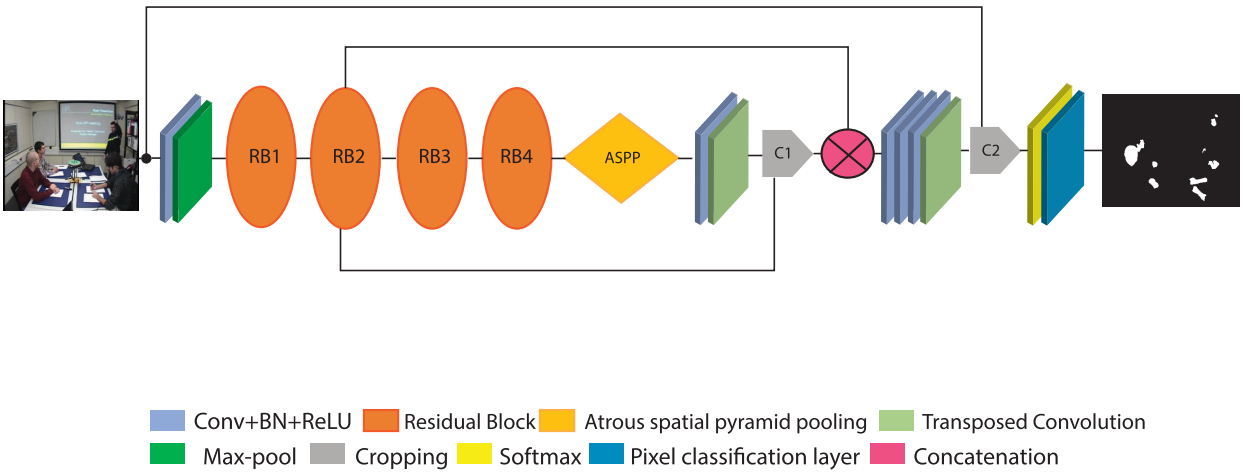


FIGURE 1. Proposed SSS-Net.

information of regions which then passed though the specifically designed ASPP module do not result in loss of small objects. Table 7 shows the valuable differences of our proposed model from Deeplabv3+ architecture.

1) PROPOSED ENCODER

In Table 4 encoder details of SSS-Net are presented. SSS-Net encoder consists of a total of 4 residual blocks, each block consists of convolution layers in sequence followed by separate batch normalization and ReLu activation layers. Every residual block comprises of two 3×3 convolutions and to reduce the size of the image each of the block is interpolated with max-pooling operation. A shortcut connection is provided to each residual block, which combines the input with result of residual block before applying ReLU in second convolution of the block. This connection enables the previous layers to get the powerful gradient signal which makes training easy for the deeper networks. Figure 10 shows a residual block of SSS-Net.

Instead of simple convolution, the final residual block uses atrous convolution [41]–[45] that enables the expanded filter's view. We used different dilation rates i.e. rate=2, rate=4 in the last two blocks. Atrous convolution track the resolution where we measure feature responses. In addition, atrous convolution provides a broader context without increasing computational expense or the number of parameters.

As down-sampling is not implemented at the atrous block, atrous spatial pyramid pooling (ASPP) [46], [47] is performed on the size same as feature response. In SSS-Net, ASPP captures multi scale contextual information and applies various dilation rates to a sequence of atrous convolutions. These rates are designed to capture the longer context. In addition, ASPP integrates image-level features to add global context information. As shown in Figure 11, there are 4 parallel operations in ASPP consisting of one 1×1 convolution and three 3×3 convolution performed with dilation rates 4,12 and 16. The stride we used for the feature maps is 16.

2) DECODER NETWORK

Decoder of SSS-Net used transposed convolution layer to up samples the features coming from the encoder part resulting in high resolution image from a low resolution image. This is followed by concatenation with the resulting low-level network features of the same resolution. On these low level features 1×1 convolution with 256 filters is applied in order to lessen the number of channels, as the resultant low-level features usually have a large number of channels and make the training of network harder. A factor of 4 is applied after concatenation to refine the features following another simple bilinear upsampling. In Table 5 decoder details of SSS-Net are presented. The diagram of a SSS-Net is shown in Figure 1.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA AND ENVIRONMENT

SSS-Net was tested for skin segmentation using five datasets of human activity recognition that are publicly available [8]. Following are the datasets that are used for the task of skin segmentation in this paper;

1) Augmented multi-party interaction (AMI)
2) In-house (SSG)
3) Event detection EDds)
4) UT-interaction (UT)
5) Laboratoire d'informatique en image et systèmes d'information (LIRIS)

These five datasets contain only a few training images. Therefore, we used data augmentation to artificially increase the amount of training data. In Table 2, a detailed description of these datasets is provided. Figures 2 to 6 show examples of segmentation results that are predicted right by our network for all the datasets. Here, the red color represents segmented skin area. As shown in these figures, our model is able to detect the skin area correctly in images in the datasets that

FIGURE 2. Example Images of AMI datasets by SSS-Net that produces good segmentation results. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

includes indoor and outdoor scenes.Also, there are some relatively poor examples of segmentation by our network that are presented in Figure 7. Cases where our network does not give good results are the skin areas near hair and beard. whereas, Skin-like background pixels are the main reason for causing false positive error, while some unfamiliar skin pixels leads to false negative errors. We also performed the performance comparison of SSS-Net with Deeplabv3+ on EdDs dataset which is presented in Table 6, and the visual results are presented in Figure 8. Table 3 shows the segmentation results by our method with other methods on the five datasets. The network was trained on the computer with Intel(R) Xeon(R) W-2133 CPU 3.60GHz, 32 GB RAM, and Nvidia 2080TI GPU, we did training and testing of the proposed network on a desktop computer. We also considered the EdDs dataset for the performance comparison of SSS-Net with Deeplabv3+. The table for the performance comparison of Deeplabv3+ and SSS-Net are shown in Table 6

TABLE 2. Details of human activity dataset for the evaluation of SSS-Net.

Name	Total Images	Training Images	Testing Images
EDds	85	60	25
LIRIS	55	30	25
SSG	45	20	25
UT	50	25	25
AMI	50	25	25

B. DATA AUGMENTATION

In this paper, skin semantic segmentation network SSS-Net is proposed. An augmentation scheme is used for the training data that included:

1) Rotation
2) Contrast enhancement

In deep neural networks, training depends on the size of the input data. In order to carry out effective training, a large amount of data is needed. When the size of the training

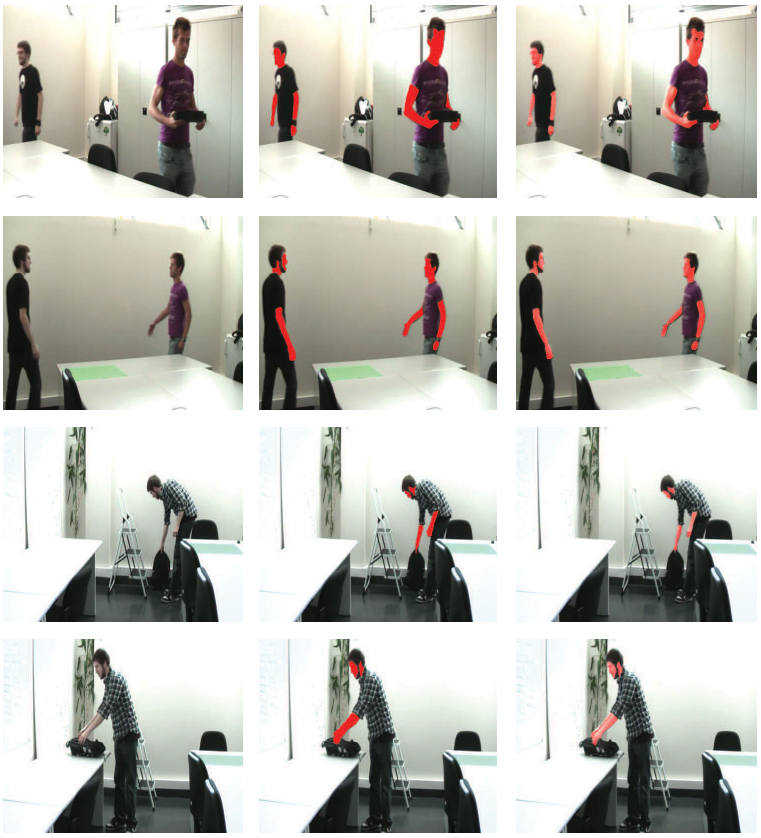


FIGURE 3. Example Images of SSG datasets by SSS-Net that produces good segmentation results. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

TABLE 3. Comparison of SSS-Net with other methods on the AMI, SSG, EdDs, UT and LIRIS datasets.

Method	AMI			SSG			EDds			UT			LIRIS		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
T CbCr (Wang et al. [32])	0.242	0.694	0.359	0.148	0.854	0.252	0.253	0.706	0.373	0.258	0.839	0.395	0.067	**0.914**	0.125
T HS (Wang et al. [32])	0.396	0.321	0.354	0.385	0.548	0.453	0.398	0.484	0.437	0.326	0.571	0.415	0.122	0.327	0.178
BAY H (Jones et al. [33])	0.531	0.804	0.639	0.515	0.493	0.504	0.626	0.502	0.557	0.330	0.590	0.423	0.147	0.647	0.239
BAY G (Jones et al. [33])	0.610	0.784	0.686	0.469	0.476	0.472	0.647	0.524	0.579	0.394	0.455	0.422	0.158	0.690	0.258
RF (Khan et al. [34])	0.503	0.930	0.653	0.436	0.766	0.558	0.502	0.685	0.580	0.284	**0.897**	0.432	0.104	0.886	0.187
ASD(Dadgostar et al. [35])	0.044	0.531	0.082	0.164	**0.902**	0.278	0.022	0.733	0.043	0.002	0.251	0.004	0.038	0.770	0.072
MMI(conaire et al. [36])	0.020	0.549	0.039	0.056	0.552	0.101	0.055	0.436	0.099	0.041	0.141	0.063	0.040	0.800	0.077
DMDA(SanMiguel et al. [8])	0.598	0.842	0.699	0.457	0.754	0.569	0.623	0.648	0.636	0.413	0.755	0.534	0.189	0.698	0.298
OR-Skip-Net(Arsalan et al. [7])	0.868	**0.982**	0.920	0.887	0.853	0.863	0.713	0.925	0.801	**0.927**	0.766	0.831	**0.869**	0.804	**0.813**
SSS-Net (proposed)	**0.875**	0.957	**0.921**	**0.890**	0.8726	**0.8795**	**0.733**	**0.933**	**0.812**	0.914	0.771	**0.8397**	0.841	0.812	**0.813**

data is small, the parameters are uncertain and the training of the network is insufficient, which seriously affects the performance of the network. One way to solve this problem is to perform data augmentation that increases the data size, alleviating this limitation. In this paper, we kept the image size same as in the dataset. We used image rotation

FIGURE 4. Example Images of EdDs datasets by SSS-Net that produces good segmentation results. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

TABLE 4. Encoder with a feature map size of the residual block. Where EConv represents the convolutional layer of the encoder. Convolutional layers with "**" means that it includes Batch Normalization (BN) and ReLU aswell.

Block in Encoder	Name and Size	No. of filters	Output feature map size	No. of parameters
Encoder Res block-1	$EConv-2a\_1^{**}/3\times3\times64$	64	$60\times80\times64$	36928
	$EConv-2b\_1^{**}/3\times3\times64$	64	$60\times80\times64$	36928
Encoder Res block-2	$EConv-2a\_1^{**}/3\times3\times64$	64	$60\times80\times64$	36928
	$EConv-2b\_1^{**}/3\times3\times64$	64	$60\times80\times64$	36928
Encoder Res block-3	$EConv-1\_1/1\times1\times128$	128	$30\times40\times128$	8320
	$EConv-2a\_1^{**}/3\times3\times128$	128	$30\times40\times128$	73856
	$EConv-2b\_1^{**}/3\times3\times128$	128	$30\times40\times128$	147584
Encoder Res block-4	$EConv-2a\_1^{**}/3\times3\times128$	128	$30\times40\times128$	147584
	$EConv-2b\_1^{**}/3\times3\times128$	128	$30\times40\times128$	147584

on these images to generate a composite image using the original training image. Each image rotates 1 degree from 0 to 360. In this way, we obtained 360 rotated images for each image, and a total of 7,200 images were obtained after rotation. To eliminate artifacts in rotated images, we first converted binary images into logical images, then used bi-cubic interpolation when we rotated these images. After rotation, we used the contrast enhancement feature and generated over 1800 images with different contrasts. Therefore, data is synthesized by expanding from 20 images

FIGURE 5. Example Images of UT datasets by SSS-Net that produces good segmentation results. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

TABLE 5. Decoder with a feature map size of the residual block. Where DConv represents the convolutional layer of the decoder. Convolutional layers with "**" means that it includes Batch Normalization (BN) and ReLU aswell.

Blocks in Decoder	Name and Size	No. of filters	Output feature map size	No. of parameters
Decoder conv. block-1	$DConv - c1\_1^{**}/1 \times 1 \times 256$	256	$30 \times 40 \times 256$	262400
Upsample	$Upsample1\_1/8 \times 8 \times 256$	256	$120 \times 160 \times 256$	4194560
Decoder conv. block-3	$DConv - c3\_1^{**}/3 \times 3 \times 256$	256	$60 \times 80 \times 256$	700672
Decoder conv. block-4	$DConv - c4\_1^{**}/3 \times 3 \times 256$	256	$60 \times 80 \times 256$	590080
Upsample	$Upsample2\_1/8 \times 8 \times 2$	2	$240 \times 320 \times 2$	258

to 9000 images. In Table 8 data augmentation details are presented.

C. NETWORK TRAINING

For training SSS-Net, we provided the images to network without any pre-processing. Considering optimizers, a very popular technique used with stochastic gradient descent (SGD) is called Momentum. Momentum not only uses the gradient of the current step to lead the search, but also mounts up the gradient of the past step to determine the direction of progress. Whereas Adam is an adaptive learning rate method that calculates individual learning rates for various parameters. SGD with momemtum appears to find a flatter minima than Adam. However, the adaptive method tends to converge to a sharper minima relatively faster. Flatter minima are better generalize than sharper minima. Although adaptive

FIGURE 6. Example Images of LIRIS datasets by SSS-Net that produces good segmentation results. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

optimizers have better training performance, but this does not mean higher accuracy for different data. Therefore, SGD with momentum is the most popular deep network optimizer [48]. In this study, we used the Stochastic Gradient Descent with Momentum (SGDM) 0.9 with an initial learning rate of 1e-3. We used L2 regularization with a weight decay of 0.0005 for training SSS-Net skin semantic segmentation network. Our network has been trained for 40 epoch, with a minimum batch size of 5 images, and as the convergence rate of our network accelerates, it shuffles after each epoch. The learning-rate decay and mini-batch size are empirically calculated to satisfy minimum loss of cross-validation and the weights and biases of ResNet18 are employed in the initialization stage of the proposed method. As we have also compared the performance of SSS-Net with Deeplab V3+ on EdDs dataset, training time on EdDs dataset was 308 minutes for 40 epochs while testing time on CPU was 1.5 second and 300 ms on GPU. The details for the training stage are presented in Table 9.

TABLE 6. Performance comparison of SSS-Net with Deeplabv3+ on EdDs dataset.

Method	F	P	R
DeepLabv3+	0.8512	0.7121	0.929
SSS-Net	0.8795	0.733	0.933

Cross-entropy loss is measured as the network training objective function. This objective function is driven by probabilities, where p stands for probability. When the obtained estimate for a certain class deviates from the actual desired class, p (the probability parameter) approaches to 1, whereas the loss is stated as the combined loss of all the pixels. Inherently, the ''non-skin'' pixels in each human activity image outweigh the ''skin'' pixels for the task of skin segmentation.

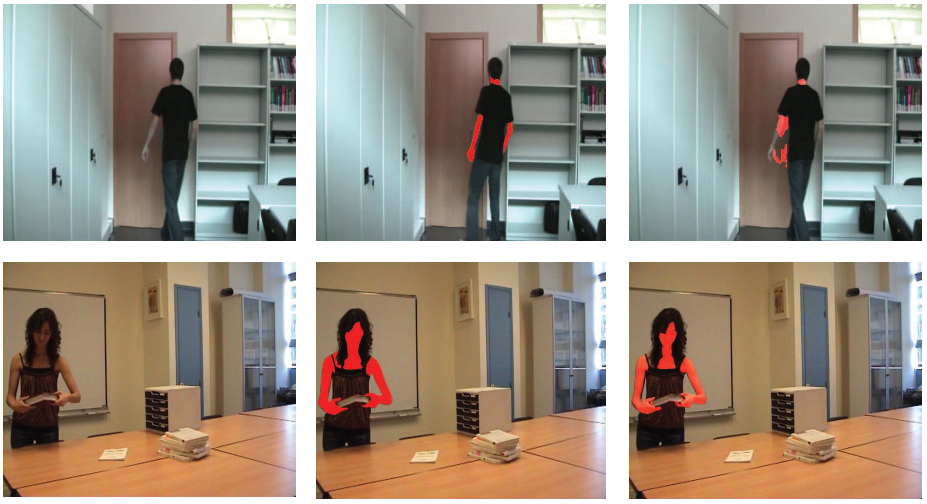


FIGURE 7. Some example Images of bad segmentation of LIRIS and SSG datasets by SSS-Net. (a) Image (b) Ground truth (c) prediction by SSS-Net. Red color presents the segmented skin area.

FIGURE 8. Visual results comparison of SSS-Net with Deeplab V3+ on EdDs dataset. (a) Shows the visual results of ground truth. (b) Shows the visual results of DeepLabv3+. (c) Shows the visual results of the SSS-Net. Red color presents the segmented skin area.

This vast amount of variance in the number of pixels among dissimilar classes may possibly lead to multiple critical problems when using the cross-entropy loss as an objective function for network training. But this problem can be overcome by class balancing as the weights are formulated to associate with each class in the loss function. Consequently, the classes with high frequency have low weights and classes with low frequency have high weights. Numerous different approaches

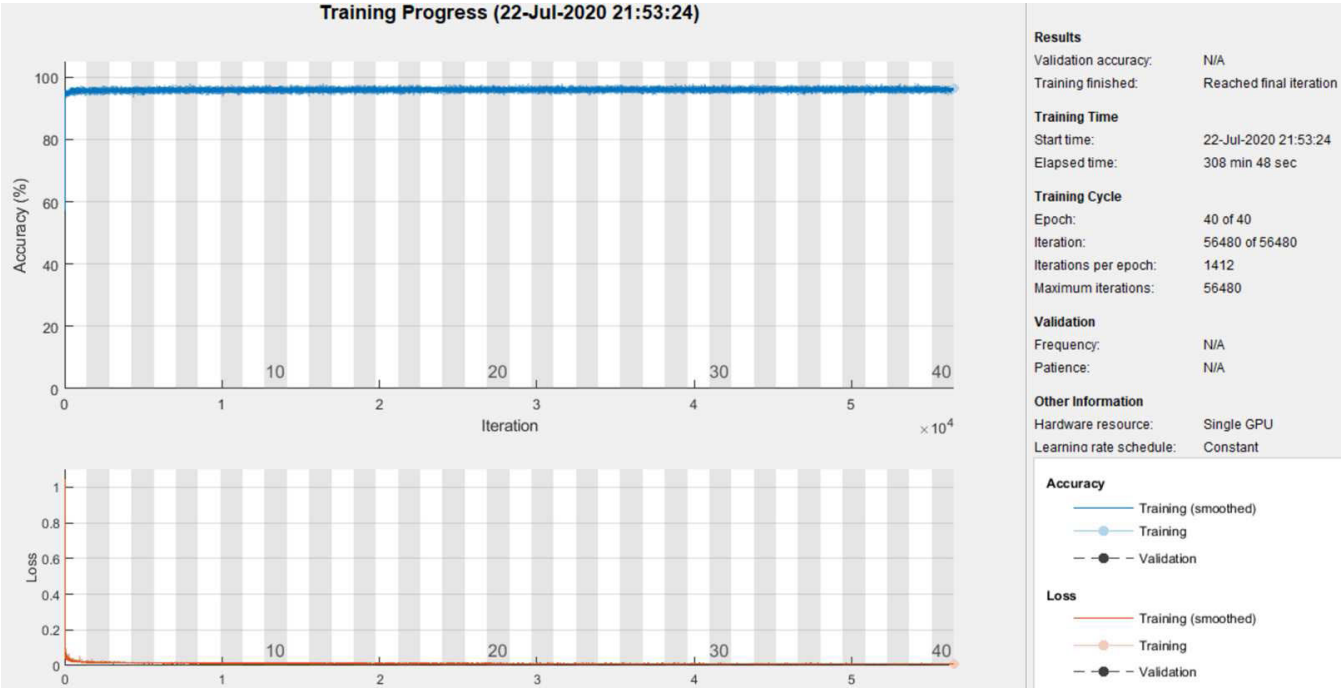


FIGURE 9. The graphs of training loss and accuracies on EdDs dataset.

TABLE 7. Difference between Deeplabv3+ and SSS-Net.

Sr No.	DeepLabv3+	SSS-Net
1	8 Residual blocks in encoder	4 Residual blocks in encoder
2	100 Number of layers in encoder and decoder	68 Number of layers in encoder and decoder
3	113 Number of connections	77 Number of connections
4	25.9 million trainable parameters	7.3 million trainable parameters

TABLE 8. Data augmentation details.

Dataset	Training images	Angle difference	Images generated by rotation	Images generated by contrast enhancement	Total
EdDs	60	3	7200	1800	9000
LIRIS	30	2	5400	1800	7200
SSG	20	1	7200	1800	9000
UT	25	1	9000	1800	10800
AMI	26	1	9000	1800	10800

to assigning these weights can be followed. In the considered approach, the classes association weights were calculated by using frequency balancing for the training of SSS-Net architectures. In this respect the corresponding weights of the classes are determined by dividing the median of the particular class frequency over the class frequency for the complete training set.

D. NETWORK TESTING

1) EVALUATION METRICES

For testing SSS-Net, we considered several assessment metrics such as recall (R), precision (P) and F-measure (F). The formulas for these protocols are given below:

$$R = TP/(TP + FN) \tag{1}$$

$$P = TP/(TP + FP) \tag{2}$$

$$F = 2RP/(R + P) \tag{3}$$

where TP represents true positive, FP represents false positive,and FN represents false negative. Here, FN are the pixels in the ground truth images, which are skin pixels but predicted from the network as non-skin pixels. In ground truth images, FP is the wrongly predicted non-skin pixel and TP are the correctly predicted skin pixels.

TABLE 9. Hyper-parameter settings for deeplabv3+ and SSS-Net architecture for training stage.

Hyper-parameter	DeepLabv3+	SSS-Net
Epochs	40	40
Mini batch size	5	5
Learning decay rate	0.0005	0.0005

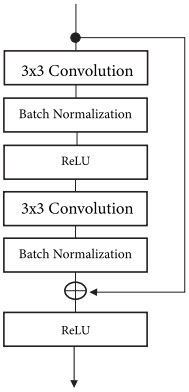


FIGURE 10. Residual block of SSS-Net.

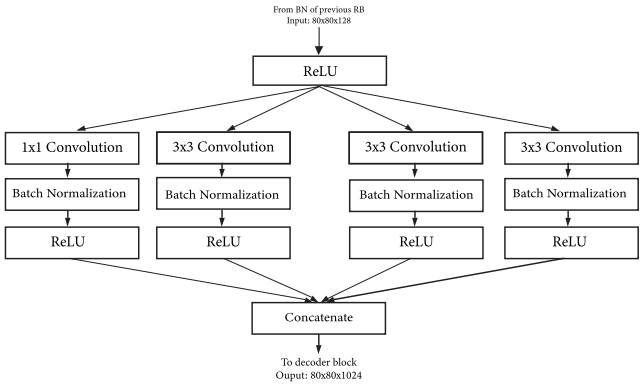


FIGURE 11. ASPP in SSS-Net. Where BN represents Batch Normalization and RB represents Residual Block.

V. DISCUSSION

In this paper, we proposed a skin semantic segmentation network (SSS-Net) for the pixel wise skin segmentation of the input images. In order to improve the network efficiency, the number of layers have been reduced at encoder level. As the task of skin segmentation is very important for human activity recognition, it is important to accurately perform skin segmentation. SSS-Net is able to capture the multi-scale contextual information and controls the signals destruction. In our network we employ the ResNet-18 architecture with 4 residual blocks only. For capturing multi scale contextual information, ASPP (Atrous Spatial Pyramid Pooling) is used in the model. ASPP applies various dilation rates to a sequence of atrous convolutions. These rates are designed to capture the longer context. In addition, ASPP integrates image-level features to add global context information. Skin segmentation is more challenging because of indoor and outdoor image scenes in the datasets. In order to measure the network efficiency, SSS-Net is evaluated on five open datasets of human activity recognition. Our network performs very well on the datasets and the metrics we chose for the evaluation of our network are P, R and F. Experimental results demonstrate the effectiveness of the techniques in our network showing that our network is outperforming the state-of-the-art methods as shown in Table 3.

VI. CONCLUSION

This paper proposed SSS-Net for skin segmentation that is able to capture the multiscale contextual information and provide results with refined edge boundaries. SSS-Net has less number of layers which results in reduced number of parameters i.e. 7.3 M which significantly lower compared to other existing networks. Furthermore, SSS-Net does not require any additional pre-processing steps. The uniqueness of this network is its ability to capture the multi-scale contextual information. We tested SSS-Net for skin segmentation on the publically available datasets of human activity recognition (AMI, SSG, EdDs, UT and LIRIS). Since these datasets contains less number of images, we adopted data augmentation techniques to increase the number of training images. The obtained results show high-quality segmentation results, indicating the effectiveness of SSS-Net for skin segmentation.

ACKNOWLEDGMENT

The authors would like to thank the team of Activity Recognition Dataset (AMI, SSG, EDds, UT and LIRIS) for keeping these databases working and making them easily available for researchers in the field of skin segmentation.

REFERENCES

[1] R. F. Rahmat, T. Chairunnisa, D. Gunawan, and O. S. Sitompul, "Skin color segmentation using multi-color space threshold," in *Proc. 3rd Int. Conf. Comput. Inf. Sci. (ICCOINS)*, Aug. 2016, pp. 391–396.

[2] M. Panwar and P. Singh Mehra, "Hand gesture recognition for human computer interaction," in *Proc. Int. Conf. Image Inf. Process.*, Nov. 2011, pp. 367–374. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050917319130

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[4] 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.

[5] G. Tripathi, K. Singh, and D. K. Vishwakarma, "Convolutional neural networks for crowd behaviour analysis: A survey," *Vis. Comput.*, vol. 35, no. 5, pp. 753–776, May 2019.

[6] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013.

[7] M. Arsalan, D. S. Kim, M. Owais, and K. R. Park, "OR-Skip-net: Outer residual skip network for skin segmentation in non-ideal situations," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112922. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419306402

[8] J. C. SanMiguel and S. Suja, "Skin detection by dual maximization of detectors agreement for video monitoring," *Pattern Recognit. Lett.*, vol. 34, no. 16, pp. 2102–2109, Dec. 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865513002936

[9] B. K. Chakraborty, M. K. Bhuyan, and S. Kumar, "A weighted skin probability map for skin color segmentation," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2016, pp. 2133–2136.
[10] Q. Zhang, M. Yang, K. Kpalma, Q. Zheng, and X. Zhang, "Segmentation of hand posture against complex backgrounds based on saliency and skin colour detection," *IAENG Int. J. Comput. Sci.*, vol. 45, pp. 435–444, 08 2018.
[11] M. Z. Osman, M. A. Maarof, and M. F. Rohani, "Improved skin detection based on dynamic threshold using multi-colour space," in *Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST)*, Aug. 2014, pp. 29–34.
[12] E. Buza, A. Akagic, and S. Omanovic, "Skin detection based on image color segmentation with histogram and k-means clustering," in *Proc. 10th Int. Conf. Electr. Electron. Eng. (ELECO)*, 2017, pp. 1181–1186.
[13] E. Hassan, A. R. Hilal, and O. Basir, "Using ga to optimize the explicitly defined skin regions for human skincolor detection," in *Proc. IEEE 30th Can. Conf. Electr. Comput. Eng. (CCECE)*, Apr. 2017, pp. 1–4.
[14] S.-K. Ueng and C.-Y. Chang, "An improved skin color model," in *Proc. Int. Conf. Appl. Syst. Innov. (ICASI)*, May 2016, pp. 1–4.
[15] V. Bhame, R. Sreemathy, and H. Dhumal, "Vision based hand gesture recognition using eccentric approach for human computer interaction," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 949–953.
[16] K. B. Shaik, P. Ganesan, V. Kalist, B. S. Sathish, and J. M. M. Jenitha, "Comparative study of skin color detection and segmentation in HSV and YCbCr color space," *Procedia Comput. Sci.*, vol. 57, pp. 41–48, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915018918
[17] L. Huang, W. Ji, Z. Wei, B.-W. Chen, C. C. Yan, J. Nie, J. Yin, and B. Jiang, "Robust skin detection in real-world images," *J. Vis. Commun. Image Represent.*, vol. 29, pp. 147–152, May 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047320315000280
[18] H. K. Al-Mohair, J. Mohamad Saleh, and S. A. Suandi, "Hybrid human skin detection using neural network and K-Means clustering tec hnique," *Appl. Soft Comput.*, vol. 33, pp. 337–347, Aug. 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494615002732
[19] A. A. Zaidan, N. N. Ahmad, H. Abdul Karim, M. Larbani, B. B. Zaidan, and A. Sali, "Image skin segmentation based on multi-agent learning Bayesian and neural network," *Eng. Appl. Artif. Intell.*, vol. 32, pp. 136–150, Jun. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0952197614000578
[20] K. Roy, A. Mohanty, and R. R. Sahay, "Deep learning based hand detection in cluttered environment using skin segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1–10.
[21] Y. Kim, I. Hwang, and N. Ik Cho, "A new convolutional Network-in-Network structure and its applications in skin detection, semantic segmentation, and artifact reduction," 2017, *arXiv:1701.06190*. [Online]. Available: http://arxiv.org/abs/1701.06190
[22] A. Dourado, F. Guth, T. Emidio de Campos, and L. Weigang, "Domain adaptation for holistic skin detection," 2019, *arXiv:1903.06969*. [Online]. Available: http://arxiv.org/abs/1903.06969
[23] A. Lumini, L. Nanni, A. Codogno, and F. Berno, "Learning morphological operators for skin detection," 2019, *arXiv:1908.03630*. [Online]. Available: http://arxiv.org/abs/1908.03630
[24] L. Leng, J. Zhang, J. Xu, K. Khan, and K. Alghathbar, "Dynamic weighted discrimination power analysis: A novel approach for face and palmprint recognition in DCT domain," *Information*, vol. 5, no. 12, pp. 467–471, 2010.
[25] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, Jan. 2017, doi: 10.1007/s11042-015-3058-7.
[26] Y. Zhang, J. Chu, L. Leng, and J. Miao, "Mask-refined R-CNN: A network for refining object details in instance segmentation," *Sensors*, vol. 20, no. 4, p. 1010, Feb. 2020.
[27] J. Chu, Z. Guo, and L. Leng, "Object detection based on multi-layer convolution feature fusion and online hard example mining," *IEEE Access*, vol. 6, pp. 19959–19967, 2018.
[28] F. Gao, K. Cao, L. Leng, and Y. Yuan, "Mobile palmprint segmentation based on improved active shape model," *J Multimed Inf Syst*, vol. 5, no. 4, pp. 221–228, Dec. 2018, doi: 10.9717/JMIS.2018.5.4.221.
[29] Y. Wu, L. Leng, and H. Mao, "Non-contact palmprint attendance system on pc platform," *J. Multimedia Inf. Syst.* vol. 5, pp. 179–188, Sep. 2018, doi: 10.9717/JMIS.2018.5.3.179.
[30] L. Leng, G. Liu, M. Li, M. K. Khan, and A. M. Al-Khouri, "Logical conjunction of Triple-Perpendicular-Directional translation residual for contactless palmprint preprocessing," in *Proc. 11th Int. Conf. Inf. Technol.*, Apr. 2014, pp. 523–528.
[31] Y. He, J. Shi, C. Wang, H. Huang, J. Liu, G. Li, R. Liu, and J. Wang, "Semi-supervised skin detection by network with mutual guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2111–2120.
[32] Y. Wang and B. Yuan, "A novel approach for human face detection from color images under complex background," *Pattern Recognit.*, vol. 34, no. 10, pp. 1983–1992, Oct. 2001.
[33] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, Jan. 2002.
[34] R. Khan, A. Hanbury, and J. Stoettinger, "Skin detection: A random forest approach," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2010, pp. 4613–4616.
[35] F. Dadgostar and A. Sarrafzadeh, "An adaptive real-time skin detector based on hue thresholding: A comparison on two motion tracking methods," *Pattern Recognit. Lett.*, vol. 27, no. 12, pp. 1342–1352, 2006.
[36] C. O. Conaire, N. E. O'Connor, and A. F. Smeaton, "Detector adaptation by maximising agreement between independent data sources," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–6.
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[38] M. Yang, B. Li, H. Fan, and Y. Jiang, "Randomized spatial pooling in deep convolutional networks for scene recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 1904–1916.
[39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122
[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[41] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*, J.-M. Combes, A. Grossmann, and P. Tchamitchian, Eds. Berlin, Germany: Springer, 1989, pp. 286–297.
[42] A. Giusti, D. C. Cireäan, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," 2013, *arXiv:1302.1700*. [Online]. Available: http://arxiv.org/abs/1302.1700
[43] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, Dec. 2013, pp. 1–7.
[44] G. Papandreou, I. Kokkinos, and P.-A. Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 390–399.
[45] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, pp. 1–7, Oct. 2015.
[46] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, 2005, pp. 1458–1465.
[47] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.)*, 2000, pp. 2169–2178.
[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

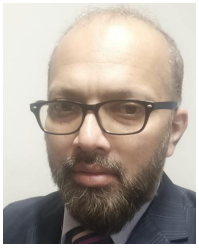
KOMAL MINHAS received the B.S. degree in software engineering from Fatima Jinnah Women University, Rawalpindi, Pakistan, in 2017. She is currently pursuing the master's degree in software engineering with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan.

TARIQ M. KHAN (Member, IEEE) received the B.S. degree in computer engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, the M.Sc. degree in computer engineering from the University of Engineering Technology, Taxila, Pakistan, and the Ph.D. degree in electronic engineering from Macquarie University, Sydney, NSW, Australia, in 2016. He is currently an Assistant Professor with the Department of Electrical Engineering, COMSATS University Islamabad. His research interests include most aspects of image enhancement, pattern recognition, medical image analysis, scene understanding, deep learning methods for image analysis, digital image processing (biometrics), and VLSI.

MUHAMMAD ARSALAN (Member, IEEE) received the B.S. degree in computer engineering from COMSATS University Islamabad, Pakistan, in 2012, and the M.S. degree in computer science from NCBAE, Lahore, Pakistan, in 2016. He is currently pursuing the Ph.D. degree in electronics and electrical engineering with Dongguk University, Seoul, South Korea. He helped to perform the experiments and analysis. His research interests include computer vision and deep learning.

SYED SAUD NAQVI received the B.Sc. degree in computer engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2005, the M.Sc. degree in electronic engineering from The University of Sheffield, U.K., in 2007, and the Ph.D. degree from the School of Engineering and Computer Science, Victoria University of Wellington, New Zealand, in 2016. He is currently an Assistant Professor with COMSATS University Islamabad. His research interests include saliency modeling, medical image analysis, scene understanding, and deep learning methods for image analysis.

MANSOOR AHMED was a Postdoctoral Fellow with Indiana University, USA. He is currently an Assistant Professor with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan. His research interests include information security and privacy, distributed computing, knowledge-based systems, data provenance, and semantic web technologies. He received the Senior Researcher Fellowship Scholarship from Indiana University for the Ph.D. studies and the Higher Education Commission (HEC) Scholarship for higher studies (Ph.D.), Austria.



HAROON AHMED KHAN (Member, IEEE) received the Ph.D. degree from Lancaster University, U.K. He is currently an Assistant Professor with the Department of Electrical Engineering, COMSATS University Islamabad, Pakistan. His research interests include most aspects of machine learning, medical image analysis, scene understanding, and deep learning methods for image analysis.

MUHAMMAD ADNAN HAIDER received the B.S. degree in computer engineering from COMSATS University Islamabad, Pakistan. He is currently pursuing the Ph.D. degree in electronics and electrical engineering with Dongguk University, Seoul, South Korea. He helped to perform the experiments and analysis. His research interests include computer vision and deep learning.

ABDUL HASEEB received the Bachelor of Information Technology (BIT) and Master of Information Technology (MIT) degrees from IQRA University, Karachi, Pakistan, in 2003 and 2005, respectively, the M.S. degree in computer engineering from the COMSATS Institute of Information Technology, Abbottabad, Pakistan, in 2008, and the Ph.D. degree in information engineering from the University of Ferrara, Italy, in 2012. He is currently an Assistant Professor with the Department of Electrical Engineering, Institute of Space Technology (IST), Pakistan. His current research interests include signal and image processing, video encoding (H.264), cross layer design, QoS, QoE, and optimization in networks.