Received July 3, 2020, accepted July 19, 2020, date of publication July 23, 2020, date of current version August 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3011406

Comparison of Deep-Learning-Based Segmentation Models: Using Top View Person Images

IMRAN AHMED[1], (Member, IEEE), MISBAH AHMAD[1], FAKHRI ALAM KHAN[1], AND MUHAMMAD ASIF[2]

[1]Center of Excellence in IT, Institute of Management Sciences, Peshawar 25000, Pakistan
[2]Department of Computer Science, National Textile University, Faisalabad 37610, Pakistan

Corresponding author: Muhammad Asif (asif@ntu.edu.pk)

This work was supported by the Institute of Management Sciences, Peshawar, Pakistan.

ABSTRACT Image segmentation is considered as a key research topic in the area of computer vision. It is pivotal in a broad range of real-life applications. Recently, the emergence of deep learning drives significant advancement in image segmentation; the developed systems are now capable of recognizing, segmenting, and classifying objects of specific interest in images. Generally, most of these techniques primarily focused on the asymmetric field of view or frontal view objects. This work explores widely used deep learning-based models for person segmentation using top view data set. The first model employed in this work is Fully Convolutional Neural Network (FCN) with Resnet-101 architecture. The network consists of a set of max-pooling and convolution layers to identify pixel-wise class labels and prediction of the mask. The second model is based on FCN called U-Net with Encoder-Decoder architecture. The encoder is mainly comprised of a contracting path, also called an encoder, which captures the context in the image and symmetric expanding path called decoder to enable accurate location. The third model used for top view person segmentation is a DeepLabV3 model also with encoder-decoder architecture. The encoder consists of trained Convolutional Neural Network (CNN) to encode feature maps of the input image. The decoder is used for up-sampling and reconstruction of output using important information extracted by the encoder. All segmentation models are firstly tested using pre-trained models (trained on frontal view data set). To improve the performance, these models are further trained using person data set captured from a top view. The output of all models consists of a segmented person in the top view images. The experimental results reveal the effectiveness and performance of segmentation models by achieving IoU of 83%, 84%, and 86% and $mIoU$ of 80% 82% and 84% for FCN, U-Net, and DeepLabv3 respectively. Furthermore, the discussion is provided for output results with possible future guidelines.

INDEX TERMS Deep learning, semantic segmentation, top view person, FCN, U-Net, DeepLab.

I. INTRODUCTION

Nowadays, image segmentation is considered an essential component in many visual-based applications that enable a better understanding of the scene [1]. It mainly involves partitioning of video frames or images into multiple objects or segments and plays a central role in many real-life applications including remote sensing [2], facial segmentation [3], autonomous driving [4], computational photography [5], indoor object segmentation [6], medical image analysis [7], [8], geo-land sensing, augmented reality [9] and object detection [10]. In literature for image segmentation, various techniques have been used, such as region growing, thresholding, watersheds, Otsu, k-means clustering, histogram-based clustering, graph cuts, and Markov random fields [11] and [12]. However, most of these former methods use low-level features and clues for object segmentation. In recent years, deep learning-based models have achieved noteworthy success with remarkable improvement in terms of time and performance accuracy. Researchers developed various neural network-based methods for object detection and classification [13]. In recent studies, deep learning-based techniques created a new generation of image segmentation models. These models are based on semantic pixel-wise

The associate editor coordinating the review of this manuscript and approving it for publication was Hong-Mei Zhang.

labeling and gained significant research interest, as found in [11]. Authors adopted Convolutional Neural Networks (CNN) [14], [15] and pixel-wise labelling for object class prediction [16], which demonstrate encouraging results in terms of accuracy and effectiveness. These models segment the object classes in the input image by predicting pixel information. It also provides a comprehensive explanation about the scene, including the information of the object class, scene semantic, location, and shape, typically based on the Fully Convolutional Network framework [17]. These models are expressed as a pixel classification problem with semantic labels (semantic segmentation).

The deep learning-based segmentation models perform pixel-level labeling for different categories of objects (e.g., human, car, road horse, sky, tree.), as shown in Figure.1. Mostly deep learning-based segmentation models are primarily developed for frontal view object segmentation, as highlighted in Figure.1. From sample images, it can be observed that the developed models predict the segmentation mask for the object when captured from close range in frontal view. The majority of the deep learning-based segmentation models recognize and characterize various objects using a number of images during its training/ learning process, mainly including person images.

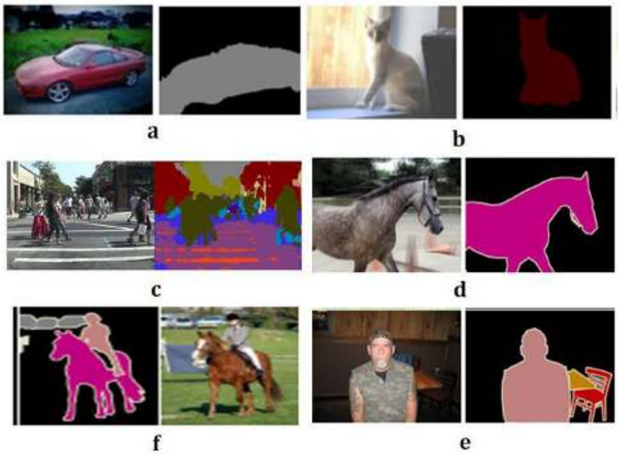

FIGURE 1. Some results of deep learning-based segmentation models including multiple objects mainly from the frontal view (a) & (b) [18], (c) [19], (d) [17] and (e) & (f) [20].

Inspired from excellent results, in this work, widely used deep learning-based segmentation models i.e. FCN [17], U-Net [21] and DeepLabV3 framework [22] have been explored for top view person segmentation. The top view perspective offers wide coverage, better visibility of the objects in the field of view, and may handle occlusion problems better than frontal view [23] and [24]. Moreover, using a single top view camera is also beneficial in terms of installation expense, energy usage, and human resource (required to monitor multiple cameras) [25]. These models were originally pre-trained using data set captured from normal or frontal view i.e. PASCAL [26] and COCO [27]. In this work, multiple person data set is recorded from the top view and used for testing and training purposes. It contains images of multiple persons with a variety of poses, scales, angles, orientations, sizes, and camera resolutions. (See Figure.6 to 8). Experimental results show that deep learning-based segmentation models efficiently segment the person in top view images using semantic features. The specific color mask is assigned using the pre-defined weights file. To decrease computation time, these models are implemented using both CPU and GPU. The main focus of the work is outlined as:

- To the best of our knowledge, in this work, for the first time, deep learning-based segmentation models are explored for top view person segmentation.
- To investigate the performance, pre-trained deep learning-based segmentation models (trained on frontal view data set) are first tested on top view person data set, which is completely different than training data set.
- To improve the performance of pre-trained models, training and testing of deep learning-based segmentation models are performed using top view data set, containing multiple person images with variation in appearance in terms of size, scale, pose, and body orientation in indoor and outdoor environments.
- The comparison of different segmentation models are made for top view person data set.
- The comparison of computation cost of segmentation models have also been made for both CPU and GPU.
- Discussion is made based on experimental results, which provides the significance of deep learning-based models for the segmentation of person in top view perspective along with possible future guidelines.

The work is mainly organized in the following subsequent sections. Section.II delivers a summary of various segmentation based methods used for object segmentation. The top view data set used for experimental purposes is discussed in Section.III. The deep learning-based models used and compared for top view person segmentation are elaborated in Section.IV. The impact and evaluation of these models are reported in Section V. The conclusion of the work, along with future guidelines, is provided in Section.VI.

II. LITERATURE REVIEW

This section briefly describes different segmentation models developed for object segmentation. It provides a summary of traditional generic, machine learning, feature, and deep learning-based methods. A good review of different segmentation techniques is also provide by [11], [28] and [29]. Mostly, researchers in the past used color, texture, and shape information for object segmentation. Some used Probabilistic Graphical Models (PGM's) & graphical models like Bayesian Network (BN) for image segmentation. These models successfully applied in object segmentation using causal relationships between random variables. Several researchers utilized Fuzzy C Means based techniques for object segmentation. Few of them combined traditional segmentation methods with machine learning techniques to enhance performance. However, recent advancement of deep learning techniques in tasks of image classification [30]–[35] and object detection [36]–[39]. These models have also adopted for object segmentation tasks. [20] used convolutional layers, mainly adopted VGG 16-layer network architecture for object segmentation. This model includes deconvolution and pooling layers employed to identify pixel-wise labels information for each class. This information was also used for the prediction of segmentation masks. The model was trained using the PASCAL VOC data set [26]. Dong *et al.* [18] employed deep learning for unified object segmentation. They used global and local context information to distinguish the ambiguous samples in the images. Another deep learning-based pedestrian semantic segmentation model is developed by [19]. This model includes the Faster R-CNN object detection module, and the branch of the network is combined for image segmentation. [17] applied a fully convolutional neural network for the object segmentation. Dvornik *et al.* [40] introduced a deep learning-based scene understanding model called BlitzNet. The model was used for object segmentation in the forward pass, allowing real-time computation. Reference [41] developed Multitask Network Cascades for instance-aware based segmentation. The model composed of three stages the first one discriminates objects, the second estimate masks for each object, and the third categorizes objects. The whole model used convolutional features of VGG-16 as backbone architecture. Wang *et al.* [42] developed convolutional neural networks (CNNs) for scene understanding based on pixel-wise segmentation model. The model used dense up-sampling convolution, which generates pixel-level prediction and hybrid dilated convolution (HDC) framework. The authors used the KITTI road and PASCAL VOC 2012 data set for segmentation tasks. Another instance aware based semantic segmentation model was developed by [43], which used the merits of FCN for segmentation. The developed model was capable of detecting and segmenting the object instances simultaneously. Reference [44] used fully convolutional neural networks (FCN) for multi-scale input image.

From the literature survey, we concluded that most of the segmentation models had been developed for frontal view objects. The deep learning-based model in literature achieved good accuracy results mainly for benchmark data sets such as COCO [27] and PASCAL VOC [26]. Some researchers also preformed top view person detection. Reference [45] employed background subtraction based methods for top view person detection and counting. Ullah *et al.* [46] used rotation invariant blob-based segmentation to track and detect people in top view images. Reference [47] developed a blob based method for tracking people in industrial environments. Reference [48] proposed an efficient detector using a lookup table and point-based transformation for top view person images. Reference [49] performed top view object detection using deep learning models. In this work, segmentation models based on deep learning architecture are examined for the top view person data set. Reference [50] used CNN based algorithm for top view person tracking.

III. DATA SET

To the best of our knowledge, mostly existing data sets used by researchers for person segmentation are mainly based on the frontal view. Some researchers used top view images, but those data sets are either not available or captured using drone cameras from different heights in which the visibility of the object is not clear. For person segmentation, a data set has been recorded in this work using a single top-view camera. It contains multiple person images against a variety of backgrounds in outdoor and indoor environments. Utilizing the top view perspective causes changes in the physical appearance of the object in terms of pose, scale, size, and orientation. Some sample images of the recorded data set are depicted from Figure.6 to Figure.8. The data set contains multiple person images recorded via Point Grey Fly Cap2 (wide-angle lens) camera and Hikvision HV-DS-2CD2T83G0-I5 (normal view) camera at the Institute of Management Sciences (IMSciences), Hayatabad, Peshawar (Pakistan). In this research work, a person is mainly focused as it is considered an essential part of video surveillance. Table.1 describes the data set.

TABLE 1. Top view person data set.

S.no	Description	
1	Color Model	RGB
2	Video type	.mpeg
3	Frame rate	20 frames per second
4	Height of camera	4 meters from the ground
5	No. of subjects	1 to 12
6	Image Resolution	640 X 480 & 1280 X 720
7	Location	Indoor/Outdoor
8	Shadows/Reflections	Yes
9	Image Format	. JPG
10	Number of Images	10,000

IV. TOP VIEW PERSON SEGMENTATION USING DL

Image segmentation is also called semantic image labeling, segments the arbitrary sized image by allocating every pixel of an input image to the label class object [11]. It generally combines the image segmentation method with object recognition techniques. Various deep learning-based segmentation models e.g. [1], [18], [21], [22] and [51] are developed for different applications. In this section, we mainly focused on prominent and widely used deep learning-based semantic segmentation models for top view person data set. The first model employed for top view person segmentation is Fully Convolutional Network (FCN) [17]. It utilized only locally connected layers, like up-sampling, pooling, and convolution. The architecture does not consist of any dense layer in order to minimize the computation time and the number of parameters. To acquire output (segmentation map), it has two paths; the first is the down-sampling path, used to capture semantic/contextual information, while the second one is the up-sampling path, which recovers spatial features.

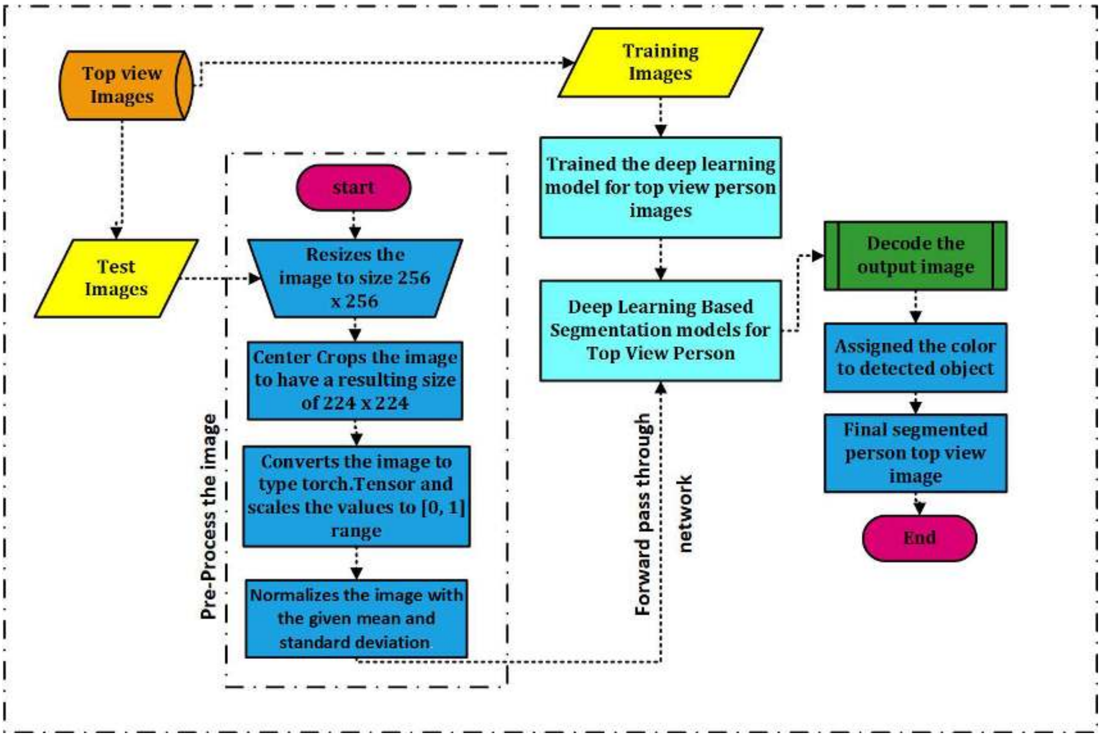
Top view Images
Training Images
Trained the deep learning model for top view person images
Test Images
start
Resizes the image to size 256 x 256
Center Crops the image to have a resulting size of 224 x 224
Converts the image to type torch.Tensor and scales the values to [0, 1] range
Normalizes the image with the given mean and standard deviation
Pre-Process the image
Deep Learning Based Segmentation models for Top View Person
Forward pass through network
Decode the output image
Assigned the color to detected object
Final segmented person top view image
End

FIGURE 2. Frame work of Deep Leaning Based Segmentation for top view person images.

The characteristic of this model is mainly the generation of segmentation map for arbitrary sized image. FCN is also used as a baseline model for other semantic segmentation based architectures, e.g., U-Net. The second model used in this work is U-Net [21] having a similar encoder-decoder architecture as FCN but having two distinct features. U-net is symmetric, and it skips the connections between the up-sampling and down-sampling path, used as a concatenation operator. The last model explored for top view person segmentation is recently developed DeepLabV3 framework. It usually shares parallel and cascaded modules of atrous convolutions [22]. The ResNet-101 architecture has been modified, using atrous convolutions it helps to keep high-resolution feature maps in deep blocks. All these models have been trained using top view person images. After training, models assign a color to object using the color variable, as shown in Figure.2. In the following subsection, deep learning-based semantic segmentation models have been discussed and compared using person data set captured from the top view. The general framework of top view person segmentation is depicted in Figure.2. It can be seen that before testing of the top view images, pre-processing and normalization is performed. The trained models require input images normalized similarly to training. The pre-processed image is further given to deep learning-based semantic segmentation models. The output image is decoded by mapping it to the corresponding assigned segmented color using the color variable. The color variable assigns pre-defined color for each segmented object.

A. FCN BASED TOP VIEW PERSON SEGMENTATION

Long *et al.* [17] presented Fully Convolutional Networks architecture (FCN) for robust segmentation by utilizing fully convolutional layers in place of last fully connected layers, as shown in Figure.3. This significant advancement enables the network to create a dense pixel-wise prediction. To obtain localization performance, up-sampled outputs are combined with high-resolution activation maps, which is further transferred to the convolution layers to construct accurate output. In this section, the FCN based semantic segmentation process has been explored for person segmentation, as explained in Figure.3. For training the model, person images captured from the top view is used. The model extracts a semantic feature map of the input image. These semantic features are usually obtained using training images, which are further used to build learned/trained semantic features. The overall flow for top view person segmentation using the FCN model is shown in Figure.3. The model allows dense prediction of arbitrary sized images. The Resnet-101 (without fully connected layers) is used as a backbone that generates features. Instead of classification scores like other deep learning-based object detection models, it outputs a spatial segmentation map, as illustrated in Figure.3. It takes an arbitrary sized input image and performs pre-processing. After forwarding to the segmentation model, it generates a segmentation map of the same size as input. The extracted features, then pass through two 1 × 1 convolutional layers which generate output with 10 × 8 × 6 dimension.

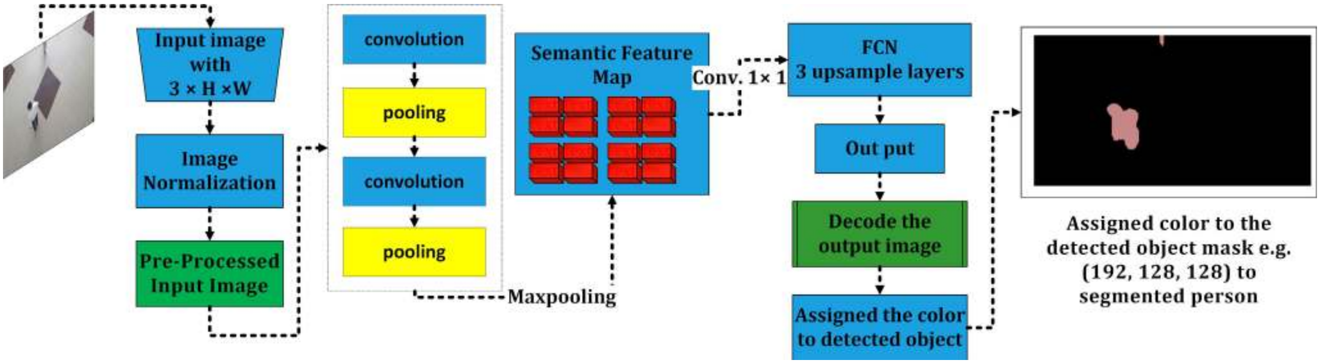
Input image with 3 × H ×W
Image Normalization
Pre-Processed Input Image
convolution
pooling
convolution
pooling
Maxpooling
Semantic Feature Map
Conv. 1× 1
FCN 3 upsample layers
Out put
Decode the output image
Assigned the color to detected object
Assigned color to the detected object mask e.g. (192, 128, 128) to segmented person

FIGURE 3. Framework of FCN architecture [17] (Resnet-101 as backbone) for top view person segmentation.

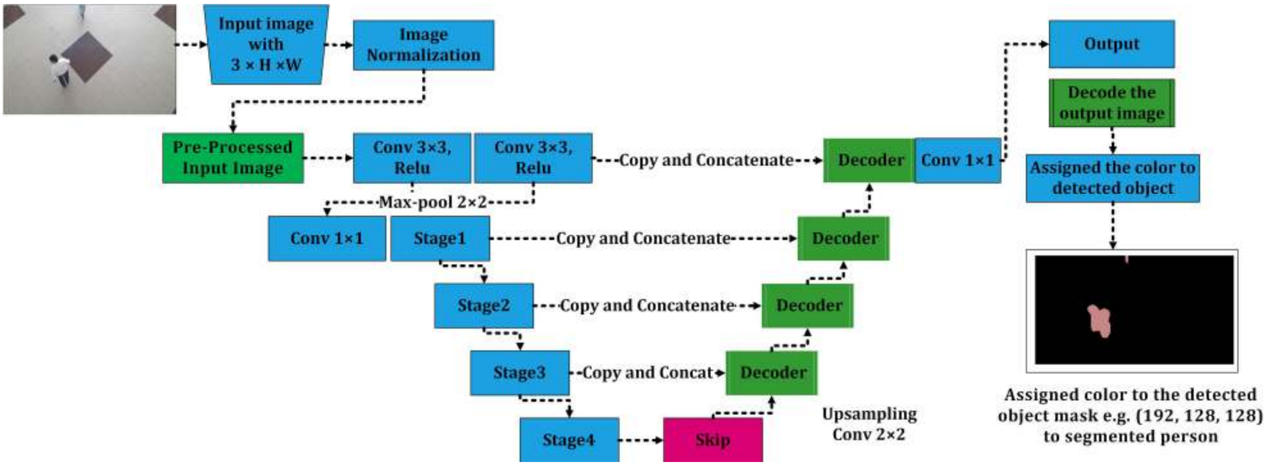

FIGURE 4. Framework of U-Net architecture [21] (left side with Resnet-101) for top view person segmentation.

The output of these convolutional layers is further processed to FCN architecture [17], which performs up-sampling. The models learning of the entire image is performed by backpropagation and dense feed-forward computation [17]. The three up-sampling layers within the network enable pixel-wise prediction. The loss function is measured for the image segmentation task using pixel-wise cross-entropy. It is widely used for image segmentation tasks. It individually evaluates loss of class predictions for pixel vector and further calculates the average value of all pixels. In case, if in the input image, an unbalanced class representation exists, it may cause some error. Reference [17] presented sample or weight loss function for each output channel to reduce class imbalance problem. The function observes each pixel separately by comparing the class prediction results with the detected segmentation class. The function is defined as [17];

$$\lambda_{loss} = \sum_{x \epsilon N} \sum_{x \epsilon L} y_{true} \log y_{pred} \tag{1}$$

where N and L represents a set of all objects and set of all class labels respectively. $log y_{pred}$ represents predicted pixels and y_{true} is the ground truth.

B. U-NET TOP VIEW PERSON SEGMENTATION

The U-Net segmentation model is developed by [21], based on the idea of FCN. Its architecture is similar to FCN encoder-decoder architecture, basically divided into three parts. The first part is the down-sampling path mainly uses Resnet101 as the backbone consists of 4 stages. Each stage primarily applies two 3 × 3 convolution with batch norm followed by 2 × 2 max-pooling, as shown in Figure.4. The horizontal bottleneck consists of two 3 × 3 convolution followed by 2 × 2 up-convolution, as depicted in Figure.4. The up-sampling path also consists of 4 stages shown as decoder containing two 3 × 3 convolutional layers followed by 2 × 2 up-sampling. The features maps become half at each stage.

As shown in Figure.4 the model skips connections between up-sampling and down-sampling paths in order to provide local and global information during up-sampling. Finally, at output 1 × 1 convolutional layer provides the segmented output, where the number of feature maps is similar to the number of desired segments. The general framework for person segmentation is explained in Figure.4, The top view person images are used for training the model; the images are pre-processed and forward to the trained model.

The input images are normalized with mini-batches of 3-channel RGB. The shape of the input images is $(N, 3, H, W)$, where N is the number of images, H, and W represent the height and width of the image. The energy function is calculated as the cross-entropy loss function combined with pixel-wise soft-max over the final feature map [21]. It is defined as;

$$p_k(x) = \frac{exp(a_k(x))}{\sum_{k'=1}^{K} exp(a'_k(x))} \tag{2}$$

In above equation, activation in feature channel is denoted by a_k. The number of classes is represented as K, while approximate maximum-function is represented by $p_k(x)$. The value of $p_k(x) \approx 1$ for maximum activation with $a_k(x)$ at k and for all other values of k it is given as $p_k(x) \approx 0$. At each position the cross entropy is then penalized as [21];

$$E = \sum_{w(x)\epsilon\Omega} \log(p_{l(x)}(x)) \tag{3}$$

where true label of each pixel is represented as $l : \Omega \rightarrow 1, \ldots\ldots\ldots\ldots.., K$ and weight map is $w : \Omega \rightarrow IR$ which gives more importance to some pixels during training [21]. The ground truth segmentation with different frequency pixels in training data set for certain classes is pre-computed through morphological operations, the weight map is given as:

$$w(x) = w_c(x) + w_o.exp\frac{-(d_1(x) + d_2(x))^2)}{2\sigma^2} \tag{4}$$

Hence, $w_c : \Omega \rightarrow IR$ represents weight map used for balancing of class frequencies. d_1 shows distance to border and d_2 represents distance from second nearest border. $w_o = 10$ and $\sigma \approx 5$. (for more details of Equation.3 and Equation.4 readers are refer to [21]).

C. DeepLabV3 TOP VIEW PERSON SEGMENTATION

Chen et al [22] recently developed deep learning-based segmentation model. The model mainly uses encoder-decoder architecture, as depicted in Figure.5. The encoder consists of the CNN model, which is used to get encoded feature maps of the input image. The decoder is used to reconstruct the output from the essential information extracted by the encoder, using the up-sampling method. It includes atrous separable convolution for each input channel along with point-wise convolution, as shown in Figure.5. To deal with multi-scale images, DeepLab utilizes a method of multiple pooling layers, also known as Spatial Pyramid Pooling (SPP). It divides the feature maps extracted from the convolutional layer into spatial bins of fixed number as input image size. In order to expand the field of view of filters, DeepLabV3 uses atrous convolution with SPP, which helps to integrate larger context without increasing the number of parameters. The model presented by [22] has been modified from previous models, with more layers of Xception backbone with depth wise dilated separable convolutions is used rather than using of max-pooling and batch normalization. The encoder module is shown in Figure.5 applying atrous convolution and encodes the multi-scale contextual information. The effective and simple decoder module refines the object boundaries and segmentation results. The atrous convolution allows the network to control features resolution, which is computed using CNN. It also helps to capture multi-scale information by adjusting filters. In case of two-dimensional signals for the input feature map x atrous convolution is applied to compute output feature map represented as y and convolution filter w as follows [22];

$$y[i] = \sum_{k} x[i + r.k]w[k] \tag{5}$$

In the above equation, r represents the atrous rate to calculate the stride, which is used for sampling of the input image(we refer reader [22] for more details of Equation.5). The value of r for standard convolution is equal to 1. For reducing computation complexity, [22] used depthwise separable convolution as depicted in Figure.5. For each input channel, depth-wise convolution performs spatial convolution along

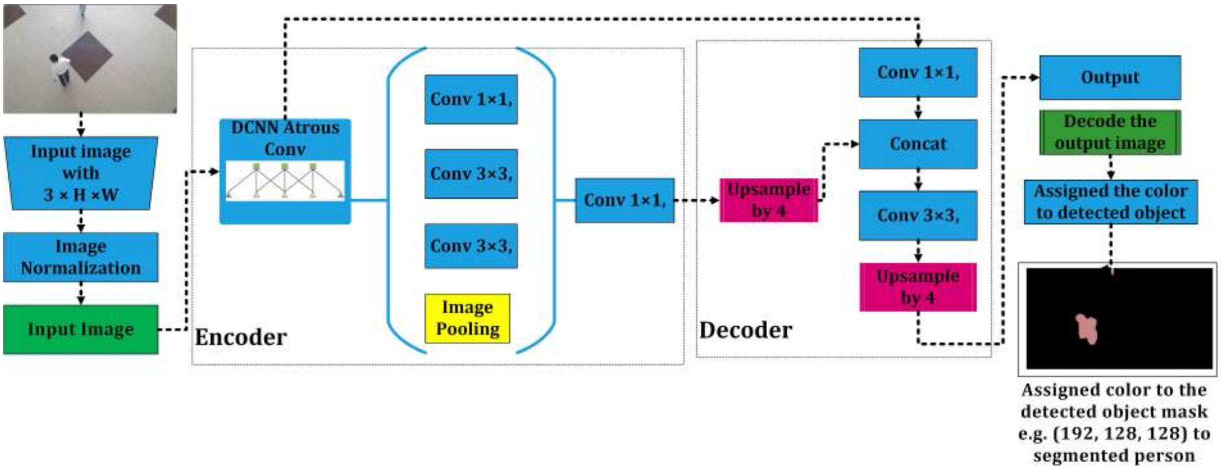

FIGURE 5. Frame work of DeepLabv3 architecture [22] for top view person segmentation.

with point-wise convolution for output. The model was originally trained using COCO and PASCAL VOC data set. It can be visualized from Figure.5 that the decoder module effectively improves the segmentation results, especially around the object boundaries. Instead of ResNet-101, Xception with significant modification [22] is used as main feature extractor or encoder. From Figure.5 that all max pooling operations are swapped with depth-wise separable convolution. The DeepLab model employs the energy function as [52]

$$E = \sum_{i} \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{6}$$

In above equation, label assignment for pixel is represented by x, i & j varies between 1 to N $\theta_i(x_i)$ is the unary function given as [53];

$$\theta_i(x_i) = -\log P(x_i) \tag{7}$$

where label probability assignment at pixel i is represented as $P(x_i)$ and calculated as [53]. To calculate all connecting pairs of image pixels, i, j following expression is used [53].

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)\Big[w_1 exp\Big(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\Big) + w_2 exp\Big(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\Big)\Big] \tag{8}$$

In above equation the value of $\mu(x_i, x_j)$ is equal to 1 if $x_i = x_j$ and otherwise 0. The rest of the expression utilized two Gaussian kernels in unlike feature spaces named as 'bilateral' kernel represented as p depends upon pixel positions and I as RGB color, the second one only deep learning-based upon pixel positions. σ_α, σ_γ and σ_γ represents hyper-parameters used to control the scale of kernels. (for more details of above equations readers are refer to [52] and [53].)

In this work, the above discussed models are trained using person data set captured from the top view. The overall algorithm for person segmentation from top view using FCN, U-Net, and DeepLabV3 models shown in Figure.2 is illustrated in the following steps. The models are tested for multiple people top view images. The segmentation models output the segmented person in top view image using assigned color information. The details of top view person segmentation models shown in Figure.3, 4 and 5, is provided as:

- The Deep learning models, i.e., FCN, U-Net, and DeepLabV3, take an RGB image as input, which is normalized using standard deviation and ImageNet mean method. The dimension of the input for each model contains batch size *Ni*, RGB channels *Ci* the width, and height is represented by *Wi* and *Hi*, respectively. After normalization of the image is pre-processed and resized.
- The models take arbitrary sized input images and produce the same sized segmented image as input. At output in the case of FCN, up-sampling is used to create feature maps, as shown in Figure.3. It helps to create the same sized output image as input. Similarly, in the case of U-Net and DeepLabV3, the given arbitrary sized top view image is down-sampled, as the encoder is based on output strides. In the case of U-Net, there is no dense layer, so images of various sizes may be used as input. In DeepLabV3, the encoded features are first up-sampled, which are further concatenated with similar low-level features. To reduced the number of input channels, 1 × 1 convolutions are applied on low-level features before concatenation, as depicted in Figure.5. Finally, 3 × 3 convolutions are applied, and the features are up-sampled. In this way, the model provides an output of the same size as that of the input image as depicted in Figure.5.
- At, testing phase, the top view person images are passed through learned models for person segmentation. Unlike other traditional blob and segmentation based models, these models used segmented color information to label the segmented object at the output, i.e., a person as shown in Figure.3, 4 and 5.
- The image is converted into an RGB image, and each segmented label class pixel is mapped to its respective color. In original models, the variable is used, which is referred as label color. It stores the color information for each class according to the index value assigned by the trained model. For example, as in this work, our focus is the person; therefore, the color variable form store's 0 index for background and the person the RGB value (192, 128, 128) is used. For other objects, different values are stored and assigned by the color variable, as defined in the original models. An empty 2D matrix for all three color channels, which means an RGB image is converted into a 2D image. Now the R, G, and B arrays are formed from the three color channels, each having the shape of the original 2D image.
- Finally, we get the output image containing a segmented object with the assigned color. At output as shown in Figure.3, 4 and Figure.5 the images are decoded into the same dimension as input. It is represented by *No* dimension of the output image, same as *Ni*, the *Co* which represents the color value assigned by the models to segmented class the *Wo* and *Ho* the width and height of the output image as same as input image *Hi* and *Wi*.

V. EXPERIMENTAL RESULTS

In this section, the testing and evaluation results of deep learning-based segmentation models for top view person images have been briefly discussed. The models have been implemented utilizing the HP core. i5 processor with 8 GB RAM using python with OpenCV 3.6. The section is categorized into three parts; the first section provides details about visualization results of segmentation models; the second section offers performance evaluation in terms of accuracy, and in the last section, time inference has been discussed.

A. VISUALIZATION RESULTS OF SEGMENTATION MODELS FOR TOP VIEW PERSON IMAGES

The visualization results of segmentation models used for top view person images have been discussed in this subsection. These models achieve better accuracy as compared to traditional segmentation models. The top view data set is used for testing purposes, containing images of multiple people in outdoor and indoor environments. It can be viewed from Figure.6 to Figure.9 that from the top view, the appearance of the person is considerably changing in terms of scale, size, orientation, and pose. The segmentation results of the discussed models for a single indoor person is depicted in Figure.6. It can be seen that models efficiently detect and segment the person in top view images as compared to the traditional segmentation techniques, which required pre-processing for noise removal (due to the background and illumination changes) for correctly segmenting the required region of interest. In this work, without using any noise/shadow removal techniques, the models give good output results for an arbitrary sized input image. In output results, the person's appearance varies in terms of size, pose, and orientation of a person's body with respect to the camera location. It can be seen from the sample images of Figure.6 (a to d), that how these models accurately segment person region in top view images.

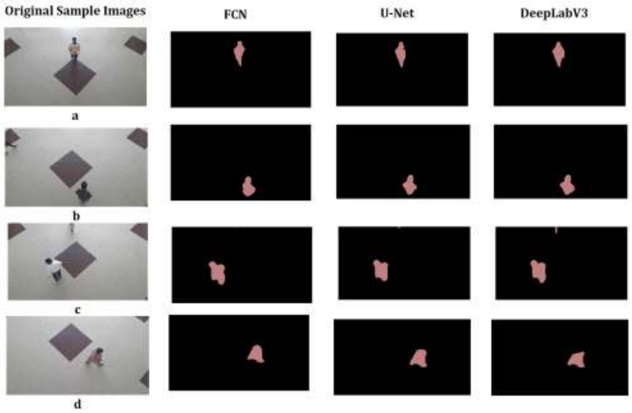

FIGURE 6. Results of segmentation models for top view person images (indoor environment).

We also tested segmentation models for multiple persons top view images as depicted in Figure.7. It can be visualized from Figure.7(a) & Figure.7(b), that models segment multiple person regions efficiently. In some cases where models are supposed to produce separate segmented region for multiple persons as highlighted in Figure.7(c), due to the close interaction of people, these models results single segmented region at the output. In Figure.7(d) and Figure.7(e), where people are close interacting with each other, and for traditional segmentation models, it is usually difficult to segment the person accurately. Here, the deep learning-based segmentation models show outstanding performance by efficiently and accurately segmenting regions containing multiple persons.

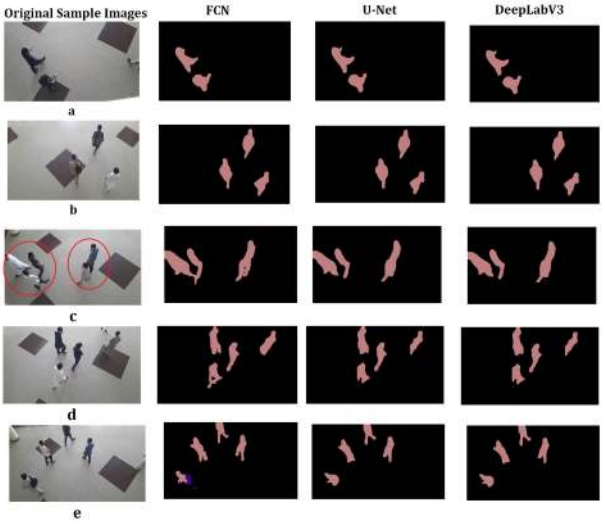

FIGURE 7. Top view multiple person semantic segmentation results (indoor environment).

The output results in the form of segmented regions are nearly same for all three models. The results of the segmentation models using a wide field of view are also shown in Figure.8. All three models show good results for multiple person images, but we also reported some failure examples. In Figure.8(c), the unsegmented person is highlighted with the help of a red circle situated lower at the right corner of the test image, which is not segmented by any of the three models. Although at different locations of the input image, the person's appearance is varied, still all three models efficiently segment regions for multiple persons, as depicted in sample images of Figure.8. Segmentation models are also tested and evaluated for outdoor images captured from the top view camera, as displayed in Figure.9. The sample images show that segmentation models show good results in

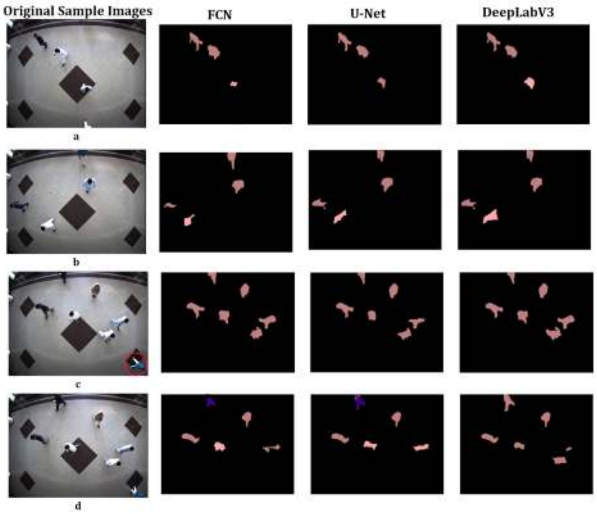

FIGURE 8. Top view multiple person semantic segmentation results for (indoor environment).

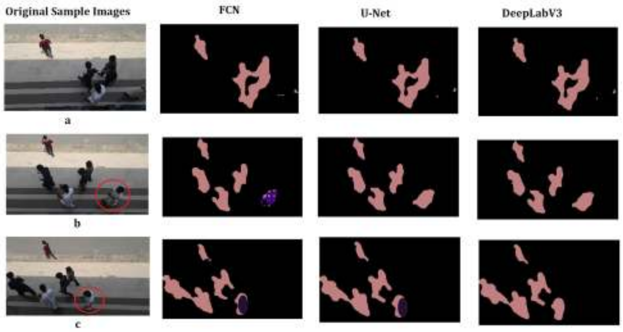

FIGURE 9. Top view multiple person semantic segmentation results for (outdoor environment/normal field of view.)

complex variations of lightning conditions and background. In Figure.9, most of the time, people are very closed to each other. As a result, a single segmented region is produced by all three models. Form sample images, it can be seen that for any segmentation algorithm, it is not easy to produce single separable regions for these kinds of overlapping objects. Overall the DeepLabV3 model shown comparatively better results than other models in outdoor environments.

B. PERFORMANCE EVALUATION

This sub-section briefly discusses the performance evaluation of segmentation models used for the top view person data set. Different matrices are available for evaluation and measurement of the accuracy of segmentation techniques [54]. The evaluation matrices used in this work is given as follows;

1) RECALL, PRECISION AND F1 SCORE ($PREC, REC, F1-SCORE$)

These are considered as popular evaluation matrices for many classical image segmentation techniques. Recall and Precision for each class is determined as;

$$Rec = \frac{tp}{tp+fn} \tag{9}$$

$$Prec = \frac{tp}{tp+fp} \tag{10}$$

The harmonic mean of precision and recall defines F1-score, given as;

$$F1-score = \frac{2Prec \times Rec}{Prec+Rec} \tag{11}$$

2) PIXEL ACCURACY (P_{acc})

It is the most widely used evaluation metric for segmentation models. It is defined as accuracy of pixel-wise prediction given as;

$$P_{acc} = \frac{\sum_{i=0}^{K}(p_i i)}{\sum_{i=0}^{K}\sum_{j=0}^{K}(p_i j} \tag{12}$$

In the above equation, K represents the total number of pixels in the test image, and p_{ii} is predicted pixels as class i, and the ground is represented as p_{ij}, the number of pixels of class i predicted as class j.

3) INTERSECTION OVER UNION (IoU)

It is also recognized as the Jaccard Index, commonly used evaluation metric to calculate the performance of segmentation models. It is generally defined as the ratio of intersection and union area between the predicted segmentation map and ground truth expressed as;

$$IoU = J(A,B) = |A \cap B| \quad / \quad |A \cup B| \tag{13}$$

In the above equation, B shows the predicted segmentation maps, and A represents ground truth.

4) MEAN-IoU ($mIoU$)

It is another widely used matric for segmentation models. It is determined as the average value of IoU overall label classes. It is generally used to report the performance of segmentation models. It usually ranges between 0 and 1 given as;

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{tp}{\sum_{j=0}^{k}fn + \sum_{j=0}^{k}fp - fn} \tag{14}$$

where k represents total classes, tp is number of true positive, and fp and fn are false positive and false negatives.

In this work, we first used the pre-trained (trained using frontal view data set) deep learning-based segmentation models to test top view person images. The results of the pre-trained models are depicted in Figure.10.

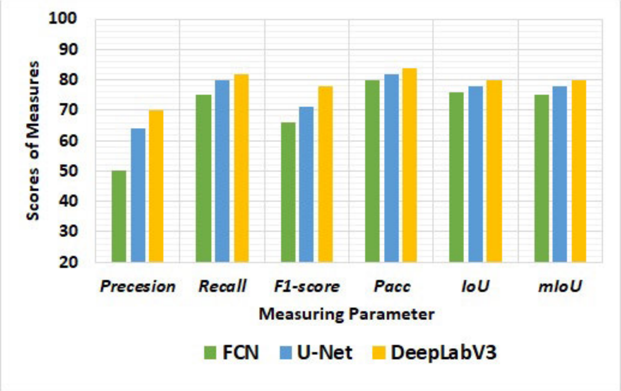

FIGURE 10. Evaluation results of pre-trained segmentation models for person images captured from top view.

To further boost segmentation models' performance, training is performed using the top view data set. Figure. 11 and 12 demonstrates the training accuracy and training loss curves of segmentation models used in this work. The result is obtained using cross-entropy and loss functions, as discussed earlier. In Figure.11, it is noted that the U-Net and DeepLabV3 accuracy curve is slightly better than the FCN model. Similarly in Figure.12, the DeepLabV3 and U-Net loss curves are decreases, which shows that it converges more easily. Based on comparison results, we concluded that U-Net and DeepLabV3 are suitable in terms of accuracy.

TABLE 2. Comparison results of segmentation models using top view person data set.

S.No	Method	Precision	Recall	F1-score	Pixel Accuracy	*IoU*	*mIoU*
1	Otsu [57]	50%	80%	70%	78%	70%	55%
2	Watershed [56]	52%	82%	74%	80%	76%	60%
3	Gaussian mixture-based model [58]	60%	82%	76%	82%	79%	70%
4	Gaussian mixture-based model [59]	60%	84%	79%	82%	80%	72%
5	Background subtraction-based [60]	65%	84%	74%	84%	80%	70%
6	Re-weighted HOG & image segmentation [55]	68%	82%	75%	82%	79%	70%
7	Adaptively splitted GMM [61]	70%	84%	75%	82%	80%	74%
8	FCN [17]	62%	92%	76%	91%	83%	80%
9	U-Net [21]	74%	92%	81%	92%	84%	82%
10	DeepLabV3 [52]	80%	96%	83%	93%	86%	84%

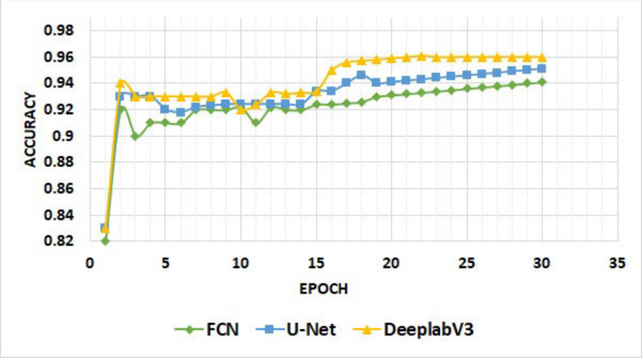

FIGURE 11. Segmentation models training accuracy.

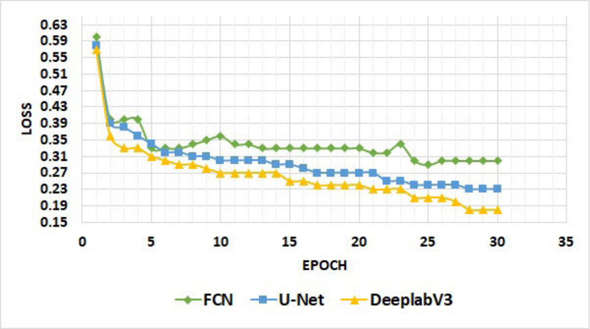

FIGURE 12. Segmentation models training loss.

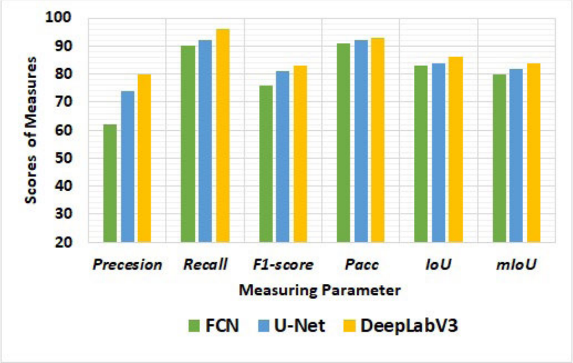

FIGURE 13. Evaluation results of segmentation models trained and tested using person data set captured from the top view.

The trained models are further evaluated using top view data set containing test images. Figure.13 summarizes the results of segmentation models trained and tested using top view person images. It can be seen that the performance of trained models is improved/enhanced as compared to pre-trained models shown in Figure.13. There is a miserable difference between the accuracy of measuring matrices.

The quantitative results of the above discussed models are also elaborated in Table.2. The overall results reveal the effectiveness of deep learning-based models compared with traditional ones. In this work, we compared the results of the well known and mostly addressed techniques, e.g., improved HOG based image segmentation model proposed by [55], watershed [56] Otsu based image segmentation [57], improved adoptive background subtraction based methods [58]–[60] and adoptive splitted GMM- based methods [61] using top view person data set. Table.2 also demonstrates the results of measuring matrices for discussed deep learning-based segmentation models along with traditional feature and background subtraction based methods.

C. COMPARISON IN TERMS OF TIME INFERENCE

All models have been compared in terms of time inference for both CPU and GPU, as seen in Figure.14 and Figure.15.

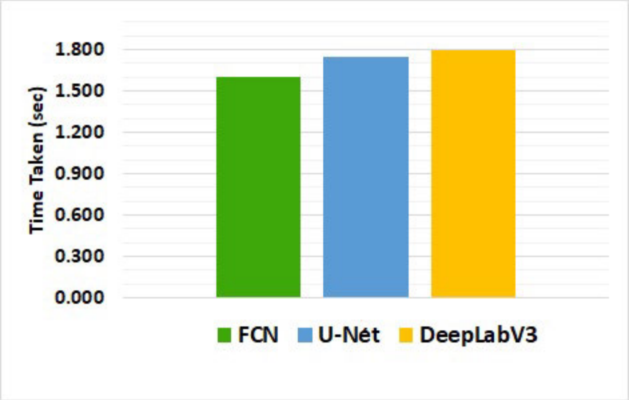

FIGURE 14. Inference time of FCN U-Net and DeepLabV3 on CPU.

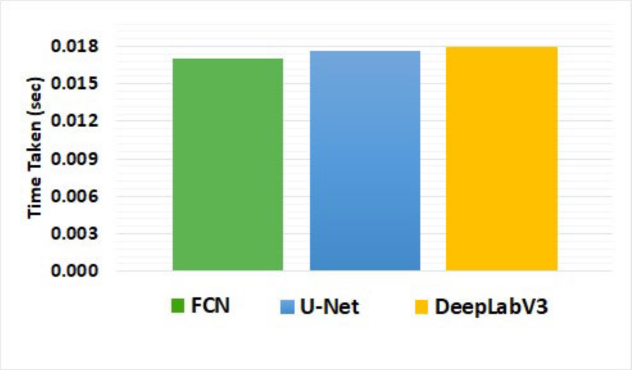

FIGURE 15. Inference time of FCN, U-Net and DeepLabV3 on GPU.

We have used an NVIDIA GTX 1080 Ti GPU for comparison and found that both models take around 1.2GB. From Figure.14 and Figure.15, it can be noticed that the DeepLabV3 model is slightly slower than FCN and U-Net.

VI. CONCLUSION

In this research work, different deep learning-based semantic segmentation models FCN, U-Net, and DeepLabV3 have been explored for top view multiple person segmentation. Although there is substantial variation in the person appearance from the perspective change of camera, the pre-trained deep learning-based segmentation models still give encouraging results. In order to further enhance the performance of segmentation models, we further trained and tested all three models using the top view person data set. The visual results show that models have the ability to segment person even when occluded and partially visible in scenes captured from the top view. Overall the accuracy of DeeLabV3 and U.Net is better than FCN. The models achieve *IoU* of 83%, 84% and 86% and *mIoU* of 80%, 82%, and 84% for FCN, U-Net and DeepLabv3 respectively. We also calculated the computational performance of both models for CPU and GPU. The inference time shows that DeepLabV3 and U-Net are slightly slower than the FCN model. The deep learning-based segmentation models perform much better than conventional segmentation methods. In the future, the work might be extended for other deep learning-based segmentation models using multiple top view object data set.

REFERENCES

[1] X. Liu, Z. Deng, and Y. Yang, "Recent progress in semantic image segmentation," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1089–1106, Aug. 2019.

[2] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 115–134, Apr. 2019.

[3] S. Saito, T. Li, and H. Li, "Real-time facial segmentation and performance capture from RGB input," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 244–261.

[4] C. C. Kaymak and A. Uçar, "A brief survey and an application of semantic image segmentation for autonomous driving," in *Handbook of Deep Learning Applications*. Springer, 2019, pp. 161–200.

[5] Y. Yoon, H.-G. Jeon, D. Yoo, J.-Y. Lee, and I. S. Kweon, "Learning a deep convolutional network for light-field image super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 24–32.

[6] C. Couprie, C. Farabet, L. Najman, and Y. LeCun, "Indoor semantic segmentation using depth information," 2013, *arXiv:1301.3572*. [Online]. Available: http://arxiv.org/abs/1301.3572

[7] Z. Zhuang, N. Lei, A. N. Joseph Raj, and S. Qiu, "Application of fractal theory and fuzzy enhancement in ultrasound image segmentation," *Med. Biol. Eng. Comput.*, vol. 57, no. 3, pp. 623–632, Mar. 2019.

[8] H. R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "Deep learning and its application to medical image segmentation," *Med. Imag. Technol.*, vol. 36, no. 2, pp. 63–71, Mar. 2018.

[9] J. Svensson and J. Atles, "Object detection in augmented reality," M.S. thesis, Math. Sci., 2018.

[10] Y. Wang, Y. Haichao, D. Gao, and J. Wang, "Image segmentation and object detection using fully convolutional neural network," U.S. Patent 10 304 193, May 28, 2019.

[11] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 2, pp. 87–93, Jun. 2018.

[12] D. Divya and T. G. Babu, "A survey on image segmentation techniques," in *Proc. Int. Conf. Emerg. Current Trends Comput. Expert Technol.* Springer, 2019, pp. 1107–1114.

[13] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[14] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2843–2851.

[15] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 297–312.

[16] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[17] 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.

[18] J. Dong, Q. Chen, S. Yan, and A. Yuille, "Towards unified object detection and semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 299–314.

[19] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Frontiers Neurorobotics*, vol. 12, p. 64, Oct. 2018.

[20] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Springer, 2015, pp. 234–241.

[22] 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.

[23] M. Ahmad, I. Ahmed, K. Ullah, I. Khan, A. Khattak, and A. Adnan, "Person detection from overhead view: A survey," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 4, pp. 567–577, 2019, doi: 10.14569/IJACSA.2019.0100470.

[24] I. Ahmed, M. Ahmad, A. Adnan, A. Ahmad, and M. Khan, "Person detector for different overhead views using machine learning," *Int. J. Mach. Learn. Cybern.*, vol. 10, pp. 2657–2668, May 2019.

[25] M. Ahmad, I. Ahmed, K. Ullah, I. Khan, A. Khattak, and A. Adnan, "Energy efficient camera solution for video surveillance," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 3, pp. 1–8, 2019, doi: 10.14569/IJACSA.2019.0100367.

[26] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Comput. Vis.* Springer, 2014, pp. 740–755.

[28] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, ''A survey on deep learning techniques for image and video semantic segmentation,'' *Appl. Soft Comput.*, vol. 70, pp. 41–65, Sep. 2018.

[29] F. Garcia-Lamont, J. Cervantes, A. López, and L. Rodriguez, ''Segmentation of images by color features: A survey,'' *Neurocomputing*, vol. 292, pp. 1–27, May 2018.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, ''SSD: Single shot multibox detector,'' in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[31] 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.

[32] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[34] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, ''Deep clustering with sample-assignment invariance prior,'' *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2019, doi: 10.1109/TNNLS.2019.2958324.

[35] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, ''Structured autoencoders for subspace clustering,'' *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, ''Rich feature hierarchies for accurate object detection and semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[37] R. Girshick, ''Fast R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[38] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[39] A. Van Etten, ''You only look twice: Rapid multi-scale object detection in satellite imagery,'' 2018, *arXiv:1805.09512*. [Online]. Available: http://arxiv.org/abs/1805.09512

[40] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, ''BlitzNet: A real-time deep network for scene understanding,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4154–4162.

[41] J. Dai, K. He, and J. Sun, ''Instance-aware semantic segmentation via multi-task network cascades,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3150–3158.

[42] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, ''Understanding convolution for semantic segmentation,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.

[43] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, ''Fully convolutional instance-aware semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2359–2367.

[44] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, ''Attention to scale: Scale-aware semantic image segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[45] M. Ahmad, I. Ahmed, K. Ullah, I. Khan, and A. Adnan, ''Robust background subtraction based person's counting from overhead view,'' in *Proc. 9th IEEE Annu. Ubiquitous Comput., Electron. Mobile Commun. Conf. (UEMCON)*, Nov. 2018, pp. 746–752.

[46] K. Ullah, I. Ahmed, M. Ahmad, and I. Khan, ''Comparison of person tracking algorithms using overhead view implemented in OpenCV,'' in *Proc. 9th Annu. Inf. Technol., Electromech. Eng. Microelectron. Conf. (IEMECON)*, Mar. 2019, pp. 284–289.

[47] I. Ahmed, A. Ahmad, F. Piccialli, A. K. Sangaiah, and G. Jeon, ''A robust features-based person tracker for overhead views in industrial environment,'' *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1598–1605, Jun. 2018.

[48] I. Ahmed, M. Ahmad, M. Nawaz, K. Haseeb, S. Khan, and G. Jeon, ''Efficient topview person detector using point based transformation and lookup table,'' *Comput. Commun.*, vol. 147, pp. 188–197, Nov. 2019.

[49] I. Ahmed, S. Din, G. Jeon, and F. Piccialli, ''Exploring deep learning models for overhead view multiple object detection,'' *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5737–5744, Jul. 2020.

[50] M. Ahmad, I. Ahmed, F. A. Khan, F. Qayum, and H. Aljuaid, ''Convolutional neural network–based person tracking using overhead views,'' *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 6, 2020, Art. no. 1550147720934738.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ''Mask R-CNN,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[53] 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.

[54] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, ''Image segmentation using deep learning: A survey,'' 2020, *arXiv:2001.05566*. [Online]. Available: http://arxiv.org/abs/2001.05566

[55] Y. S. Salas, D. V. Bermudez, A. M. L. Peña, D. G. Gomez, and T. Gevers, ''Improving hog with image segmentation: Application to human detection,'' in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Springer, 2012, pp. 178–189.

[56] A. Bleau and L. J. Leon, ''Watershed-based segmentation and region merging,'' *Comput. Vis. Image Understand.*, vol. 77, no. 3, pp. 317–370, Mar. 2000.

[57] N. Otsu, ''A threshold selection method from gray-level histograms,'' *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[58] Z. Zivkovic, ''Improved adaptive Gaussian mixture model for background subtraction,'' in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2, Aug. 2004, pp. 28–31.

[59] P. Suo and Y. Wang, ''An improved adaptive background modeling algorithm based on Gaussian mixture model,'' in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 1436–1439.

[60] A. B. Godbehere, A. Matsukawa, and K. Goldberg, ''Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation,'' in *Proc. Amer. Control Conf. (ACC)*, Jun. 2012, pp. 4305–4312.

[61] R. H. Evangelio, M. Patzold, I. Keller, and T. Sikora, ''Adaptively splitted GMM with feedback improvement for the task of background subtraction,'' *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 5, pp. 863–874, May 2014.

IMRAN AHMED (Member, IEEE) received the B.Sc. degree in computer science and mathematics from the Edwardes College, the M.Sc. degree in computer science from the University of Peshawar, the M.S. degree in IT from IMsciences, Peshawar, Pakistan, with major research in computer vision, and the Ph.D. degree in computer science major from the University of Southampton, U.K. He is currently working as an Assistant Professor with the Institute of Management Sciences, Peshawar. His research interests include machine learning, data science, computer vision, feature extraction, digital image and signal processing, medical image processing, biometrics, pattern recognition, data mining, and deep learning. He has attended several national and international conferences in these areas. He has been acting as a Reviewer in journals, such as the IEEE Transactions on Industrial Electronics, IEEE Access, the *Journal of Ambient Intelligence*, and Elsevier.

MISBAH AHMAD received the B.S. degree in telecommunication systems from Islamia College University, Peshawar, Pakistan, in 2015, and the M.S.C.S. degree from the Institute of Management Sciences, Peshawar, in 2019. Her research interests include computer vision, image processing, machine learning, deep learning, and data science. She is involved as a Referee for many reputed international journals and conferences.

FAKHRI ALAM KHAN received the Ph.D. degree in computer science from the Institute of Scientific Computing, University of Vienna, Austria, in 2010. He is currently an Associate Professor with the Institute of Management Sciences, Peshawar, Pakistan. His research interests include scientific workflows provenance, energy efficiency in WSN, multimedia technologies, nature inspired meta-heuristic algorithms, and workflow parameters significance measurement.

MUHAMMAD ASIF received the M.S. and Ph.D. degrees from AIT, in 2009 and 2012, respectively, on the HEC Foreign Scholarship. He was a Research Scholar with the Computer Science and Information Management Department, Asian Institute of Technology, Thailand. During the course of time, he was a Visiting Researcher with the National Institute of Information Tokyo, Japan. He is currently the Chairman of the Department of Computer Science, National Textile University, Faisalabad, Pakistan. He has worked on some projects, including the Air Traffic Control System of Pakistan Air force. He is a permanent member of the Punjab Public Service Commission (PPSC) and an Advisor and a Program Evaluator at the National Computing Education Accreditation Council (NCEAC), Islamabad. He has been serving as a Reviewer of a number of reputed journals and also authored a number of research papers in reputed journals and conferences. He has been serving as an Associate Editor for IEEE Access and the prestigious journal of the IEEE.