IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 7, 2019, accepted September 8, 2019, date of publication September 12, 2019, date of current version October 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941026

Performance Analysis of Semantic Segmentation Algorithms for Finely Annotated New UAV Aerial Video Dataset (ManipalUAVid)

S. GIRISHA[1], (Member, IEEE), MANOHARA M. M. PAI[1], (Senior Member, IEEE), UJJWAL VERMA[2], (Senior Member, IEEE), AND RADHIKA M. PAI[1], (Senior Member, IEEE)

[1]Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India
[2]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India

Corresponding authors: Manohara M. M. Pai (mmm.pai@manipal.edu), Ujjwal Verma (ujjwal.verma@manipal.edu), and Radhika M. Pai (radhika.pai@manipal.edu)

This work was partially supported by IBM India research grant which was utilized for purchase of drone used in this project.

ABSTRACT Semantic segmentation of videos helps in scene understanding, thereby assisting in other automated video processing techniques like anomaly detection, object detection, event detection, etc. However, there has been limited study on semantic segmentation of videos acquired using Unmanned Aerial Vehicles (UAV), primarily due to the absence of standard dataset. In this paper, a new UAV aerial video dataset (ManipalUAVid) for semantic segmentation is presented. The videos have been acquired in a closed university campus, and fine annotation is provided for four background classes viz. constructions, greeneries, roads, and waterbodies. Also, the performance of four semantic segmentation approaches: Conditional Random Field (CRF), U-Net, Fully Convolutional Network (FCN) and DeepLabV3+ are analysed on ManipalUAVid dataset. It is seen that these algorithms perform competitively on UAV aerial video dataset and achieves an mIoU of 0.86, 0.86, 0.86 and 0.83 respectively.

INDEX TERMS Convolutional neural networks, semantic segmentation, shot boundary detection, UAV video.

I. INTRODUCTION

Semantic segmentation refers to the process of assigning a class label to each pixel in an image, enabling a high-level description of the image. It has been used for various applications such as autonomous driving [1], medical image processing [2], photo editing etc. The availability of affordable priced UAV has led to an increased focus on utilizing UAV for surveillance oriented applications. As compared to traditional CCTV based approaches which would require multiple CCTVs to monitor a particular area, a single UAV can monitor a larger area. Moreover, UAV can be rapidly deployed and is particularly advantageous to be used for temporary major events such as a concert or a marathon [3]. However, there is no standard dataset for semantic segmentation of videos acquired using Unmanned Aerial Vehicles (UAV). Although there have been several studies analysing UAV videos, most of these studies have focussed on object detection [4], object tracking [5] and human action recognition [3]. Typically, the semantic segmentation algorithms are evaluated on standard datasets such as PASCAL [6], COCO [7] etc, containing generic scenes. In addition, there exists dataset such as KITTI [8], cityscapes [9] etc., focussing on complex urban street scenes. To the best of our knowledge, there is no standard dataset for semantic segmentation of scene on images acquired using UAV.

A typical outdoor scene as viewed from a UAV consists of multiple objects (such as people, vehicle, etc.) interacting with each other or with their environment (roads, building, etc). This work is focussed on semantic segmentation of scene background (roads, building, etc.) as the definition of foreground objects could be application-specific. Moreover, the semantic segmentation of scene background can assist in other autonomous tasks such as anomalous activity detection, action recognition, and other high-level aerial video understanding by providing spatial information.

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

The existing UAV datasets are designed for specific applications such as tracking [5], anomaly activity recognition [3], action recognition [10] etc. These datasets are acquired at various locations such as college campus, parking lot, or open stadium. The creators of these datasets have focussed on tracking/detection/recognition of foreground objects such as a person, cars, etc. and subsequent high-level tasks such as human-object interaction, anomalous activity recognition etc. However, the background in these datasets is limited to a few locations. For instance, the background in the EPFL mini drone dataset [3] contains roads, building and lawns only, while okutama action dataset [10] contains baseball field and greenery as the background. Also, Stanford dataset [5] contains road, lawns, and buildings as the background. Therefore, labelling the existing UAV dataset would limit the background labels. As a result, a new UAV dataset is created by grouping the different background elements into four generic categories (viz. greenery, construction, road and water body). Although the data is acquired at the institute campus, the data is general enough to represent a typical background of an outdoor image acquired using UAV. Even though the Stanford Drone dataset [5] contains semantic labels for the background to study object-space interaction, the labels are limited to road, roundabout, sidewalk, grass, building, and bike rack.

Typically, probabilistic models such as Conditional Random Fields (CRF) are utilized to incorporate correlation information between pixels for semantic segmentation. In recent years, Convolutional Neural Network (CNN) has been widely used for semantic segmentation [2], [11], [12]. To perform semantic segmentation, the common CNN architecture such as VGG [13], AlexNet [14] is modified by converting last fully connected layers into convolutional layers. The ability of Convolutional Neural Network to capture deeply learned features and context information has made it a prominent choice for semantic segmentation. Recently, a combination of CRF and CNN have achieved excellent performance in semantic segmentation as demonstrated in public benchmark such as Pascal VOC [6].

The key contributions of this work are as follows:

- A new dataset (ManipalUAVid) is developed for semantic segmentation of UAV aerial videos by acquiring aerial videos from a UAV in a closed university campus.
- Performance evaluation of four different semantic segmentation approaches (CRF and three CNN architectures) are done on this new UAV video dataset.

The ManipalUAVid dataset presented in this paper is an extended version of the dataset presented in the earlier work [15]. This updated ManipalUAVid covers more locations and also contains more videos. Moreover, four class annotation is presented for the videos as compared to two class annotation in [15]. Besides, the performance evaluation of only CNN based method was presented in [15], whereas in this work the performance of four different semantic segmentation approaches (CRF, and three CNN architectures) are compared.

This paper is organized as follows: Section 2 presents the recent developments in semantic segmentation and also summarizes the existing datasets. Section 3 describes the acquisition protocol and labelling policy for the new UAV aerial video dataset. Section 4 presents the methodology for semantic segmentation of UAV aerial videos. Results obtained on the UAV aerial video dataset and a detailed discussion of the same are presented in Section 5. Finally, the conclusion of the work is presented in Section 6.

II. RELATED METHODS

Accuracy of semantic segmentation depends on individual pixel and its correlation with the neighbouring pixels. The ability of CRF and CNN techniques to capture the relationship between a pixel and its neighbours have made it a popular choice for semantic segmentation. This section summarizes the two approaches for semantic segmentation (CRF, CNN). A more detailed description can be found in [16]. In addition, a brief overview of the existing dataset for semantic segmentation is presented in this section.

CRFs, a variant of Markov Random Field (MRF), consists of clique potentials which are conditioned on input features [17]. In literature, CRFs are widely studied for semantic segmentation and scene understanding because of their ability to capture spatial information [18]–[20]. Different handcrafted features like colour, texture, edge etc. can be encoded as potential energy in CRF model [21]. In addition, higher-order potentials are often employed [22] to improve segmentation accuracy. CRF post-processing improves segmentation accuracy. However, learning the CRF is time-consuming as it requires repeated inference steps [23].

CNN is a modification of multilayer perceptron and is widely used for object recognition, image classification, tracking [2], [11]–[14], [24] etc. A typical CNN architecture for semanctic segmentation consists of encoding layers which are used to extract features and decoding layers to infer the class labels [2], [11], [12]. Some examples of this encoder-decoder architecture are U-net [2], Segnet [12] etc. In addition, there also exist other CNN architectures such as DeepLab [11] where the filters were modified to have dilations capturing the increased spatial information with a reduced number of parameters. In another approach called Fully Convolutional Network, the popular CNN architectures (AlexNet [14], VGG net [13], GoogLeNet [25]) has been modified for semantic segmentation by replacing the last fully connected layers with convolutional layers [26].

Several authors have explored combining both CRF and CNN for semantic segmentation [27]–[31]. These algorithms have the ability to map complex input features to output segmentation map by learning deep features from CNN and uses CRFs to refine the output by modelling the interactions of output variables.

In this work, the performance of CRF, U-Net, FCN and DeepLabV3+ approaches for semantic segmentation are analysed on the new UAV aerial videos dataset (Figure 1).

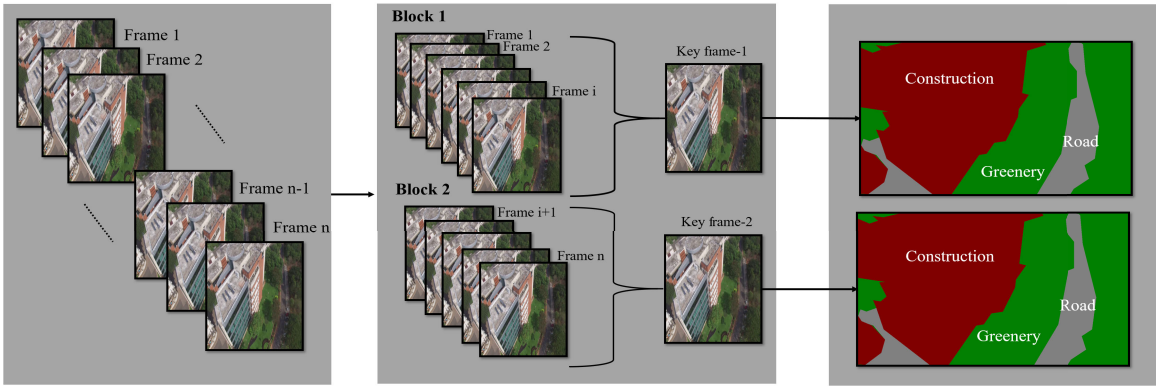


FIGURE 1. Overview of the semantic segmentation method for UAV aerial videos.

Existing Semantic Segmentation Datasets: There are several popular semantic segmentation datasets like MS COCO [7], Pascal VOC [6], CamVid [32], KITTI [8] and CityScape [9]. MS COCO focuses on semantic segmentation of general scenes with class labels as person, car, animal etc. Pascal VOC provides a benchmark and consists of classes like bus, car, cow etc. KITTI is semantic segmentation benchmark, which consists of common city scenes. Cityscape is a benchmark suite consisting of various target classes such as a human, vehicle, construction etc, for urban street scenes. CamVid is a video dataset with 32 target classes captured from the perspective of a driving vehicle. A more exhaustive review of the existing dataset can be found in [16]. It is observed that there is no standard dataset for semantic segmentation of UAV aerial videos. There are some studies on semantic segmentation of aerial images but these are limited to SAR images [33]. Also, popular UAV aerial video datasets like Stanford dataset [5] focus on multiclass target tracking, or anomaly detection [3], which lacks the annotations for semantic segmentation. Recently, a new data set was proposed for semantic segmentation of SAR images [17]. This SAR image dataset contains annotations for road, water, built-up area, and vegetation. However, semantic segmentation of SAR image is a challenging task due to the presence of speckle noise and scattering phenomena. An alternative approach to identify the road, water-bodies, built-up area, etc. is by analyzing UAV aerial videos. Indeed, semantic segmentation of UAV aerial videos can be utilized for a more finer analysis of the region.

III. MANIPALUAVID: MANIPAL UAV AERIAL VIDEO DATASET

Analyzing UAV aerial video has been limited to detection and tracking of objects [5]. In this work, a new UAV aerial video dataset is developed for semantic segmentation. This dataset is named as Manipal UAV aerial video (ManipalUAVid) dataset after the location where the videos have been collected. The purpose of this dataset is to serve for the semantic segmentation on UAV aerial images enabling better scene understanding.

TABLE 1. Summary of the ManipalUAVid dataset.

Total number of videos	33
Minimum duration	30 sec
Maximum duration	12 min
Total number of keyframes	667
Number of class considered for annotations	4 (Road, Construction, Greenery and Water Bodies)
Image resolution	1280x720 pixels
Frames per second	29
Approximate altitude	20-30 mts

TABLE 2. Location details of the ManipalUAVid dataset.

Locations	Number of videos	Number of key frames
Library	1	44
Innovation centre	13	208
Canteen	2	54
Student plaza	2	33
Hostel	4	117
Swimming pool	2	51
Approach road to innovation centre	3	57
Approach road to AB1 block	6	103

UAV aerial videos are captured by using DJI Phantom 3 Professional drone with 1280 $\times$ 720 resolution at 29 frames per second. The videos are captured in the campus of Manipal Institute of Technology, Manipal, India at an approximate altitude of 25 meters. Videos are captured from six different locations such as parking lot, swimming pool, library, academic blocks, etc. at different timings during the day. The videos have been acquired under different weather conditions such as sunny, cloudy and after rains, etc. In total, 33 videos have been acquired with the minimum duration of a single video being about 30 seconds and the maximum duration of 12 minutes. Out of 33 videos, 11 videos were acquired in morning, 14 videos in the afternoon and 8 videos in the evening. A summary of the data set is shown in Table 1 and 2.

TABLE 3. Class details of ManipalUAVid dataset.

Class	Objects considered under the class	Percentage of pixels	Number of Pixels
Greenery	Grass, Trees, Ground	41.4	231,596,236
Road	Roads, Parking, Vehicles, Pedestrians, Side Walk	51.5	288,096,768
Construction	Constructions, Compounds, Wall, Fence	5.82	32,538,152
Water Body	Swimming pools	1.28	7,075,992

In this work, four classes of objects are considered for semantic segmentation (Greenery, Road, Construction and Waterbody). Various objects grouped under particular class is shown in Table 3. Labelling individual pixels (fine annotations, the annotations are provided by domain experts with the help of LabelMe annotation tool) is a challenging and time-consuming task. In case of video, fine and/or coarse annotation for each frame would be redundant. Indeed, a small number of finely labelled data achieves the same performance for semantic segmentation as compared to a large number of coarse labelled data [34]. In continuation of the standard practice followed for video semantic segmentation dataset, annotations are provided for certain frames (keyframes) in the video in ManipalUAVid. While labelling the dataset, "Foreground class must not have any hole" policy is followed [9]. Moreover, labelling individual pixels at the boundary of two classes (especially trees) of objects is a challenging task because of the ambiguous object boundary. Few original frames and corresponding ground truth masks are shown in Figure 2. The complete dataset is available at https://github.com/uverma/ManipalUAVid.

IV. SEMANTIC SEGMENTATION OF UAV AERIAL VIDEOS: CRF AND CNN

This section describes the methodology followed for semantic segmentation of UAV aerial videos. As discussed earlier, four approaches (CRF, U-Net, FCN and DeepLabV3+) are studied for semantic segmentation of UAV aerial videos. Analyzing each frame would be time-consuming and redundant, therefore a keyframe is identified using shot boundary (Section IV-A). Subsequently, semantic segmentation is performed on these keyframes using CRF (Section IV-B), U-Net (Section IV-C), FCN (Section IV-D) and DeepLabV3+ (Section IV-E).

A. SHOT BOUNDARY DETECTION

Processing every frame of a video at 29 fps is redundant and time-consuming. Hence it is advantageous to process keyframes that represent a group of frames in a video. Keyframes are identified by using the shot boundary detection algorithm (Figure 1). To compute shot boundary each colour frame of the video is divided into a non-overlapping grid of size 16 × 16. The histogram difference is computed between two corresponding windows of two consecutive frames by using Chi-square distance. These histogram differences of all corresponding windows of two consecutive frames are averaged to find the average histogram difference between two consecutive frames which is given as follows,

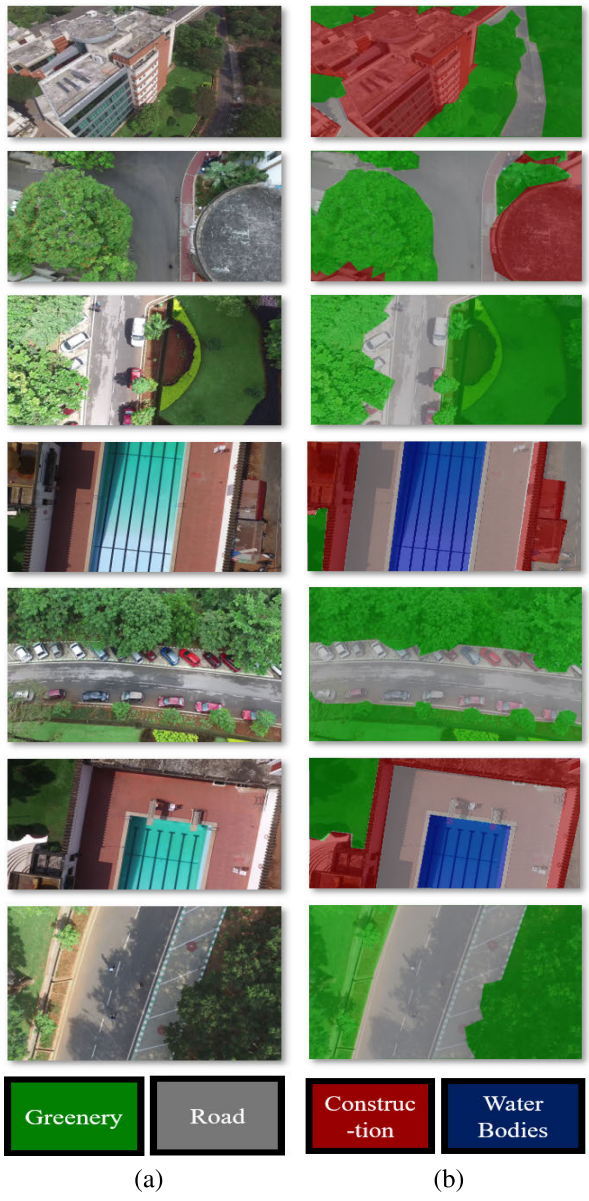


FIGURE 2. Few images of UAV aerial video dataset along with corresponding four class annotation.

$$D_i = \frac{1}{N}\sum_{p=1}^{N}\frac{(H_i(I_p) - H_{i+1}(I_p))^2}{H_i(I_p)} \quad (1)$$

where $H_i(I_p)$ represents the histogram of p^{th} image patch I_p of i^{th} frame. Similarly, $H_{i+1}(I_p)$ represents the histogram of p^{th} image patch I_p of $(i+1)^{th}$ frame, D_i represents the average histogram difference between i^{th} and $(i+1)^{th}$ consecutive image frames and N represents the total number of grids in an image. For an image of size 1280 × 720, 3600 grids are obtained. This difference D_i is calculated for every pair of consecutive frames and shot boundary is identified as

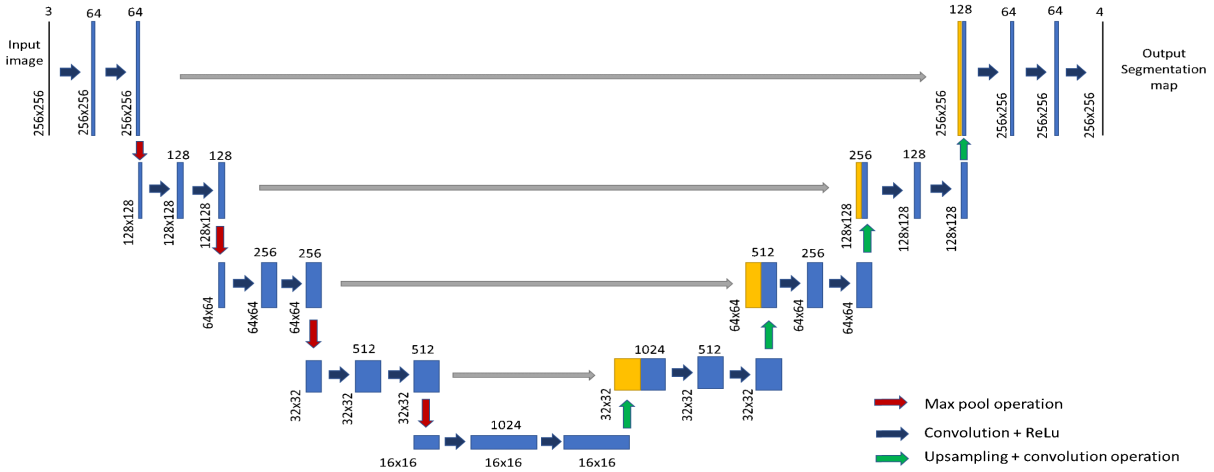


FIGURE 3. The modified U-Net architecture [15]: Contracting path followed symmetric expanding path.

follows.

$$\text{Shot boundary} = \begin{cases} True & if\ |D_i - D_{i+1}| > T_{shot} \\ False & otherwise \end{cases} \quad (2)$$

where T_{shot} is determined experimentally. The shot boundary is identified as the frame where $D_i - D_{i+1}$ is greater than the threshold. All frames within one shot are known as a shot and this shot is represented by the middle frame (key frame) in that shot.

B. CRF: TextonBoost

Conditional Random Field (CRF) is a probabilistic approach for semantic segmentation incorporating the correlation among the pixels. The energy of CRF consists of unary potential energy $\varphi(x_i)$ and a pairwise potential energy $\lambda(x_i, x_j)$.

$$E(x) = \sum_i \varphi(x_i) + \sum_{i,j} \lambda(x_i, x_j) \quad (3)$$

where, x_i represents the i^{th} input pixel. The unary potential energy captures features which are local to pixel itself. Pairwise potential energy acts as a smoothness term adding penalties based on neighbouring pixels.

One of the widely used CRF approaches utilizes a novel feature called ''texture-layout filter'' which incorporates texture, context and layout information [21]. This feature response is computed by first generating a texton map. This map is generated by convolving the image with a 17-dimensional filter bank at varying scale and then applying K-means clustering algorithm. Subsequently, for each texton, t, the area in the rectangle mark that matches t is computed. This process is repeated for each texton and rectangular mask, thus generating a texton histogram of area response for the rectangular region, texton pair. Finally, an adapted version of the joint boost algorithm is used to compute the texture layout potential.

In addition to the texture layout filter response, two unary potentials (colour and location) and one pairwise potential (edge) are also considered to refine the results. The parameters for each of these potentials are learned independently. The aim of the model is to identify the most probable class label for a given pixel. Once the parameters are learnt, the labels are inferred by using alpha-expansion graph-cut algorithm. More details about this approach can be found in [21].

C. CNN: U-Net

One of the popular CNN architecture for semantic segmentation is U-Net [2], which derives its name from its U shaped architecture (the contracting path followed by symmetric expanding path). The contracting path is similar to a typical convolutional neural network architecture viz. convolution operation followed by ReLU activation function and then max pooling. The expansive path consists of deconvolution layers, flowed by concatenation with corresponding cropped feature map from the contracting path. Subsequently, two consecutive 3×3 convolution operation followed by a ReLU activation function is applied. The network relies on data augmentation and gradient descent for training. The main advantage of the U-Net architecture is that it is not dependent on large dataset. However, due to the unpadded convolutions, the output image is smaller than the input by a constant border width. At the last layer, SoftMax activation is used to obtain the probability distribution of each pixel to every class.

The U-Net architecture proposed in [2] was designed to process the grey scale image of size 572×572. In the proposed work, the existing U-Net architecture is modified to process colour image of size 256×256 by using three filters for each color channels at the input layer (Figure 3). In addition, the last layer is modified for multiclass classification. Note that the padding is considered for each layer by assuming that values outside the bound are the same as the boundary pixels.

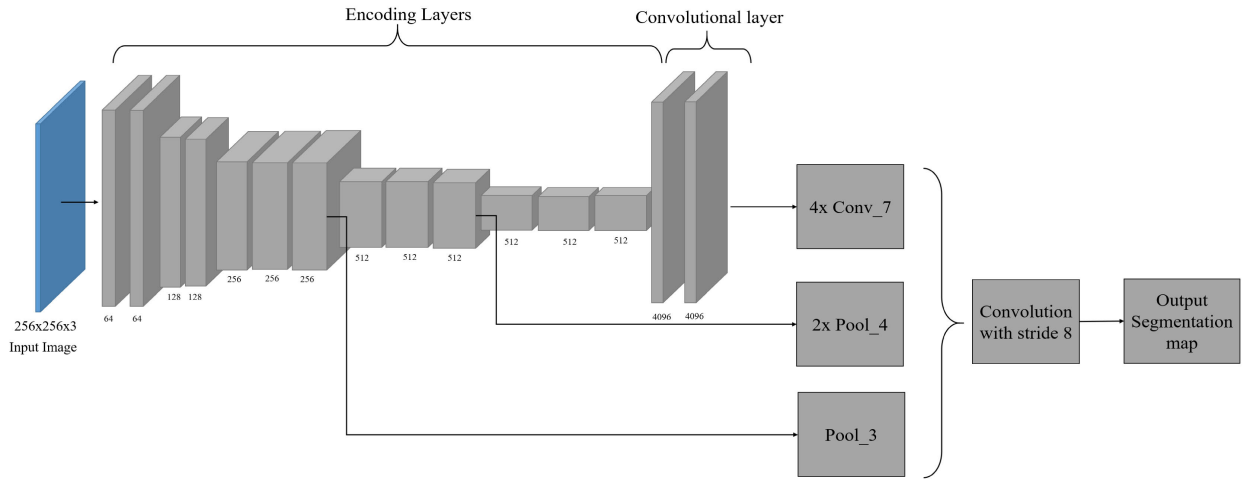


FIGURE 4. The FCN-8 architecture [20].

D. CNN: FCN-8 WITH VGG16 BACKBONE

Fully convolutional neural network [20] is a widely used semantic segmentation algorithm which adopts classification networks like AlexNet [14], the VGG net [13], and GoogLeNet [25]. These classification networks are used as feature extractor and the last fully connected layers are replaced by fully convolutional layers to obtain segmentation output. Different versions of FCN exists such as FCN-8, FCN-16 and FCN-32. All these three FCN architecture uses same downsampling path but differs in their respective upsampling path. The downsampling path is used to recover semantic/contextual information while the upsampling path recovers the spatial information.

In the present study, FCN-8 is used for semantic segmentation with VGG16 (Figure 4) as the backbone architecture because of its simple structure. The downsampling path is similar to the VGG-16 architecture except for the last fully connected layers. In FCN-8, the dense layer of VGG16 is replaced with fully convolutional layer. To obtain the output image of the same size as the input image, segmentation map is obtained by using transposed convolution layer with stride 8 at the last layer. The FCN-8 architecture is shown in Figure 4.

E. CNN: DeepLabV3+

DeepLabV3+ [35] is another encoder-decoder based CNN architecture that achieved state-of-the-art semantic segmentation performance on Pascal VOC and CityScape dataset. The earlier iteration DeepLabV3 uses several parallel atrous convolutions with different rates, thus capturing multiscale contextual information. DeepLabV3+ utilizes the output of DeepLabV3 as the encoder output. In addition, DeepLabV3+ consists of a decoder module which improves the segmentation at object boundaries.

The encoder module for DeepLabV3+ consists of the last ResNet block which is duplicated and arranged in cascade. Also, image-level features are included in atrous spatial pyramid module, consisting of atrous convolution of different rates. The last feature map of DeepLabV3 (before logits) is used as the encoder output for DeepLabV3+.

The encoder feature maps are upsampled bilinearly and concatenated with the corresponding low-level feature maps from the backbone network. Subsequently, 1×1 convolution is applied to reduce the number of channels. Then, 3×3 convolution is applied followed by bilinear upsampling. More details about DeepLabV3+ can be found at [35]. The architecture of DeepLabV3+ is shown in Figure 5.

V. RESULTS AND DISCUSSION

As discussed earlier, a keyframe is first identified, and then the four semantic segmentation approaches (CRF, U-Net, FCN-8, DeeplabV3+) are applied on these keyframes. This section discusses the results obtained for semantic segmentation of UAV aerial videos. For this study, the dataset was manually split into training, validation and test images so as to capture variations of different scenes. This manual split ensured that each split contains data from all the locations under different conditions. The use of this manual split instead of a random split is in accordance with the policy adapted for standard video semantic segmentation datasets [9]. In total, the training split contains 535 images, while validation and test split contains 66 and 66 images respectively. The performance of these algorithms are evaluated by using mean Intersection over Union (mIoU), Pixel Accuracy (PA), mean Pixel Accuracy (mPA), Precision, Recall and F1-score, by comparing it with the ground truth, as defined below.

$$mIoU = \frac{\sum_i x_{ii}}{C(\sum_i \sum_j x_{ij} + \sum_j x_{ji} - x_{ii})} \tag{4}$$

$$PA = \frac{\sum_i x_{ii}}{\sum_i \sum_j x_{ij}} \tag{5}$$

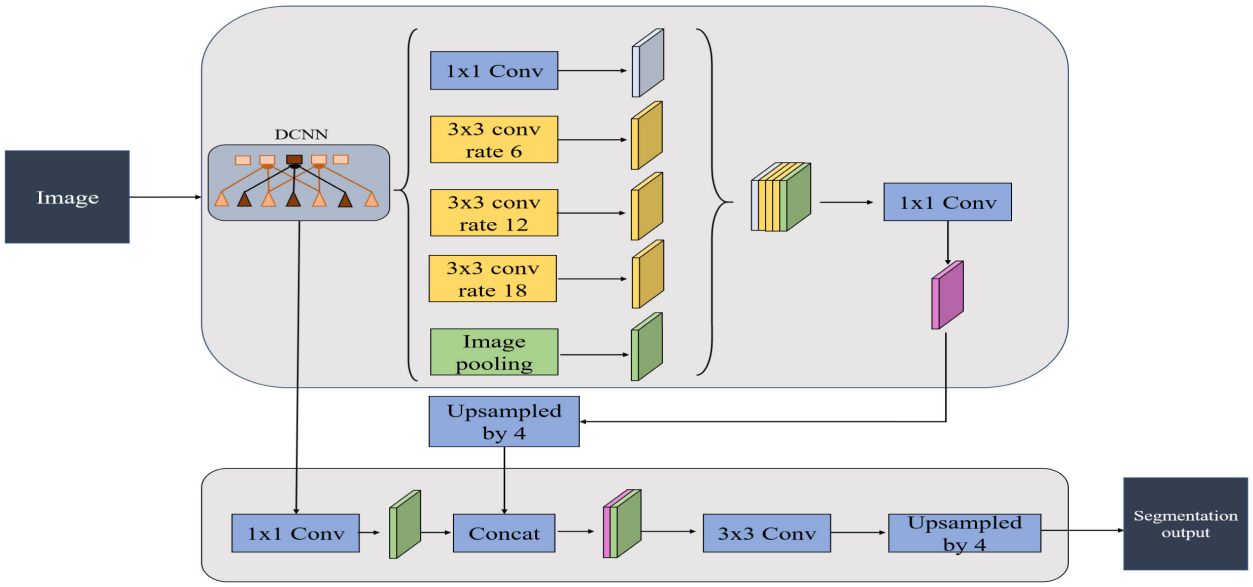


FIGURE 5. The DeepLabV3+ architecture [35].

$$mPA = \frac{1}{C}\frac{\sum_i x_{ii}}{\sum_j x_{ij}} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

where C is the number of classes (four in this study). x_{ij} represents the number of pixels belonging to class i and predicted as class j. TP, FP and FN are true positives, false positives and false negative respectively. All these models were trained on Intel Xeon Silver 4110 CPU clocked at 2.10 GHz with 32 GB RAM and Nvidia GeForce GTX 1080Ti GPU.

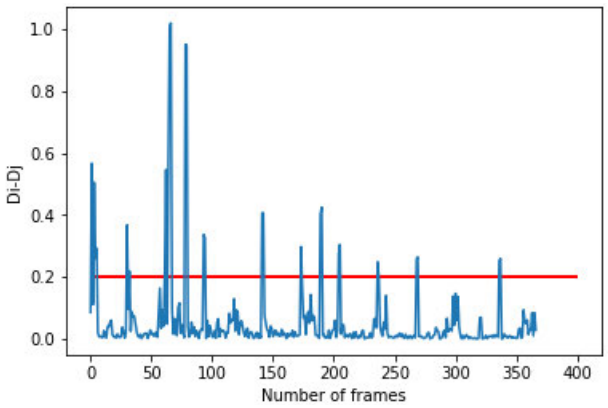


FIGURE 6. Variation of histogram difference ($D_i - D_{i+1}$) for a particular video.

A. SHOT BOUNDARY

The shot boundary is identified by comparing the histogram difference between two consecutive frames with a threshold T_{shot} (Section:IV-A). Figure 6 shows the variation of the histogram difference ($D_i - D_{i+1}$) for a particular video. The peak at regular interval signifies the presence of a significant change in the two consecutive frames. T_{shot} value is experimentally determined to be 0.2 as shown in Figure 6. If the histogram difference is greater than T_{shot}, a shot boundary is identified. All the frames in between two shot boundaries is called as a shot.

B. CRF: TextonBoost

The CRF based approach [21] which includes the texture, colour, location and edge potential is applied for semantic segmentation of keyframes identified. The parameters (the number of textons, the number of boosting rounds) for this model are estimated experimentally as explained below. In this study, the number of textons for the texture layout feature is set to different values and the corresponding pixel-wise segmentation accuracy is calculated on the test set images (Figure 7). It is observed that the segmentation accuracy increases as the number of textons increases to $k = 300$ and subsequently decreases. As a result, the value of number textons is set to $k = 300$. Another parameter in this model is the number of boosting rounds used in joint boost algorithm for computing texture layout potential. The effect of varying boosting round on pixel-wise segmentation accuracy is analysed and is shown in Figure 7. It is seen that the segmentation accuracy increases until the number of boosting rounds equals to 700 and decreases later on. In this work, the number of boosting rounds is set to 700.

To evaluate the performance of semantic segmentation precision, recall, F1-score, class confusion matrix, mIoU, PA, mPA and ROC curves are computed. The precision,

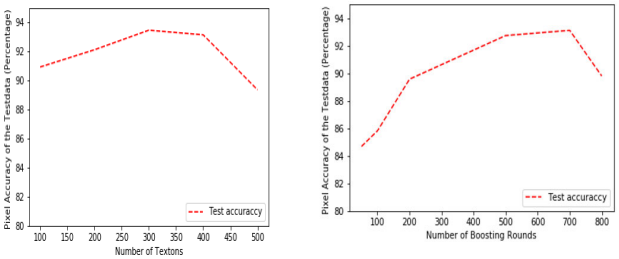


FIGURE 7. Variation of pixel wise segmentation accuracy with respect to number of textons and boosting rounds.

TABLE 4. Precision, Recall and F1-score of TextonBoost algorithm on ManipalUAVid.

Algorithms	Precision	Recall	F1-Score
Greenery	0.90	0.89	0.89
Road	0.94	0.94	0.93
Construction	0.82	0.82	0.82
Water bodies	0.93	0.93	0.93

TABLE 5. PA, mPA and mIoU of TextonBoost, U-Net, FCN-8 and DeepLabV3+ algorithms on ManipalUAVid.

Algorithms	PA	mPA	mIoU
TextonBoost	0.90	0.89	0.86
U-Net	0.93	0.92	0.86
FCN-8	0.92	0.93	0.86
DeepLabV3+	0.91	0.90	0.83

recall and F1-score obtained on the ManipalUAVid dataset is shown in Table 4. In addition, metrics such as mIoU, PA and mPA are shown in Table 5. It is observed that the average F1-score of the CRF based approach is 0.892. This approach classifies water bodies and roads accurately with F1-score of 0.93 followed by greeneries. However, a slightly low F1-score for construction class is obtained as few pixels belonging to construction class is misclassified into road and greenery class as shown in the confusion matrix (Figure 8). The ROC curve of the system is shown in Figure 10.

It can be observed that the water bodies (swimming pool) is accurately segmented compared to other classes due to the distinct colour and texture features (Figure 9). Moreover, the CRF model obtains an accurate segmentation while capturing the fine details of the greenery class. Indeed, the foliage (leaves) are generally scattered and sparse, but CRF model is able to capture these fine details as shown in Figure 9. However, a poor segmentation is obtained in case of variation in colour or texture of roads such as wet roads, presence of shadow etc. as shown in Figure 9.

C. CNN: U-Net RESULTS

The U-Net architecture is utilized for semantic segmentation of keyframes identified. The model is trained from scratch and no transfer learning is used. The batch size is fixed to 5 because of memory constraints. The training

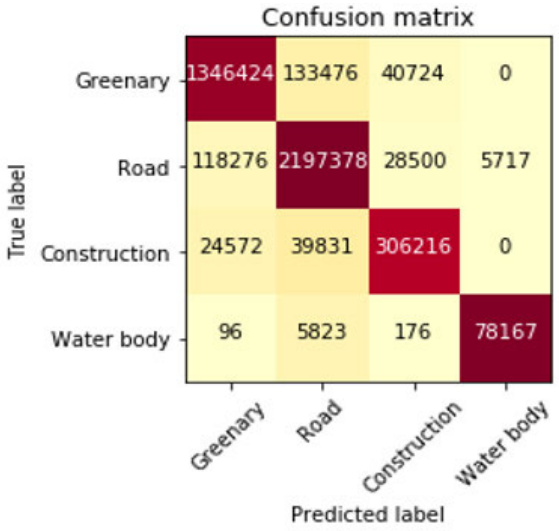


FIGURE 8. Confusion matrix for TextonBoost classifier for ManipalUAVid with four classes.

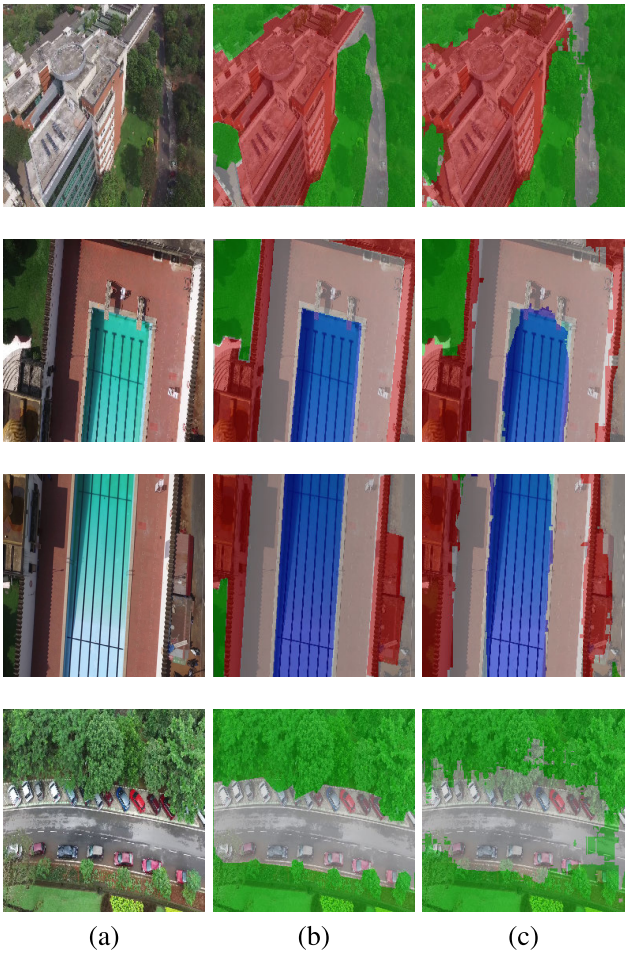


FIGURE 9. (a) Original images. (b) Ground truth images. (c) Semantically segmentated images using TextonBoost algorithm. Last row shows the poor results of TextonBoost algorithm.

procedure is not dependent on data augmentation as in [2], due to the availability of sufficient training samples.

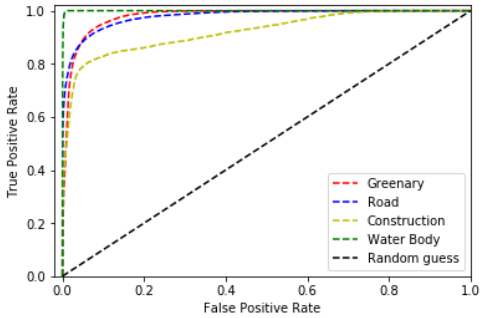
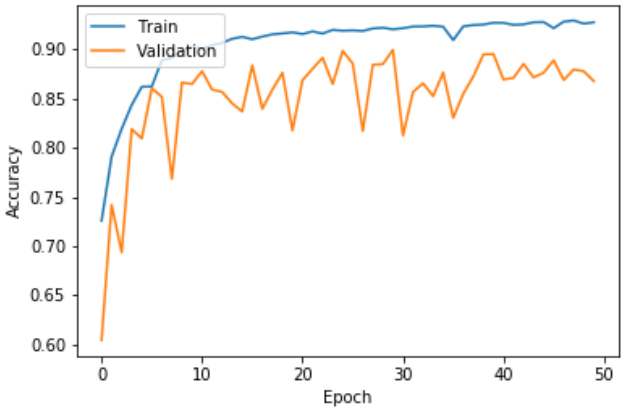
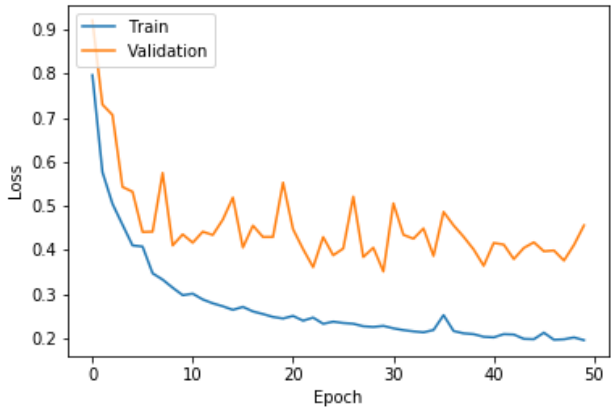


FIGURE 10. ROC curves obtained for TextonBoost algorithm.



(a)



(b)

FIGURE 11. (a) Accuracy and (b) loss curve of U-Net architecture for semantic segmentation of ManipalUAVid.

Categorical cross entropy is used as the loss function and weights are initially assigned using the normal distribution [36].

The model is trained for 50 epochs. The loss and accuracy curve for training and validation set are shown in Figure 11. Dropout regularization is used to avoid overfitting of the model.

The performance metrics (precision, recall and F1-score) of U-Net model on ManipalUAVid dataset is shown

TABLE 6. Precision, Recall and F1-score of U-Net on ManipalUAVid.

Algorithms	Precision	Recall	F1-Score
Greenery	0.95	0.91	0.93
Road	0.96	0.96	0.96
Construction	0.70	0.85	0.77
Water bodies	0.91	0.99	0.95

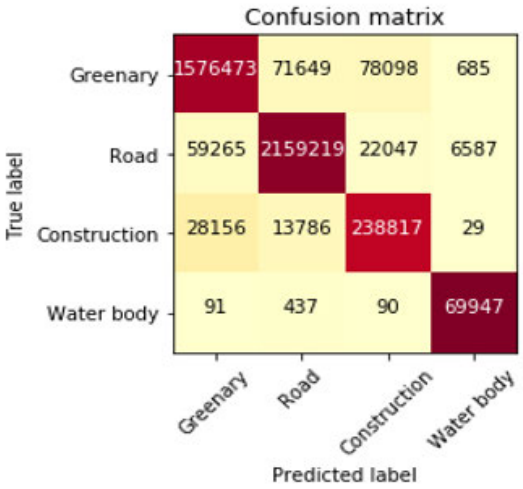


FIGURE 12. Confusion matrix for U-Net architecture for four class semantic segmentation of ManipalUAVid.

in Table 6. In addition, mPA, PA and mIoU for U-net architecture is shown in Table 5. A high F1-score is obtained for three classes viz. greenery, road and water bodies. However, a low F1-score of 0.77 is obtained for construction class. This low F1-score for the construction class is primarily due to low precision (misclassification of road and greenery pixels as construction pixels as shown in the confusion matrix in Figure 12). The ROC curve for the system is shown in Figure 14. The ROC curve for water bodies has the largest area under its curve which is followed by the curve for road, curve for greenery and lastly curve for construction class.

Some semantic segmentations of UAV aerial videos obtained using U-Net architecture are shown in Figure 13. It can be seen that U-Net architecture is able to identify the pixels belonging to four classes. However, the U-Net architecture fails to capture the fine details as compared to the CRF based approach. This substantiates the previous finding in the literature. Moreover, in some images, the road pixels are misclassified as constructions (bottom row in Figure 13). This misclassifications is primarily due to similar colour/textures features.

D. CNN: FCN RESULTS

Semantic segmentation is performed on keyframes identified in Section V-A by using FCN-8. No transfer learning is adopted and initial weights are assigned based on normal distribution [36]. Categorical cross entropy is used as the loss function. The model is trained for 35 epochs. Model overfitting is prevented by the usage of dropout layers.

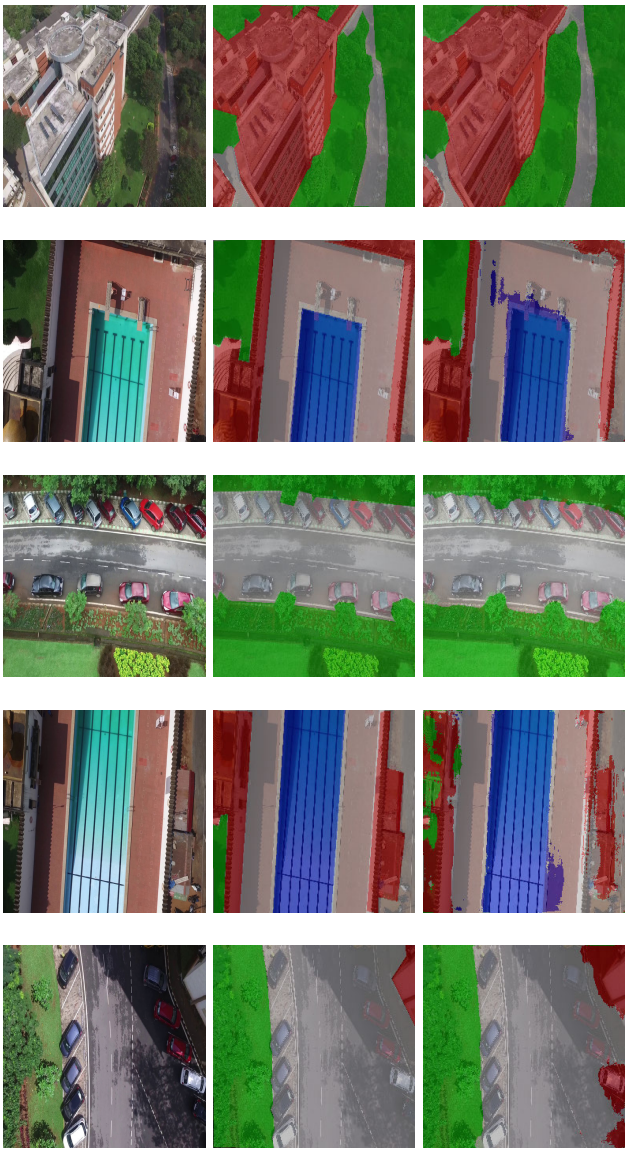


FIGURE 13. (a) Original image. (b) Ground truth image. (c) Semantic segmentation results by applying U-Net.

TABLE 7. Precision, Recall and F1-score of FCN-8 on ManipalUAVid.

Algorithms	Precision	Recall	F1-Score
Greenery	0.93	0.91	0.92
Road	0.96	0.95	0.95
Construction	0.71	0.91	0.80
Water bodies	0.94	0.97	0.95

The respective loss and accuracy curve for training and validation set is shown in Figure 15.

The different performance metrics such as precision, recall and F1-score for FCN-8 architecture on ManipalUAVid dataset is shown in Table 7. Other metrics such as mPA, PA and mIoU for FCN-8 are shown in Table 5. The ROC curve of FCN-8 is shown in Figure 18.

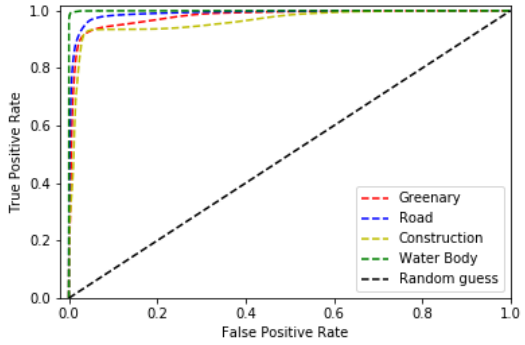


FIGURE 14. ROC curves obtained for U-Net architecture for semantic segmentation of ManipalUAVid.

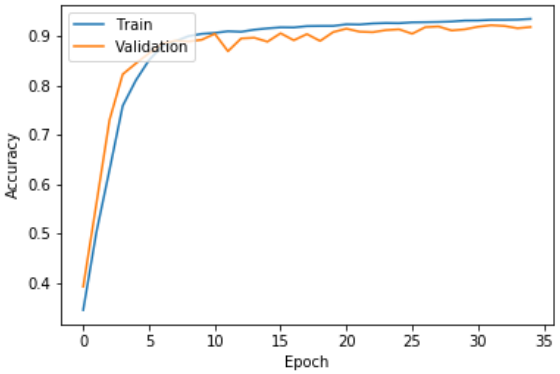


FIGURE 15. (a) Accuracy and (b) loss curve of FCN-8 architecture for semantic segmentation of manipalUAVid.

A high F1 score is obtained for all four classes except construction class indicating a robust segmentation. In addition, a high mPA is observed as compared to TextonBoost, U-Net and DeepLabV3+ methods. The usage of deep CNN architecture VGG16 enabled the model to learn deep features which resulted in less false positives for all the four classes as compared to U-Net. However, mIoU of FCN-8 is similar to TextonBoost and U-Net.

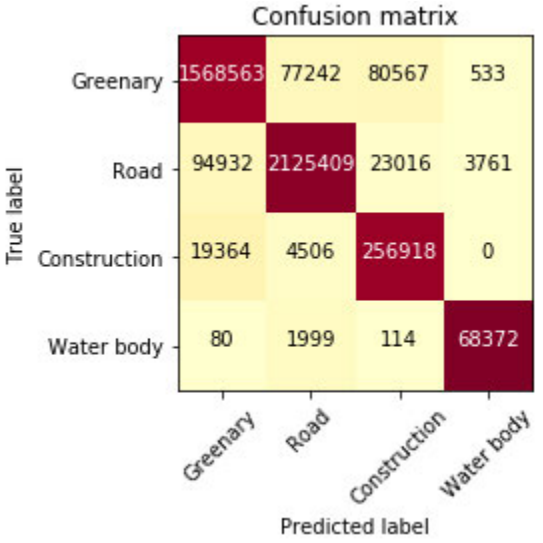


FIGURE 16. Confusion matrix for FCN-8 architecture for four class semantic segmentation of ManipalUAVid.

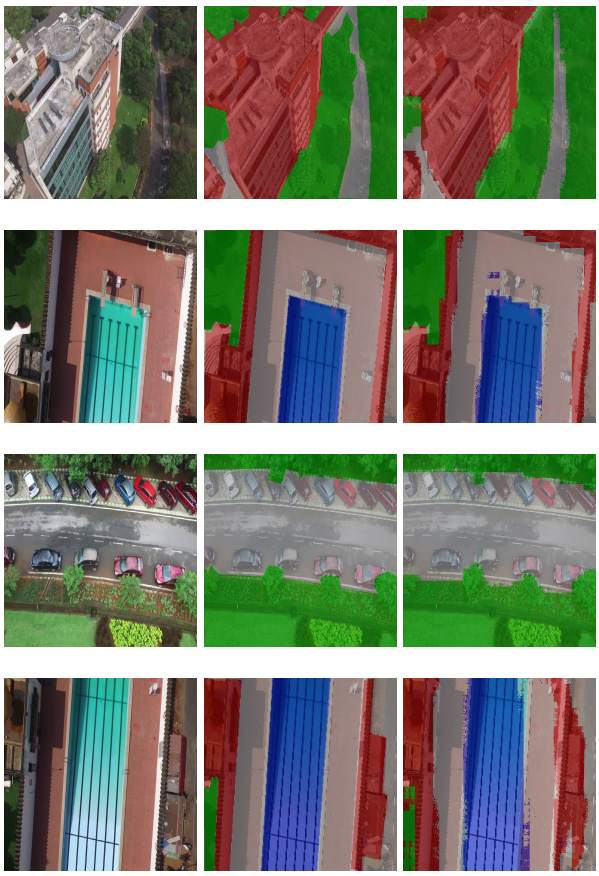


FIGURE 17. (a) Original image. (b) Ground truth image. (c) Semantic segmentation results by applying FCN.

Few sample results of FCN-8 semantic segmentation on UAV aerial videos is shown in Figure 17. The confusion matrix of FCN-8 model is shown in Figure 16.

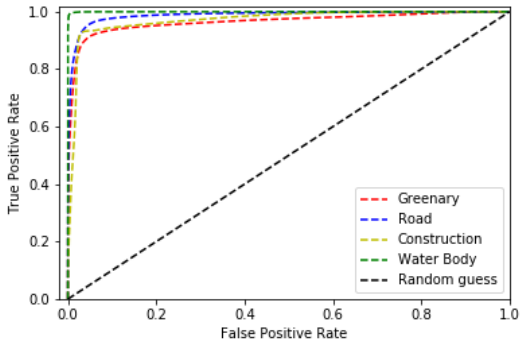
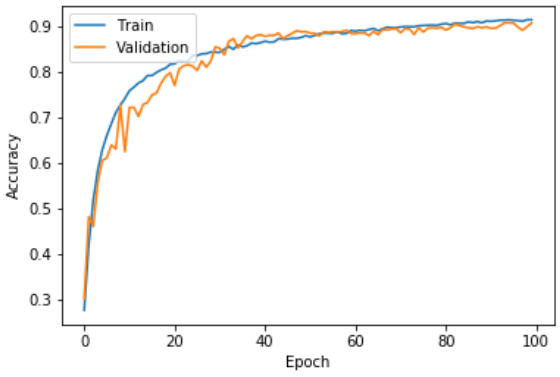
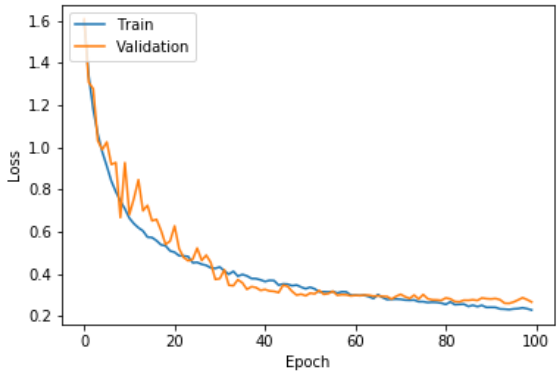


FIGURE 18. ROC curves obtained for FCN architecture for semantic segmentation of ManipalUAVid.



(a)



(b)

FIGURE 19. (a) Accuracy and (b) loss curve of DeepLabV3+ architecture for semantic segmentation of manipalUAVid.

E. CNN: DeepLabV3+

DeepLabV3+ is utilized to perform semantic segmentation of the keyframes identified in the ManipalUAVid aerial video dataset. The model is trained for 100 epochs on the dataset using categorical cross-entropy loss with Adam optimizer. Transfer learning is not utilized due to the availability of sufficient dataset. The model loss and accuracy curve for training and validation set is shown in Figure 19.

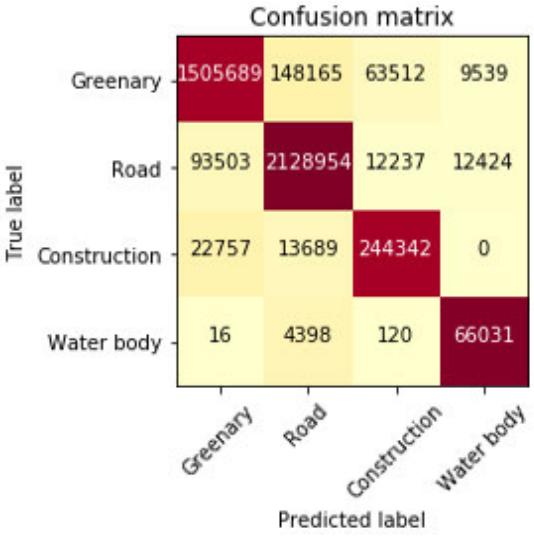


FIGURE 20. Confusion matrix for DeepLabV3+ architecture for four class semantic segmentation of ManipalUAVid.

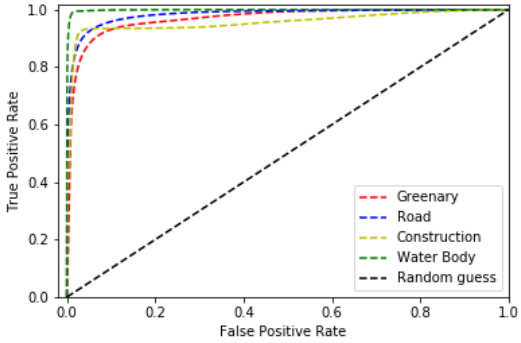


FIGURE 21. ROC curves obtained for DeepLabV3+ architecture for semantic segmentation of ManipalUAVid.

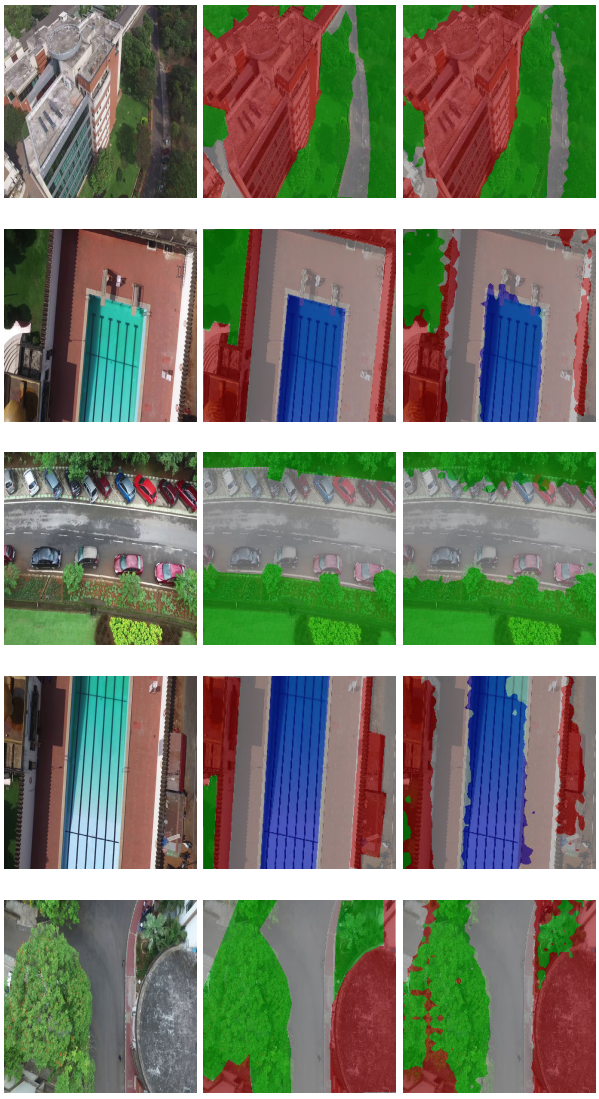


FIGURE 22. (a) Original image. (b) Ground truth image. (c) Semantic segmentation results by applying DeepLabV3+.

TABLE 8. Precision, Recall and F1-score of DeepLabV3+ on ManipalUAVid.

Algorithms	Precision	Recall	F1-Score
Greenery	0.93	0.87	0.90
Road	0.93	0.95	0.94
Construction	0.76	0.87	0.81
Water bodies	0.75	0.94	0.83

The performance metrics such as precision, recall and F1-score using DeepLabV3+ approach is shown in Table 8. A high F1 score (greater than 0.8) is observed for almost all class. PA, mPA and mIoU metrics utilizing DeepLabV3+ is shown in Table 5. A mIoU of 0.83 is obtained along with a mPA of 0.90 indicating a good segmentation. The ROC curve of DeepLabV3+ is shown in Figure 21. The area under the curve for all the four class is large indicating the high accuracy of the algorithm. The confusion matrix for DeepLabV3+ is shown in Figure 20. Few segmentation outputs obtained using DeepLabV3+ are shown in Figure 22. It can be seen that the four classes are segmented accurately with a sharp class boundary which is consistent with the previous findings. However, the occurrence of slightly higher false positives for construction and water body class has resulted in low precision of 0.76 and 0.75 respectively (bottom row in Figure 22).

As discussed earlier, semantic segmentation of aerial images has been studied in [17] using PolSAR images. In their study, the SAR images were semantically segmented into four classes viz water, vegetation, road, built up area and others using CNN based approaches. A pixel accuracy of 84% was obtained in their study. However, the presence of clouds limits the visibility of the Po1SAR images.

Instead, videos acquired using UAV can provide complementary information for semantic segmentation. Moreover, the semantic segmentation of UAV aerial videos using CRF and CNN (U-Net, FCN and DeepLabV3+) based

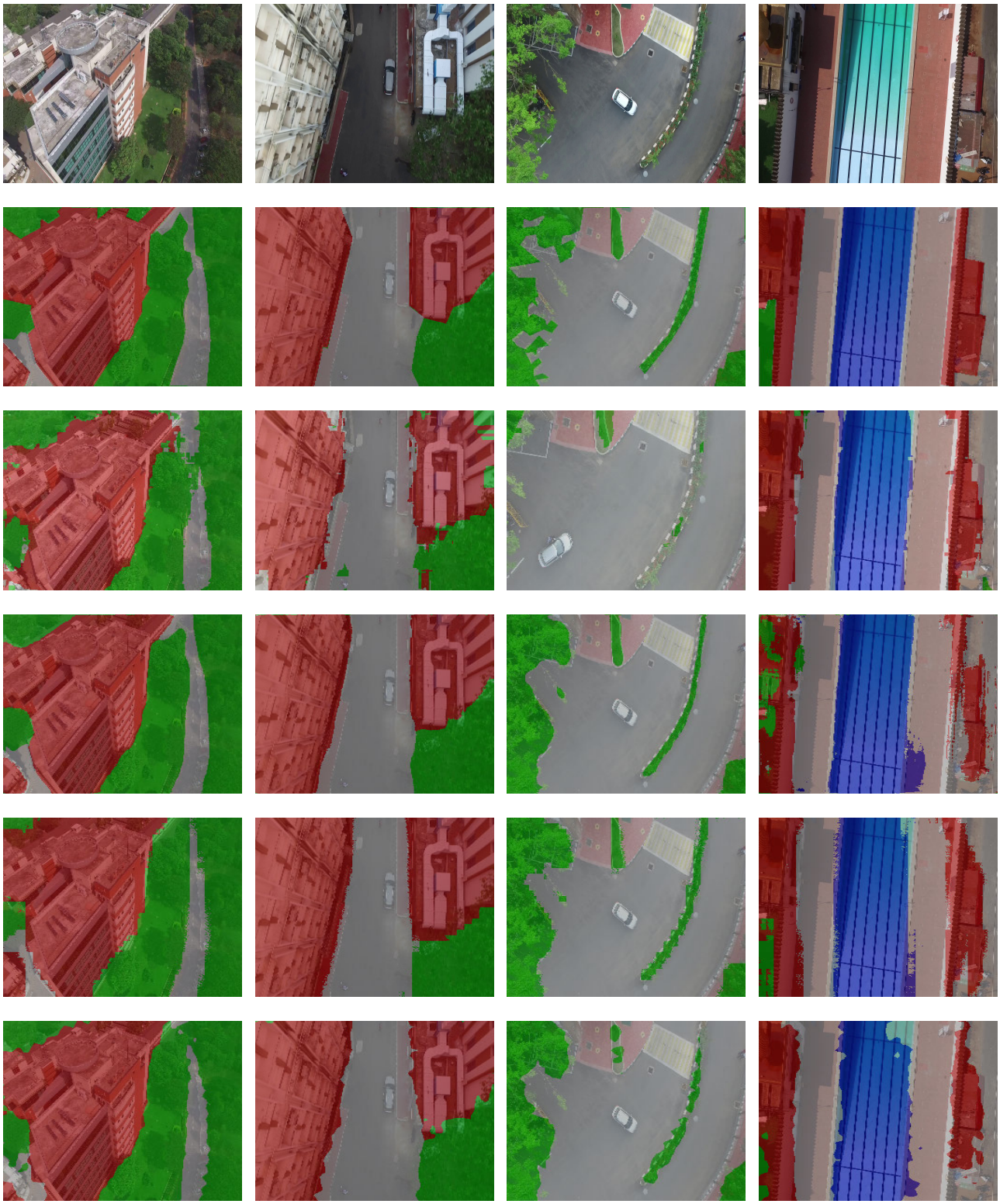


FIGURE 23. First row represents the original image. Second row shows the corresponding ground truth image. Third, fourth, fifth and sixth row shows the results obtained by applying Texton Boost, U-Net, FCN-8 and DeepLabV3+ algorithm.

approaches have achieved a higher pixel accuracy of 90% and 92% (average of U-Net, FCN and DeepLabV3+) respectively.

VI. CONCLUSION

Semantic segmentation is an important tool for scene understanding and plays a dominant role in various applica-

tions such as autonomous driving, object tracking, activity recognition, etc. However, a standard UAV aerial video dataset is essential to evaluate the performance of semantic segmentation algorithms. The new dataset consisting of aerial videos (more than 2 hours duration) acquired in a closed university campus is created as part of this work and is made available in public domain. The dataset contains fine annotations with four background classes (greenery, roads, construction and water bodies) describing the general outline of the scene for 650+ keyframes. The videos are captured with high-resolution cameras at different locations, during different weather conditions and different time of the day. The UAV aerial video dataset created through this research work has given a standard platform for analysing the performance of the developed semantic segmentation algorithms.

The aerial video dataset created as part of this work has been used to analyze the performance of the four standard approaches for semantic segmentation viz Conditional Random Field (CRF), U-Net architecture [2], FCN-8 [20] and DeepLabV3+ [35] architecture. For the CRF based approach, the texture layout feature along with colour, location and edge potential is incorporated [21]. It has been demonstrated that these approaches perform competitively for semantic segmentation with a pixel accuracy of 0.90 (CRF), 0.93 (U-Net), 0.92 (FCN) and 0.91 (DeepLabV3+). However, the CRF and DeepLabV3+ based method captures finer details as compared to CNN based approach, which is consistent with the previous findings. The comparative study of CRF, U-Net, FCN-8 and DeepLabV3+ has helped in gaining the insight of these algorithms and hence created an opportunity to develop state-of-the-art algorithm for multiclass semantic segmentation.

REFERENCES

[1] M. Teichmann, M. Weber, M. Zoellner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.

[2] 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.*, 2015, pp. 234–241.

[3] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 4, May 2015, pp. 1–6.

[4] W. G. Aguilar, M. A. Luna, J. F. Moya, V. Abad, H. Parra, and H. Ruiz, "Pedestrian detection for UAVs using cascade classifiers with meanshift," in *Proc. IEEE 11th Int. Conf. Semantic Comput. (ICSC)*, Jan. 2017, pp. 509–514.

[5] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 549–565.

[6] M. Everingham, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[7] 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. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3213–3223.

[10] M. Barekatain, M. Mart, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 28–35.

[11] 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. 2017.

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

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

[14] 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.

[15] S. Girisha, M. Pai, U. Verma, and R. Pai, "Semantic segmentation of UAV aerial videos using convolutional neural networks," in *Proc. IEEE Int. Conf. Artif. Intell. Knowl. Eng.*, Jun. 2019, pp. 21–27.

[16] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: https://arxiv.org/abs/1704.06857

[17] W. Wu, H. Li, X. Li, H. Guo, and L. Zhang, "PolSAR image semantic segmentation based on deep transfer learning—Realizing smooth classification with small training sets," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 977–981, Jun. 2019.

[18] B.-S. Kim, P. Kohli, and S. Savarese, "3D scene understanding by voxel-CRF," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1425–1432.

[19] X. He and S. Gould, "An exemplar-based CRF for multi-instance object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 296–303.

[20] N. D. Reddy, P. Singhal, and K. M. Krishna, "Semantic motion segmentation using dense CRF formulation," in *Proc. Indian Conf. Comput. Vis. Graph. Image Process.*, 2014, p. 56.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Dec. 2007.

[22] Y. Liu, S. Piramanayagam, S. T. Monteiro, and E. Saber, "Semantic segmentation of remote sensing data using Gaussian processes and higher-order CRFS," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5454–5457.

[23] M. T. Teichmann and R. Cipolla, "Convolutional CRFs for semantic segmentation," 2018, *arXiv:1805.04777*. [Online]. Available: https://arxiv.org/abs/1805.04777

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.

[25] 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.*, Jun. 2015, pp. 1–9.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[27] A. G. Schwing and R. Urtasun, "Fully connected deep structured networks," 2015, *arXiv:1503.02351*. [Online]. Available: https://arxiv.org/abs/1503.02351

[28] R. Vemulapalli, O. Tuzel, and M.-Y. Liu, "Deep Gaussian conditional random field network: A model-based deep network for discriminative denoising," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4801–4809.

[29] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*. [Online]. Available: https://arxiv.org/abs/1412.7062

[30] 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.

[31] X. Qi, J. Shi, S. Liu, R. Liao, and J. Jia, "Semantic segmentation with object clique potential," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2587–2595.
[32] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 44–57.
[33] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei, "Remote sensing image scene classification using bag of convolutional features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1735–1739, Oct. 2017.
[34] A. Zlateski, R. Jaroensri, P. Sharma, and F. Durand, "On the importance of label quality for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1479–1487.
[35] 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.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

S. GIRISHA received the B.E. degree from the Srinivas School of Engineering, VTU, Belgaum, and the master's degree in computer science and engineering from NMAMIT, Nitte, India. He is currently pursuing the Ph.D. degree in computer vision and machine learning with the Manipal Institute of Technology. His research interests include image segmentation, object detection, and deep learning for computer vision.

MANOHARA M. M. PAI received the Ph.D. degree in computer science and engineering. He has been a Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India, for the last 27 years. He holds six patents to his credit. He has published 80 articles in national and international journals/ conference proceedings. He has published two books, guided five Ph.D. students, and 65 master thesis. His research interests include data analytics, cloud computing, the IoT, computer networks, mobile computing, scalable video coding, and robot motion planning. He is a Life Member of ISTE and a Life Member of the Systems Society of India. He is also the Chair of the IEEE Mangalore Sub-Section, in 2019.

UJJWAL VERMA received the Ph.D. degree in image processing from Télécom ParisTech, University of Paris-Saclay, Paris, France, and the M.S. degree in signal and image processing from IMT Atlantique, France. He is currently an Assistant Professor with the Department of Electronics and Communication Engineering, Manipal Institute of Technology, India. His research interests include variational methods in image segmentation, action recognition, and deep learning methods for scene understanding. He is a recipient of the ISCA Young Scientist Award by the Indian Science Congress Association (ISCA), in 2018, a professional body with the Department of Science and Technology, Government of India. He is the Joint Secretary of the IEEE Mangalore Sub-Section, in 2019.

RADHIKA M. PAI received the Ph.D. degree from the National Institute of Technology Karnataka, Surathkal, India. She is currently a Professor with the Department of Information and Communication Technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. She has a teaching experience of above 27 years. She has published about 50 articles in national /international journals/conferences. She has guided three Ph.D. students and several master thesis. Her research interests include data mining, big data analytics, character recognition, sensor networks, and e-learning.