SEMANTIC SEGMENTATION OF UAV IMAGES BASED ON U-NET IN URBAN AREA

A. Majidizadeh [1, *] ,H. Hasani [1], M. Jafari [1]

[1] Department of Geodesy and Surveying Engineering, Tafresh University, 79611-39518, Tafresh, Iran – (h.hasani, Jafari, a.majidizadeh)@tafreshu.ac.ir

Commission IV, WG IV/3

KEY WORDS: Semantic Segmentation, UAV, Deep Learning, Convolutional Neural Network, Encoder-Decoder Architecture, U-Net.

ABSTRACT:

Semantic segmentation of aerial data has been one of the leading researches in the field of photogrammetry, remote sensing, and computer vision in recent years. Many applications, including airborne mapping of urban scenes, object positioning in aerial images, automatic extraction of buildings from remote sensing or high-resolution aerial images, *etc*., require accurate and efficient segmentation algorithms. According to the high potential of deep learning algorithms in the classification of complex scenes, this paper aims to train a deep learning model to evaluate the semantic segmentation accuracy of UAV-based images in urban areas. The proposed method implements a deep learning framework based on the U-Net encoder-decoder architecture, which extracts and classifies features through layers of convolution, max pooling, activation, and concatenation in an end-to-end process. The obtained results compare with two traditional machine learning models, Random Forest (RF) and Multi-Layer Perceptron (MLP). They rely on two steps that involve extracting features and classifying images. In this study, the experiments are performed on the UAVid2020 semantic segmentation dataset from the ISPRS database. Results show the effectiveness of the proposed deep learning framework, so that the U-Net architecture achieved the best results with 75.15% overall accuracy, compared to RF and MLP algorithms with 52.51% and 54.65% overall accuracy, respectively.

1. INTRODUCTION

In recent years, photogrammetry and remote sensing have achieved significant progress in the field of information extraction from databases (image, video, point cloud datasets, *etc.*). In the last three decades, data semantic segmentation has been one of the most challenging tasks in the photogrammetry and remote sensing community (Yuan et al. 2021). Semantic segmentation, as part of scene understanding, describes assigning each segment in an image to a class label (e.g., car, tree, vegetation, road, sky, person, *etc*.). In other words, semantic segmentation by dividing images into meaningful semantic objects has the task of labeling each image pixel into a predefined set of classes. Segmentation is used in a wide range of various urban applications, including; urban planning (Yao et al. 2019), land cover mapping (Matikainen and Karila, 2011), building and road extraction from high-resolution aerial images (Wei et al. 2017; Li et al. 2015), automated driving of autonomous vehicles (obstacle detection, pedestrian detection, traffic signs, etc.) (Wang et al. 2021). However, with the progress of deep learning networks and the significant improvement of their performance in recent years, Semantic segmentation in the understanding of urban aerial scenes has become an essential and, at the same time, challenging issue (Minaee et al. 2021).

Segmentation techniques can be classified into two categories: traditional approaches based on manual feature extraction and processes based on deep learning. Traditional methods are multi-step classifying images into several regions by considering the similarity between pixels. In these methods, the steps of extraction and classification of pixel-based features are done independently; firstly, the pixel-based features are extracted, and then the classification process is done by one of the typical classifiers. Although these methods can segment the image into separate parts, there is no semantic communication with the segmented areas in their results (Shi and Malik, 2000; Rother et al. 2004; Ullah and Alaya Cheikh, 2018). Consequently, semantic segmentation models that are a part of supervised methods, and are also carried out in an end-to-end framework, have gained the attention of researchers in the last decade (Garcia-Garcia et al. 2018; Neupane et al. 2021). In deep learning, due to the high capacity of feature learning, deep semantic segmentation methods, especially convolutional neural networks, have achieved high performance (Chen et al. 2014; He et al. 2016).

In this work, we used a symmetric encoder-decoder convolutional network (U-Net) (Ronneberger et al. 2015) for semantic segmentation of urban aerial scene images. U-Net is a well-known deep learning model which contains two parts: encoder and decoder. Its architecture is such that the encoder extracts the features through convolutional layers, max pooling, and activation. Next, the extracted feature vectors are decoded by the decoder module through convolution, transfer, concatenation, and activation layers. Each block of U-Net architecture consists of two convolutional layers with Batch Normalization and Rectified Linear Unit (ReLU) activation function. In addition, the features of each convolution block in the encoder are connected to the corresponding convolution block in the decoder by Concatenated layers to step-by-step obtain high-resolution features (Liu et al. 2019).

The structure of the paper is as follows: In Section 2, the related work is presented in the semantic segmentation of UAV images.

[*] Corresponding author
E-mail address: a.majidizadeh@tafreshu.ac.ir (A. Majidizadeh).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-W1-2022-451-2023 | © Author(s) 2023. CC BY 4.0 License.

Then, in Section 3, the proposed method will be presented in detail. In Section 4, we evaluate the performance of the implemented strategies. Finally, in Section 5, the conclusion is presented.

2. RELATED WORK

The development and appearance of various machine learning methods occurred in the 1990s. At that time, the support vector machine algorithm performed well in many fields. Soon after the 2000s, the era of big data began, which helped the development of various learning algorithms and made deep learning the focus of researchers' attention (Cox and Dean, 2014). Most early methods (Silberman et al. 2011; Khan et al. 2014) performed processing operations on the input data in several steps. At first, pre-processing is performed; then, features are extracted from the input data, and finally, the extracted features are used as input to the machine learning algorithms (such as Gray Level Co-occurrence Matrix (GLCM) and Histograms of Oriented Gradients (HOG)). In these approaches, the feature extraction step encounters challenges such as the generation of feature space with high dimensions and the existence of additional and dependent features, *etc*.

Recently, with the development of deep learning models, there has been an effort to solve the challenges in the field of machine learning in such a way that the extraction and classification of features automatically during the training operation by the learning algorithm itself, in an end-to-end process (Krizhevsky et al. 2012).

A wide range of studies have been published in semantic segmentation and classification of remotely sensed data. These approaches can be classified based on the type of data worked on, the method used to classify the segments, the segmentation type, and the object detection and tracking method. In this section, the related work is presented into two parts: semantic segmentation of UAV data based on traditional and deep learning methods.

2.1 Semantic segmentation based on traditional machine learning approaches

In these approaches, the segmentation process is performed in two separate steps feature extraction and classification. Morandozzo and Mellagni (2012) presented a combined approach based on corresponding feature extraction by scale-independent feature transform (SIFT) algorithm and feature classification with a support vector machine (SVM) classifier to identify vehicles in UAV images (Moranduzzo and Melgani, 2012). They (2014) proposed a method of extracting HOG gradients histogram features using horizontal and vertical filtering in the problem of vehicle detection based on UAV images (Moranduzzo and Melgani, 2014). Ammour et al. (2017) presented an automatic vehicle detection method in UAV images based on the extraction of candidate regions and the classification process. In this approach, a pre-trained convolutional neural network (VGG16) has been used to extract features. The classification operation of "car" and "non-car" has also been done by training the SVM support vector machine algorithm (Ammour et al. 2017). Bhatnagar et al. (2020) segmentation of UAV images using machine and deep learning approaches (combination of random forest algorithm and convolutional neural network methods) was presented for mapping marsh plant communities (Bhatnagar et al. 2020).

2.2 Semantic segmentation based on deep learning approaches

In these approaches, feature extraction and classification are carried out automatically in an end-to-end process. Micheal et al. (2021) presented a method based on deep learning to detect and track objects from UAV-based data. The proposed approach included deep supervised object detector training for object detection and long-short-term memory training for detecting detected objects (Micheal et al. 2021). Girsha et al. (2019) used the fully convolutional neural network (FCN) and U-Net convolutional encoder-decoder network for semantic segmentation of UAV video frames (Girsha et al. 2019). Wang et al. (2019) presented a deep learning method for semantic segmentation of drone video frames. The approach proposed is a combination of the fully convolutional network (FCN) and convolution long-short-term memory (Conv-LSTM) for semantic segmentation of video frames (Wang et al. 2019). Qiu et al. (2017) used a combination of three deep fully convolutional neural networks for semantic segmentation of video frames. The proposed framework combines 2D and 3D fully convolutional networks and convolutional long short-term memory (Qiu et al. 2017). Valipour et al. (2017) proposed and implemented an approach called fully convolutional recurrent networks for real-time segmentation of video sequences with the aim of video semantic segmentation. The architecture proposed in this paper consists of a fully convolutional network and a recurrent unit that operates on a sliding window of time data (Valipour et al. 2017). Kentsch et al. (2020) used a transfer learning technique and Multi-Layer Perceptron (MLP) classifier to analyze forest images obtained by drones (Kentsch et al. 2020).

In the recent two decades, due to the appearance of large-scale data, the traditional machine learning methods have been affected by the challenges of the feature extraction space. These approaches had low computational speed and accuracy in fitting training data. Due to the high dimensions of the training data, they required a system with increased memory, which was uneconomical. The development of deep learning approaches and their remarkable efficiency compared to traditional methods has been one of the research interests in various scientific fields in recent years.

In this paper, we propose a U-Net convolutional encoder-decoder neural network for semantic segmentation of UAV-based aerial images that can extract and classify features during the training operation in an end-to-end process. U-Net is a network with two encoder-decoder modules that, for image semantic segmentation, first, image features are extracted and encoded by the encoder module. Next, the extracted features are decoded using the decoder module. Finally, the prediction vectors of each class are presented through a classifier layer.

3. METHODOLOGY

The proposed method includes four key steps, whose flowchart is shown in Figure 1.

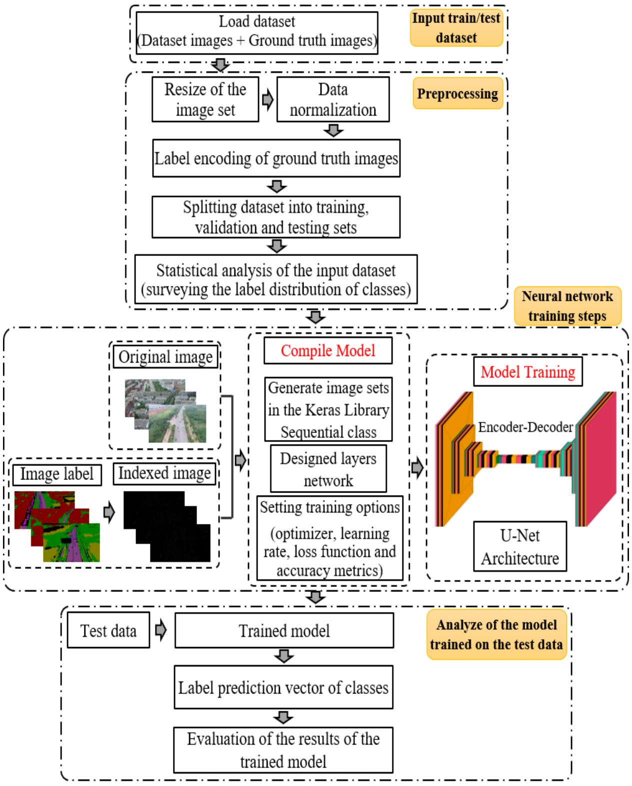

Load dataset
(Dataset images + Ground truth images)
Input train/test dataset
Resize of the image set
Data normalization
Preprocessing
Label encoding of ground truth images
Splitting dataset into training, validation and testing sets
Statistical analysis of the input dataset (surveying the label distribution of classes)
Neural network training steps
Original image
Image label
Indexed image
Compile Model
Generate image sets in the Keras Library Sequential class
Designed layers network
Setting training options
(optimizer, learning rate, loss function and accuracy metrics)
Model Training
Encoder-Decoder
U-Net Architecture
Test data
Trained model
Analyze of the model trained on the test data
Label prediction vector of classes
Evaluation of the results of the trained model

Figure 1. Flowchart of the proposed method.

The proposed model is a convolutional encoder-decoder deep learning structure. We have used the U-Net encoder-decoder neural network for semantic segmentation of high-resolution urban UAV-based images. This research aim is the semantic segmentation of urban UAV-based images using deep learning methods and comparison with traditional machine learning methods. For this purpose, after loading the original and the ground truth images through the pre-processing operation, we make the initial settings related to the neural network training process on the input dataset. The next step concerns tuning the parameters, defining the layers, and training the neural network. Then, we have the operation of training the neural network, during which the spatial features of the input image sets are extracted. Then they are given to a classifier. Finally, we provide the test image sets as input to the trained neural network and evaluate the performance of the proposed segmentation model in terms of segmentation of data classes and labeling by analyzing the prediction vector of the labels with the ground truth map. In this research, we used Random Forest (RF) and Multi-Layer Perceptron (MLP) to compare with the proposed approach.

Implementation of the proposed approach is based on four following steps:

1. Loading input data: the original and the corresponding ground truth images from the UAVid2020 dataset (Lyu et al. 2020).

2. Pre-processing: initial settings of the model training process on the input dataset. Settings include resizing input images, data normalization, label encoding (converting the labels into numerical form to present to the neural network), split data into train/test/validation, and statistical analysis of class label distribution.

3. Training of the neural network: after performing pre-processing on the input image sets and splitting them into training and test data, we generate the image sets by setting the dimensions, data class, and batch size in the Keras Library's Sequential class. The layers of the neural network are also defined. In addition, the optimizer, loss function, accuracy metric, and learning rate parameters on the neural network are set. Training is carried out, so feature extraction and classification are done end-to-end during the training operation.

4. Performance evaluation: test images are fed to the trained model and get the prediction vector of the classes in the output. The results of semantic segmentation are compared with ground truth maps. To analyze the performance of the proposed segmentation model with machine learning methods, we calculate the segmentation metrics (accuracy, IOU of each class, MeanIOU, and MeanBFScore).

3.1 U-Net convolutional encoder-decoder architecture

In this paper, U-Net neural network is used for semantic segmentation of urban aerial images. The U-Net architecture consists of two encoder-decoder modules. The encoder module extracts the spatial features from the training data, and the decoder module generates the prediction vector of data class labels from the encoded components. U-Net neural network module consists of four blocks. Each encoder block is separated by a Max Pooling layer, and each decoder block is separated by a Transposed convolutional layer. In the neural network, Max Pooling and Transposed convolutional layers with size (2,2), stride two are applied. Each block consists of two successive convolution layers (3,3), two batch normalization layers, and two activation functions of the Rectified Linear Unit (ReLU). Also, each decoded feature is connected to the corresponding feature map in the encoder through a Concatenate layer. Finally, using the softmax activation function, the vector of data class labels is predicted during the up-sampling process. The neural network architecture is shown in figure 2.

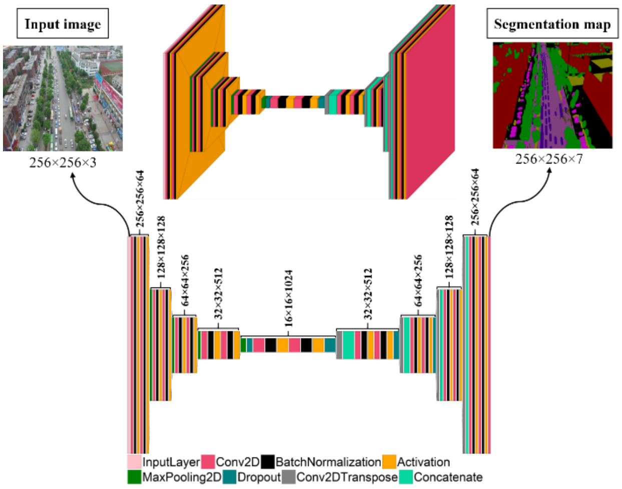


Figure 2. U-Net architecture (dimensions of input images 256×256×3 and dimensions of output segmentation map 256×256×7). We specified the size and number of channels of the feature maps above all layers of each network block.

3.2 Traditional machine learning

Random Forest (RF) and Multi-Layer Perceptron (MLP) machine learning algorithms are implemented to compare with the proposed approach. Figure 3 shows the flowchart of traditional machine learning methods. In these methods, feature extraction and classification are done independently.

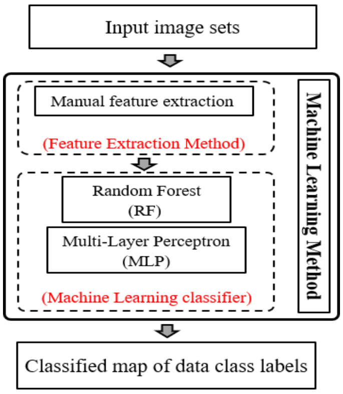

Input image sets
Manual feature extraction
(Feature Extraction Method)
Random Forest (RF)
Multi-Layer Perceptron (MLP)
(Machine Learning classifier)
Machine Learning Method
Classified map of data class labels

Figure 3. Flowchart of machine learning approaches.

Random Forest is trained in two steps; first, the data are sampled independently, and a decision tree is built for each sample. The decision trees are then combined for training, and each tree generates a classification prediction vector as output. Finally, one vote is taken for each predicted result, and the prediction with the most votes is selected as the final classification outcome. figure 4 shows the structure of the tenth decision tree from the trained Random Forest (RF).

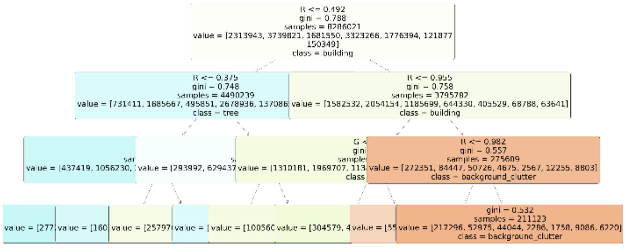

R <= 0.492
gini = 0.788
samples = 8286021
value = [2313943, 3739821, 1681550, 3323266, 1776394, 121877 150349]
class = building
R <= 0.375
gini = 0.748
samples = 4490239
value = [731411, 1685667, 495851, 2678936, 137086
class = tree
R <= 0.955
gini = 0.758
samples = 3795782
value = [1582532, 2054154, 1185699, 644330, 405529, 68788, 63641]
class = building
value = [437419, 1056230,
value = [293992, 629437
value = [1310181, 1969707, 113
class
R <= 0.982
gini = 0.557
samples = 775609
value = [272351, 84447, 50726, 4675, 2567, 12255, 8803]
class = background_clutter
value = [277
value = [160
value = [25797
value = [
value = [10056C
value = [304579, 4
value = [55
gini = 0.532
samples = 211123
value = [217296, 52975, 44044, 2286, 1758, 9086, 6220]
class = background_clutter

Figure 4. The structure of one of the decision trees from our random forest algorithm (tenth decision tree).

Multi-Layer Perceptron neural network is another classifier that is implemented for comparison. It consists of two Dense consecutive layers (fully connected layers). Next to each dense layer, a ReLU activation layer and a dropout layer are defined (figure 5). Since we have a multi-class segmentation, the softmax activation function is used as the last layer to discriminate the classes. We used a neural network including an input layer with three units (RGB values), six hidden layers with 512 units, and an output layer with seven units (number of class labels). Each input of the neural network is a 65536-dimensional array obtained from an image with dimensions of 256×256.

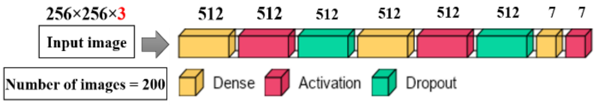
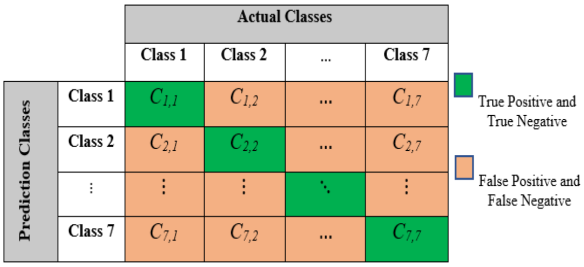


Figure 5. Multi-Layer perceptron architecture.

3.3 Evaluation metric

Performance evaluation is an essential issue in the classification process. The common way to analyze trained models is the evaluation of the diagonal and non-diagonal elements of the confusion matrix (figure 6 shows a confusion matrix for seven label classes.).



		Actual Classes			
		Class 1	Class 2	...	Class 7
Prediction Classes	Class 1	$C_{1,1}$	$C_{1,2}$	...	$C_{1,7}$
	Class 2	$C_{2,1}$	$C_{2,2}$	...	$C_{2,7}$
	⋮	⋮	⋮	⋱	⋮
	Class 7	$C_{7,1}$	$C_{7,2}$	...	$C_{7,7}$

🟩 True Positive and True Negative

🟧 False Positive and False Negative

Figure 6. Confusion matrix.

To evaluate the experimental results in this paper, we calculated overall accuracy, the intersection-over-union (IoU) score for each class, the mean IoU, and the MeanBFScore (Mean Boundary F1 Score) (equations 3, 5, 6, and 7). The Mean BFScore measures the matching of the predicted boundary of the objects with the ground truth boundary over the entire dataset (Csurka et al. 2013).

$$Precision_i = \frac{C_{ii}}{\sum_{j=1}^{N_{cls}} C_{ij}}, \tag{1}$$

$$Recall_i = \frac{C_{ii}}{\sum_{j=1}^{N_{cls}} C_{ji}}, \tag{2}$$

$$Overall\ Accuracy = \frac{\sum_{i=1}^{N_{cls}} C_{ii}}{\sum_{i=1}^{N_{cls}} \sum_{j=1}^{N_{cls}} C_{ij}}, \tag{3}$$

$$F_1Score = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{4}$$

$$IOU_i = \frac{target \cap prediction}{target \cup prediction} = \frac{C_{ii}}{\sum_{j=1}^{N_{cls}} C_{ij} + \sum_{\substack{j=1 \\ i \neq j}}^{N_{cls}} C_{ji}}, \tag{5}$$

$$Mean_{IOU} = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} IOU_i, \tag{6}$$

$$MeanBFScore = \frac{1}{N_{cls}} \sum_{i=1}^{N_{cls}} \mathrm{F_1Score}_i, \tag{7}$$

where N_{cls} = Number of classes
C_{ii} = The number of pixels belong to class i that are correctly labeled as class i
$C_{ij\ (ji)}$ = Number of pixels belong to class i (j) but classifier labels them as class j (i).

4. EXPERIMENTAL RESULTS

In this section, we analyze the details of training models, including the data set used, setting up deep and machine learning models, and evaluating the results of different methods.

4.1 Dataset

The dataset used in this work is the UAVid2020 dataset from the ISPRS database. UAVid2020 is a high-resolution UAV semantic segmentation dataset (high-resolution 4K images in oblique views) focusing on urban scenes. The images in this dataset are annotated into eight classes, Building, Road, Tree, Low Vegetation, Moving Car, Static Car, Background Clutter, and Human. The size of the original images is 3840×2160 pixels or 4096×2160 pixels (Lyu et al. 2020). Due to RAM limitations, we resized the training and test images to 256×256 pixels. Also, by reducing the dimensions of the images, the number of training pixels related to the human class in each image is significantly reduced, affecting the evaluation. Therefore, the human class is removed from evaluation, and we use seven classes to perform semantic segmentation of images. The total data used in this work includes 200 training images and 70 test images.

4.2 Results

In this work, the U-Net model is implemented in Python version 3.6 using the Sequential class of the Keras library and uses TensorFlow as the backbone. To train the neural network, we compiled categorical_crossentropy loss function, adam optimizer, and accuracy metric on the neural network. U-Net neural network training has been done in 30 epochs with a learning rate of 0.01. Moreover, to train the RF, the number of trees is considered to be 100, and used the Gini function to measure the quality. This algorithm is implemented using the Scikit-learn (Sklearn) machine learning library. The Sequential class of the Keras Library is also used to implement MLP. To model training, categorical_crossentropy loss function, adam optimizer, and accuracy metric are compiled on the neural network. Neural network training is done in 30 epochs.

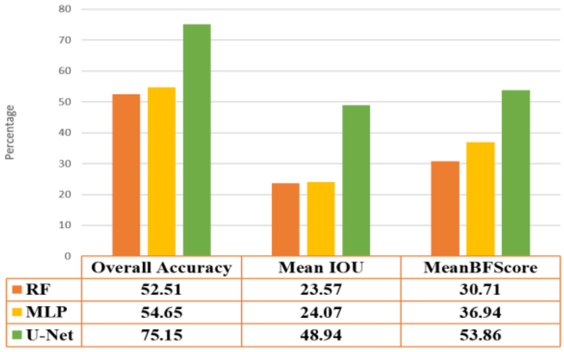
In this section, the potential of the U-Net neural network is evaluated in comparison with traditional machine learning algorithms in the semantic segmentation of UAVid2020 urban UAV-based image sets. For this purpose. Seventy (70) images are used as test data (these data are not contributing in the training process.). The experimental results of semantic segmentation obtained from the U-Net deep learning neural network and two other machine learning algorithms are presented in Table 1 and the statistical graph in Figure 7. Table 1 shows the IoU metric for each dataset class.

IOU Metric	Methods		
	MLP	RF	U-Net
Background clutter	0.2099	0.2026	0.4054
Building	0.4758	0.4530	0.7627
Road	0.2577	0.2578	0.6040
Tree	0.4584	0.4278	0.5035
Low vegetation	0.2738	0.2862	0.4934
Moving car	0.0095	0.0162	0.4443
Static car	0	0.0063	0.2123

Table 1. IOU metric results for each class.

Table 1 proves the high ability of U-Net in discriminate of all seven classes of complex urban areas. It depicts a low overlap percentage of the segmentation results of RF and MLP machine learning algorithms with the ground truth, especially in the classes with small samples (static and moving cars). At the same time, the U-Net deep learning network has a very high performance in overlapping with ground truth data, even the category of moving and static cars.

Figure 7 shows a comparison of three metrics of overall accuracy, mean IoU, and MeanBFScore (mean boundary F Score) for U-Net, MLP, and RF models.



	Overall Accuracy	Mean IOU	MeanBFScore
🟧 RF	52.51	23.57	30.71
🟨 MLP	54.65	24.07	36.94
🟩 U-Net	75.15	48.94	53.86

Figure 7. Comparison of OA, MIOU, and MeanBFScore metrics for three RF, MLP, and U-Net models.

Analyzing figure 7 shows the superiority of the U-Net deep learning network in all three measures of OA, MIOU and MeanBFScore compared to traditional machine learning algorithms (RF, MLP). The proposed U-Net neural network has performed more than 20% better than traditional methods in the semantic segmentation of urban aerial images in all three-evaluation metrics.

The visual results of semantic segmentation of the proposed model with machine learning models on three images from the test dataset are shown in figure 8.

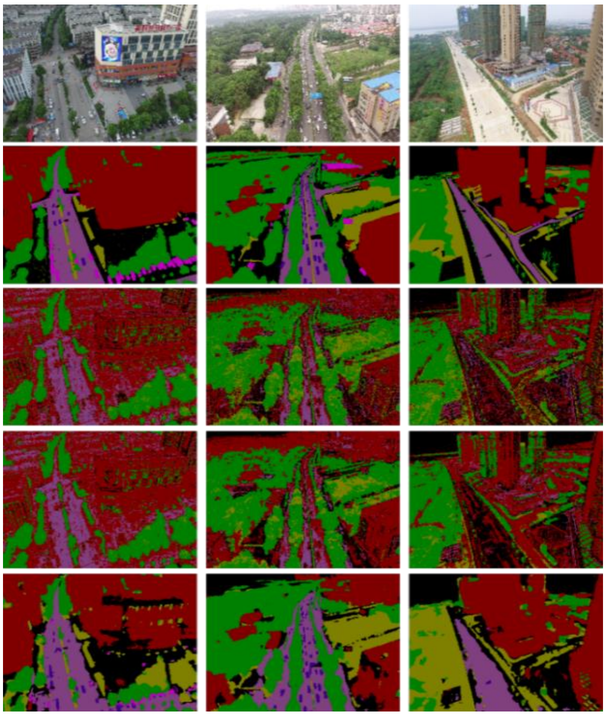


Figure 8. Segmentation results of U-Net, RF, and MLP models on the UAVid2020 test set. The first row shows the original image. The second row shows the ground truth images. The three last rows show the results of the RF, MLP, and U-Net models, respectively.

By investigating the visual results (the 3rd and 4th rows of figure 8), we can find that the two traditional machine learning approaches, RF and MLP, perform poorly in the semantic segmentation of UAV images and have problems matching the overlap boundaries with ground truth. Also, they performed the worst in classifying objects with small sample sizes, such as static and moving cars. On the other hand, the U-Net neural network has had an acceptable performance in the semantic segmentation of urban UAV-based images. According to the 5th row in figure 8, this model performs better than the other two approaches in the semantic segmentation of most classes, especially the small-scale classes of the moving car and, to some extent, the static car.

5. CONCLUSIONS

Our goal in this study was to perform semantic segmentation of UAV-based aerial imagery based on a deep learning approach. In this paper, we proposed a U-Net convolutional encoder-decoder architecture for semantic segmentation of urban aerial imagery. Experiments were performed on the UAV-based UAVid2020 dataset. The experiments performed on the test data showed the effectiveness of U-Net convolutional encoder-decoder architecture compared to Random Forest (RF) and Multi-Layer Perceptron (MLP) machine learning algorithms. The U-Net neural network provided acceptable performance in segmenting all seven data classes (especially those with smaller dimensions (static and moving car)). The results of the test dataset segmentation show that U-Net architecture with an overall accuracy of 75.15%, Mean IOU of 48.94%, and MeanBfScore score of 53.86% has a better balance in matching the predicted boundaries and ground reality in comparison with traditional machine learning classifiers.

For future works, it is suggested to use a transfer learning approach along with convolutional neural network training to achieve more accurate semantic segmentation. In this research, we resized the training dataset due to memory limitations. For more accurate semantic segmentation of urban scenes, it is suggested to use a set of higher-resolution images in a system with better hardware.

REFERENCES

Ammour, N., Alhichri, H., Bazi, Y., Benjdira, B., Alajlan, N. and Zuair, M., 2017. Deep learning approach for car detection in UAV imagery. *Remote Sensing*, *9*(4), p.312.

Bhatnagar, S., Gill, L. and Ghosh, B., 2020. Drone image segmentation using machine and deep learning for mapping raised bog vegetation communities. *Remote Sensing*, *12*(16), p.2602.

Chen, X., Xiang, S., Liu, C.L. and Pan, C.H., 2014. Vehicle detection in satellite images by hybrid deep convolutional neural networks. *IEEE Geoscience and remote sensing letters*, *11*(10), pp.1797-1801.

Cox, D.D. and Dean, T., 2014. Neural networks and neuroscience-inspired computer vision. *Current Biology*, *24*(18), pp.R921-R929.

Csurka, G., Larlus, D., Perronnin, F. and Meylan, F., 2004. What is a good evaluation measure for semantic segmentation?. *IEEE PAMI*, *26*(1). http://dx.doi.org/10.5244/C.27.32.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P. and Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing*, *70*, pp.41-65.

Girisha, S., MM, M.P., Verma, U. and Pai, R.M., 2019, June. Semantic segmentation of uav aerial videos using convolutional neural networks. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)* (pp. 21-27). IEEE.

Yuan, X., Shi, J. and Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, *169*, p.114417.

He, T., Huang, W., Qiao, Y. and Yao, J., 2016. Text-attentional convolutional neural network for scene text detection. *IEEE transactions on image processing*, *25*(6), pp.2529-2541.

Kentsch, S., Lopez Caceres, M.L., Serrano, D., Roure, F. and Diez, Y., 2020. Computer vision and deep learning techniques for the analysis of drone-acquired forest images, a transfer learning study. *Remote Sensing*, *12*(8), p.1287.

Khan, S.H., Bennamoun, M., Sohel, F. and Togneri, R., 2014, September. Geometry driven semantic labeling of indoor scenes. In *European Conference on Computer Vision* (pp. 679-694). Springer, Cham.

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), pp.84-90.

Li, E., Femiani, J., Xu, S., Zhang, X. and Wonka, P., 2015. Robust rooftop extraction from visible band images using higher order CRF. *IEEE Transactions on Geoscience and Remote Sensing*, *53*(8), pp.4483-4495.

Liu, Y., Gross, L., Li, Z., Li, X., Fan, X. and Qi, W., 2019. Automatic building extraction on high-resolution remote sensing imagery using deep convolutional encoder-decoder with spatial pyramid pooling. *IEEE Access*, *7*, pp.128774-128786.

Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A. and Yang, M.Y., 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, *165*, pp.108-119. doi.org/10.1016/j.isprsjprs.2020.05.009.

Matikainen, L. and Karila, K., 2011. Segment-based land cover mapping of a suburban area—Comparison of high-resolution remotely sensed datasets using classification trees and test field points. *Remote Sensing*, *3*(8), pp.1777-1804.

Micheal, A.A., Vani, K., Sanjeevi, S. and Lin, C.H., 2021. Object detection and tracking with UAV data using deep learning. *Journal of the Indian Society of Remote Sensing*, *49*(3), pp.463-469.

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N. and Terzopoulos, D., 2021. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

Moranduzzo, T. and Melgani, F., 2012, July. A SIFT-SVM method for detecting cars in UAV images. In *2012 IEEE*

International Geoscience and Remote Sensing Symposium (pp. 6868-6871). IEEE.

Moranduzzo, T. and Melgani, F., 2014. Detecting cars in UAV images with a catalog-based approach. *IEEE Transactions on Geoscience and remote sensing*, *52*(10), pp.6356-6367.

Neupane, B., Horanont, T. and Aryal, J., 2021. Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. *Remote Sensing*, *13*(4), p.808.

Qiu, Z., Yao, T. and Mei, T., 2017. Learning deep spatio-temporal dependence for semantic video segmentation. *IEEE Transactions on Multimedia*, *20*(4), pp.939-949.

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

Rother, C., Kolmogorov, V. and Blake, A., 2004. " GrabCut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, *23*(3), pp.309-314.

Shi, J. and Malik, J., 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), pp.888-905.

Silberman, N. and Fergus, R., 2011, November. Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 601-608). IEEE.

Ullah, M., Mohammed, A. and Alaya Cheikh, F., 2018. PedNet: A spatio-temporal deep convolutional neural network for pedestrian segmentation. *Journal of Imaging*, *4*(9), p.107.

Valipour, S., Siam, M., Jagersand, M. and Ray, N., 2017, March. Recurrent fully convolutional networks for video segmentation. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 29-36). IEEE.

Wang, L., Li, R., Wang, D., Duan, C., Wang, T. and Meng, X., 2021. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sensing*, *13*(16), p.3065.

Wang, Y., Lyu, Y., Cao, Y. and Yang, M.Y., 2019, July. Deep learning for semantic segmentation of UAV videos. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium* (pp. 2459-2462). IEEE.

Wei, Y., Wang, Z. and Xu, M., 2017. Road structure refined CNN for road extraction in aerial image. *IEEE Geoscience and Remote Sensing Letters*, *14*(5), pp.709-713.

Yao, H., Qin, R. and Chen, X., 2019. Unmanned aerial vehicle for remote sensing applications—A review. *Remote Sensing*, *11*(12), p.1443.