# COMPARATIVE ANALYSIS OF DEEP LEARNING-BASED TECHNIQUES FOR UAV IMAGE SEMANTIC SEGMENTATION

**Muhammed Enes Atik**

Assistant Professor (PhD) Muhammed Enes Atik;
Department of Geomatics Engineering
Department of Geomatics Engineering, Faculty of Civil Engineering, Istanbul Technical University, 34469 İstanbul, Türkiye
telephone: +902122853414, e-mail: atikm@itu.edu.tr

***Abstract***
*Unmanned aerial vehicle (UAV) images have become essential data for rapid national natural resource surveys, emergency mapping, and disaster monitoring. With the development of sensor technology in UAVs, aerial images containing significant differences in the appearance of objects and complex backgrounds have opened up a new field of research, such as semantic segmentation. Deep learning-based semantic segmentation techniques have begun to be used successfully for UAV images. In this study, the performance of existing deep learning networks in the semantic segmentation of UAV images will be comparatively analyzed. Existing algorithms such as DeepLabv3 with different backbones were examined in the study. All experiments will be carried out using the UAVid dataset containing aerial images. Thus, it aims to present comprehensive research on UAV image semantic segmentation and analyze the performances of existing algorithms in detail. ResNet-50 achieved an overall accuracy and F1 score of 80.91 and 65.66, respectively.*

***Keywords:*** *UAV, Aerial Image, Semantic Segmentation, Deep Learning.*

## INTRODUCTION

Applications related to the United Nations (UN) Sustainable Development Goals (SDGs) for environmental monitoring purposes are becoming widespread. For these applications, real-world data with high spatial resolution are very useful data sources (Isiler et al., 2023). Unmanned aerial vehicle (UAV) images have become essential data for many applications, such as rapid national natural resource surveys, emergency mapping, and disaster monitoring. In recent years, visual scene understanding using UAV imagery has become an important research area (Lyu et al., 2020). Semantic segmentation is one of the main tasks of scene understanding. Semantic segmentation of multispectral satellite imagery and high-resolution UAV imagery has many applications. Artificial intelligence techniques developed recently have provided significant advantages in processing large images (Atik and Ipbuker, 2021). Learning-based approaches have substantial advantages over traditional rule-based methods (Atik and Duran, 2022; Majidizadeh et al., 2023). Thanks to the high feature learning capacity of deep learning, successful semantic segmentation results can be produced (Majidizadeh et al., 2023). For this purpose, many methods and data sets (Lyu et al., 2020; Nigam et al., 2018) have been presented in the literature. However, there is still a large research need due to advances in sensor technology and hardware specifications.

This study comparatively analyzes the performance of five different backbones integrated with the DeepLabv3+ architecture. These backbones are ResNet-18, ResNet-50, Xception, MobileNetv2 and Inception-ResNet-v2. The recently published UAVid dataset is used as the dataset. The dataset was obtained in the urban area and contains eight classes.

## MATERIAL AND METHODS

### UAVid Dataset

The UAVid dataset is published by Lyu et al. (2020) for semantic segmentation tasks. DJI phantom 3 pro and DJI phantom 4 are used for data collection. The image size is 3850 × 2160 pixels and the flight height is 50 meters. The dataset is splitted into 400 images for training, 140 for validation, and 300 for test. The dataset includes eight classes in urban area: building, road, static car, tree, low vegetation, human, moving car and background clutter. The sample images are presented in Figure 1.


*Figure 1. The samples from the UAVid dataset (Lyu et al., 2020).*

## DeepLabv3+

Because DCNNs use consecutive pool and convolution layers, the input feature maps in images for segmentation tasks become smaller. Images lose their contextual information as a result of this. It produces outputs with blurry object boundaries and limited spatial resolution of the estimates. The DeepLab model addresses this issue using Atrous convolutions and Atrous Spatial Pyramid Pooling (ASPP) modules (Chen et al., 2017). DeepLabv3 does not need to adjust the feature view resolution. It controls how densely features are computed in convolutional networks. Consequently, multi-scale object features in fully convolutional networks can be calculated without reducing feature map sizes. For multi-scale context information extraction in the ASPP network, four parallel atrous convolutions with different atrous ratios are applied to the feature map (Atik et al., 2022). Several architectures can be utilized with the DeepLab model to extract features. In this study, ResNet18, ResNet50, Xception, Mobilenetv2, and Inception-ResNetv2 are used for feature extraction.

ResNet (He et al., 2016) enhanced the CNN architecture with multiple stacked residual units to enable network optimization and enhance accuracy with considerably more depth. Such neural networks are based on the idea of skipping connections, which is central to residual blocks. Creating an alternative path for the gradient to follow skipping connections reduces the issue of the gradient disappearing. ResNet-34 was the first ResNet architecture, and it added shortcut links to change a flat network into its network equivalent. The VGG neural networks using convolutional networks with 3 × 3 filters served as the model for the ResNet-34 architecture. ResNets are more straightforward and contain fewer filters than VGG networks. ResNet versions are named according to the number of weight layers they contain (He et al., 2016).

The Xception architecture (Chollet, 2017) is an “extreme” version of the Inception architecture. Inception proposes a more efficient process by separating the process of looking for cross-channel and spatial correlations into sub-processes. The Inception module maps the input data into three or four small fields by computing cross-correlations through 1 × 1 convolutions. Instead of using numerous 1 × 1 convolutions, Xception can be redesigned as a large 1 × 1 convolution. The whole CNN architecture of Xception is built on highly separable convolution layers. The spatial correlations of each output channel are mapped independently in the Xception architecture after the correlations between the channels are matched using 1 × 1 convolution (Atik et al., 2022).

MobileNet (Howard et al., 2017) is designed to be used as a fast and efficient method for real-time applications. Since it does not require any particular operator, it is a successful model in multiple image classification and detection tasks and has low data processing capacity and low latency. When extracting features from images using convolutional filters, the MobileNet architecture employs a technique known as depthwise separable convolutions rather than the conventional convolutional operation. Deeply separable convolutions divide it into two filtering layers and a combining layer using a different layer. With just a slight loss in accuracy, MobileNetv2 (Sandler et al., 2018), which was created based on the MobileNetv1 network, uses three deep separable convolutions at a computational cost that is eight to nine times less

costly than that of regular convolutions. To reduce information loss in layers that apply the linear transformation, MobileNetv2 also includes linear bottleneck layers.

A deep convolutional neural network architecture called Inception-ResNet-v2 (Szegedy et al., 2017) combines the ideas of two strong models, ResNet and Inception. Inception-ResNet-v2 employs a variation of the Incepiton modules to extract features from the input images. Also, residual connections in ResNet are incorporated into the Inception modules in Inception-ResNet-v2, which improves the modules' functionality. It can capture both high-level features and fine-grained information with the combination of ResNet and Inception components, producing predictions that are highly accurate.

## RESULTS AND DISCUSSION

This study compares the performance of DeepLabv3+ algorithms integrated with different backbones for UAV image segmentation. Input images for DL consist of patches of 500 × 500 pixels obtained from images in the data set. As training parameters, 20 epochs, 8 batch sizes, 0.001 learning rate, and stochastic gradient descent with momentum (SGDM) optimizer were selected. The learning rate is reduced by a factor of 0.1 every 6 epochs. "CrossEntropyLoss" was applied as loss function by assigning weight to each of the classes. All experiments were carried out in MATLAB environment. All experiments are performed by using a computer with GPU-supported RTX 4080 graphics card and 64 GB RAM. Algorithm performances were evaluated with overall accuracy, F1 score, and IoU metrics.

According to the overall dataset metrics (Table 1), the highest overall accuracy was achieved with DeepLabv3+ ResNet50 with 80.91%. DeepLabv3+ Inception-ResNet-v2 achieved the highest F1-score (65.85%). Xception has the lowest results in both metrics. However, it has very similar values to MobileNetv2. According to the results, ResNet architectures improve accuracy for semantic segmentation of UAV images. While the Xception method has low metrics, the Inception-ResNet-v2 architecture has one of the most successful results due to ResNet combined with the Inception module. As expected, MobileNetv2 is one of the algorithms with the lowest accuracy metrics because it proposes a simpler and less accurate architecture for fast evaluation.

*Table 1. The general metrics for UAVid dataset.*

| Method | Overall Accuracy (%) | F1 Score (%) |
|---|---|---|
| DeepLabv3+ ResNet18 | 79.81 | 64.17 |
| DeepLabv3+ ResNet50 | **80.91** | 65.66 |
| DeepLabv3+ Xception | 77.03 | 60.21 |
| DeepLabv3+ MobileNetv2 | 77.26 | 60.30 |
| DeepLabv3+ Inception-ResNet-v2 | 80.32 | **65.85** |

Class-based results are presented according to the IoU metric (Table 2). The algorithms achieve the highest accuracy as expected for classes with many instances such as building, road, and tree. Because in deep learning, the number of instances is an important factor that improves accurate prediction. Human class has the fewest instances, so the algorithms did not learn enough. ResNet-50 achieved the highest IoU in the Human class with 13.82% IoU. The lowest IoU values were obtained in the human class. ResNet-50 and Inception-ResNet-v2 algorithms generally produce the highest IoU values. ResNet-18 has the highest IoU in the road and tree classes. Moving car class is better predicted than static car. The classification results are illustrated in Figure 2.

*Table 2. The class-based IoU values for UAVid dataset. The values are in %.*

| Class | ResNet18 | ResNet50 | Xception | MobileNetv2 | Inception-ResNet-v2 |
|---|---|---|---|---|---|
| Background clutter | 53.44 | **57.33** | 50.79 | 51.81 | 55.35 |
| Building | 79.22 | **81.40** | 78.61 | 76.11 | 80.71 |
| Road | **71.94** | 71.85 | 68.73 | 70.33 | 71.81 |
| Tree | **72.94** | 72.81 | 67.82 | 67.52 | **72.94** |
| Low vegetation | 54.43 | **56.45** | 51.62 | 50.92 | 53.89 |
| Moving car | 57.67 | 56.96 | 41.41 | 55.08 | **60.37** |
| Static car | 46.28 | 42.74 | 42.65 | 40.24 | **48.68** |
| Human | 9.57 | **13.82** | 6.53 | 8.94 | 11.93 |
| **mIoU** | 55.67 | 56.67 | 51.02 | 52.62 | **56.96** |

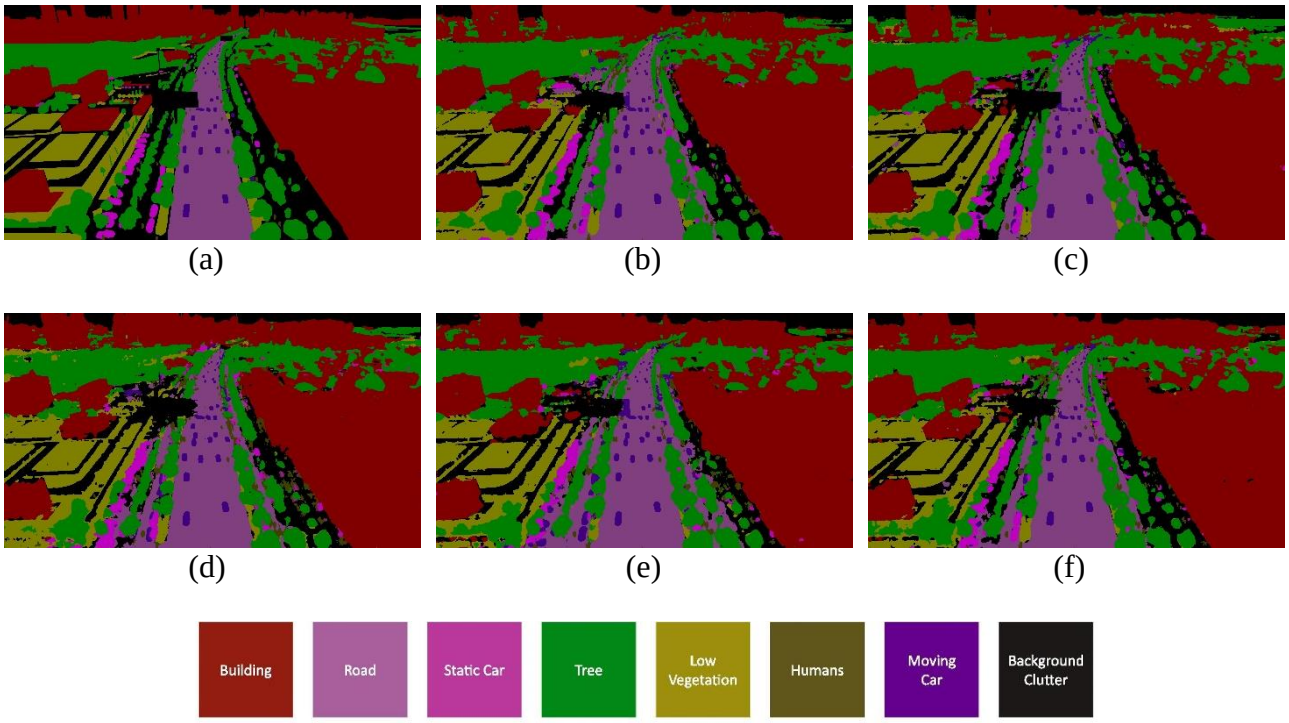


*Figure 2. The classified UAV image for each algorithm. (a) Ground truth; (b) ResNet-18; (c) ResNet-50; (d) Xception; (e) MobileNetv2; (f) Inception-ResNet-v2.*

## CONCLUSIONS

This study presents comparative research on the semantic segmentation of UAV images with DeepLabv3+ architecture using different backbones. Deep learning techniques have been shown to offer promising results for multi-class semantic segmentation problems. In future studies, data sets can be developed with varying angles of shooting such as oblique images. Additionally, point clouds obtained with LiDAR sensors can be fused with UAV imagery for more successful semantic segmentation. Deep learning techniques have significant potential for feature extraction, especially from large UAV images.

## REFERENCES

Atik, M. E., & Duran, Z. (2022). Selection of relevant geometric features using filter-based algorithms for point cloud semantic segmentation. Electronics, 11(20), 3310.

Atik, S. O., & Ipbuker, C. (2021). Integrating convolutional neural network and multiresolution segmentation for land cover and land use mapping using satellite imagery. Applied Sciences, 11(12), 5551.

Atik, S. O., Atik, M. E., & Ipbuker, C. (2022). Comparative research on different backbone architectures of DeepLabV3+ for building segmentation. Journal of Applied Remote Sensing, 16(2), 024510-024510.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Isiler, M., Yanalak, M., Atik, M. E., Atik, S. O., & Duran, Z. (2023). A Semi-Automated Two-Step Building Stock Monitoring Methodology for Supporting Immediate Solutions in Urban Issues. Sustainability, 15(11), 8979.

Lyu, Y., Vosselman, G., Xia, G. S., Yilmaz, A., & Yang, M. Y. (2020). UAVid: A semantic segmentation dataset for UAV imagery. ISPRS journal of photogrammetry and remote sensing, 165, 108-119.

Majidizadeh, A., Hasani, H., & Jafari, M. (2023). Semantic segmentation of UAV images based on U-NET in urban area. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 10, 451-457.

Nigam, I., Huang, C., & Ramanan, D. (2018, March). Ensemble knowledge transfer for semantic segmentation. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1499-1508). IEEE.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4510-4520).

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

## BIOGRAPHY

**Muhammed Enes Atik** received a Ph.D. degree in geomatics engineering from Istanbul Technical University, Istanbul, Turkey, in 2022. He is currently working as an Assistant Professor with the Department of Geomatics Engineering, at Istanbul Technical University. His research interests include photogrammetry, deep learning, machine learning, point cloud segmentation, LiDAR, UAVs, and remote sensing.