iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82236

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Binary Vegetation Mapping from High-Resolution Satellite Imagery Using Deep Learning: A Comparative Study

Harish Kundar[1], Meera Jagadeesh H[2], Meghana D G[3], Vaishnavi S Kedhlaya[4]
Department of Artificial Intelligence and Machine Learning, Alva's Institute of Engineering and Technology, Mangalore, India

Abstract: ***Semantically segmented images can be processed pixel-wise and have many applications in satellite image interpretation, environmental monitoring, and city planning. However, performing segmentation on high-resolution images is difficult because there are differences in scene complexities and scales. This paper discusses a comparison between two popular deep learning models: U-Net and DeepLabV3+ when working with high- resolution images and binary semantic segmentation tasks based on high-resolution satellite images. While U-Net incorporates the symmetrical encoder-decoder structure with skip connections, DeepLabV3+ incorporates atrous convolutional networks and Atrous Spatial Pyramid Pooling (ASPP) modules to include multiscale contextual features. Both models are benchmarked according to various metrics, including accuracy and Intersection over Union (IoU), for the binary class (vegetation/non-vegetation) of the DeepGlobe dataset. According to the findings, DeepLabV3+ outperforms U-Net through higher IoUs and consistency in segmentation tasks.***
Index Terms: Semantic Segmentation, U-Net, DeepLabV3+, Deep Learning, DeepGlobe Dataset

I. INTRODUCTION

Semantic Segmentation is one of the basic steps in computer vision, wherein each pixel is assigned a class label in an image. Unlike other approaches, such as image classification and object detection, semantic segmentation offers finer granularity in understanding scene structures. Semantic Segmentation is important for applications like remote sensing, self-driving cars, urban development, environmental protection, and med- ical imaging, among others.

Due to fast technological advances, satellite images can be captured with high resolution, thereby offering an excellent source for visual information. Nevertheless, obtaining useful information from this kind of data is quite a complicated task, due to different factors like illumination changes, differences in object scale, presence of occlusion, as well as complicated backgrounds. One of the examples when the problem arises is the process of vegetation detection, which needs proper distinction between vegetation and other regions in the image. That is why deep learning approaches have been developed.

FCNs marked a breakthrough in computer vision tech- nology regarding prediction on the pixel level without any incorporation of fully connected layers [3]. Various deep learning models have been designed based on the idea of FCNs to improve the accuracy of the results achieved through segmentation tasks. Among those neural network models, the most influential one may be the U-Net model [1] that gained popularity due to its structure made up of encoder-decoder blocks with skip connections.

On the other hand, DeepLabV3+ [2] employs atrous/dilated convolution, allowing for enlargement of the receptive field size while keeping the spatial resolution intact. Additionally, the use of ASPP module means that the DeepLabV3+ model is capable of extracting multiscale feature maps from the image, making it very effective for analyzing images with differently sized objects.

However, despite all of the aforementioned facts, picking the most appropriate way of implementing either one of these methodologies to implement high-resolution semantic segmentation is no easy feat. To illustrate the problem at hand better, upon further consideration of the structure of the U- Net methodology, one finds that it is extremely suitable for detecting fine-grained information; however, it is unsuitable for global information context detection.

In terms of the study at hand, our primary area of interest will center around the process of extracting vegetation from satellite imagery through binary classification. In this case, the image will essentially be divided into two categories, which are vegetation and non-vegetation. Although the problem might appear simple, it does hold some considerable importance in the context of the environment. In this regard, a thorough comparative study between U- Net and DeepLabV3+ models on high-resolution images will be conducted in this paper. The comparative study between both models will be done based on the application of both qualitative and quantitative approaches including accuracy and IoU score.

©IJRASET: All Rights are Reserved

The main contributions of this paper are summarized as follows:

1) Detailed comparison of U-Net and DeepLabV3+ for binary segmentation in high-resolution remote sensing imagery.
2) Performance assessment based on accuracy and IoU measurements.
3) Examination of the merits and demerits of both models in tackling complex scenes.

II. LITERATURE REVIEW

There have been several milestones in semantic segmentation due to the rise of deep learning models. Initially, classical computer vision techniques and manually engineered features were used in semantic segmentation, but they had their limitations with dealing with complicated visual phenomena. With Fully Convolutional Networks [3], there was a leap of progress that enabled pixel-level predictions through the use of convolutional layers. Fully Connected layers were replaced by convolutional layers, thereby making the network capable of processing images of any resolution.

The difference between the architecture of U-Net and FCN is that the presence of skip connections is a feature present in the U-Net architecture. This leads to the symmetry of the U-Net architecture. In addition, the skip connections enable capturing of spatial information that could not be captured owing to downsampling. As a result of the ability to capture spatial information, the U-Net architecture can be applied in a number of fields, including image segmentation in medicine and remote sensing.

In order to overcome this problem, DeepLabV3+ [2] came into existence, utilizing the atrous (or dilated) convolution approach, which helps to increase the receptive field while keeping the spatial dimension intact. Further, the Atrous Spatial Pyramid Pooling (ASPP) architecture is designed to improve the model's performance by extracting multi-scale context information through parallel atrous convolutions with varying dilation factors.

Algorithm PSPNet [4] adopted the strategy of pyramid pooling for gathering global information from images having objects at various sizes by the help of images at various scales. Likewise, the algorithm SegNet [8] adopted the strategy of an encoder-decoder architecture using max-pooling indices.

The deep learning paradigm was further advanced by the introduction of the residual network (ResNet), allowing for the training of deep neural networks using skip connections. It has become very popular for use as a backbone network in several DeepLab-based segmentations. Several successes of the recent surveys [6], [7] demonstrate the role of the use of multi-scale information and context information in the field of research of semantic segmentation. It is known that the use of deep learning approaches in remote sensing research provides better results than conventional methods, which is related to the high efficiency of deep learn- ing [10]. There are several studies dedicated to the problem of generating vegetation maps and land cover classification in the case of binary segmentation (vegetation/non-vegetation).

From the above literature review, it is clear that although U- net is quite efficient in extracting finer details, it is effective when dealing with smaller data sets, whereas DeepLabV3+ is more effective in dealing with complex situations due to its capability to perform local and global analysis. There is an urgent need to conduct a comparison between these two models in terms of vegetation detection in satellite images through binary classification.

III. METHODOLOGY

A. Dataset and Preprocessing

The experiments will be carried out using the DeepGlobe data set [5]. The DeepGlobe data set [5] can be seen as one of the standard large-scale data sets for studying satellite images. In the mentioned data set, there are satellite images of high resolution annotated by pixel-level labeling with a purpose of different purposes like road extraction, land cover classification, and building detection. Each picture in the dataset has an associated ground-truth mask that encodes pixel-level class labels. The high-resolution nature of the dataset presents some difficulties, including class imbalance, detecting fine boundaries, and noisy annotations.

For the purpose of this analysis, the multi-class labels in DeepGlobe data have been transformed to a binary classifica- tion problem. Classes have been categorized into two groups, one being vegetation and the other being non-vegetation. Vegetation consists of regions that are covered by trees, grass, or other plant-related entities, whereas non-vegetation consists of all remaining categories, including buildings, roads, barren land, and water bodies. The following preprocessing steps have been performed to tackle the aforementioned problems. All the images are resized into a uniform size that is compatible with the neural network input. Normalization of pixel values facilitates efficient conver- gence during the training process. Data augmentation methods such as flipping horizontally, flipping vertically, rotating, and scaling randomly have also been implemented to enhance the generalizability of the model. This data set is then partitioned into training and validation sets to avoid bias in the performance assessment of the models.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. U-Net Architecture

The U-Net is a type of convolutional neural network used for semantic segmentation [1]. The U-Net is symmetric in terms of its encoder and decoder architecture; the former extracts features and the latter localizes them.

The encoder uses several rounds of convolution, after which comes ReLU activation and max-pooling. Each time max- pooling is applied, the spatial dimensions become smaller, and the feature channel size increases, making learning possible for the neural network.

In the decoder, there is up-sampling that makes use of transposed convolution, where the spatial resolution of the feature maps is recovered. An important feature of U-Net architecture is the use of skip connections, where features maps from the encoder get concatenated with those in the decoder.The mathematical notation for the convolution function is provided below:

$$y = f(W * x + b) \quad (1)$$

Here, x is the input feature map, W is the convolution filter weight, b denotes the bias, while f refers to the activation function.

Although U-Net is a good technique when it comes to capturing fine-grained structures, it falls short in terms of modeling long-range dependencies due to its small receptive field.

C. Architecture of DeepLabV3+

The DeepLabV3+ architecture is a highly sophisticated semantic segmentation network that belongs to the Deep Lab series and is characterized by an encoder-decoder framework [2]. This architecture aims to overcome the limitations of the conventional convolutional networks by using atrous or dilated convolution.

Atrous convolution provides a large receptive field without decreasing the spatial resolution, thereby allowing the model to learn contextual information at different scales. Atrous convolution is represented mathematically as:

$$y[i] = \sum_k x[i + r \times k]w[k] \quad (2)$$

where r represents the dilation rate.

An important module used in DeepLabV3+ is the ASPP module, which employs several parallel atrous convolution operations with varying dilation factors. This gives the network the capability to learn from multi-scale features, making it efficient to segment object features regardless of their size.

The encoder performs high-level features learning by em- ploying a deep backbone architecture (like the ResNet model). On the other hand, the decoder enhances segmentation bound- aries by combining both high-level and low-level features. In comparison to earlier architectures, this provides higher accuracy in detecting boundaries.

In summary, DeepLabV3+ is very efficient in complicated scenes due to the network's capability to analyze multi-scale contexts.

D. Training Parameters

Training is done on both U-Net and DeepLabV3+ with a supervised learning technique, where labeled pixel data is used. Categorical cross-entropy is used as a loss function since it quantifies the divergence between the predicted and the actual distribution:

$$L = -\sum y log(\hat{y}) \quad (3)$$

Here, y denotes the ground truth label, while $\hat{y}$ denotes the probability of prediction.

The use of the Adam optimization algorithm is recom- mended because of its adaptability to different learning rates to facilitate quicker convergence. The learning rate scheduler is introduced to decrease the learning rate when training performance stagnates.

Multiple epochs are employed to train the model, using validation data to track training progress and avoid overfit- ting. Techniques such as batch normalization and dropout are applied to increase generalizability.

Performance evaluation is done using various metrics, in- cluding accuracy and Intersection over Union (IoU), a critical metric in the segmentation task, as it calculates the degree of overlap between the predictions and the ground truth:

$$IoU = \frac{TP}{TP + FP + FN} \quad (4)$$

In the equation above, TP stands for true positive, FP for false positive, and FN for false negative.

Training ensures both models are tested under the same conditions.

IV. RESULTS AND DISCUSSION

A. *Qualitative Analysis*

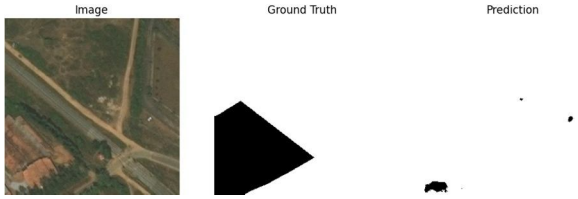


Fig. 1. U-Net segmentation results

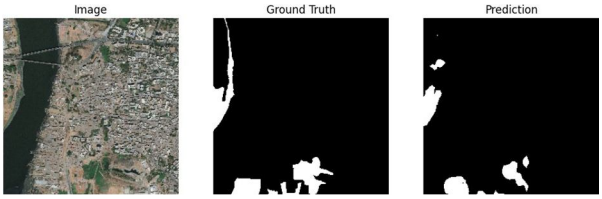


Fig. 2. DeepLabV3+ segmentation results

Qualitative results also reveal obvious differences between the two methods. Both models are assessed using the binary segmentation mask (i.e., vegetation/non-vegetation). As seen from the results obtained by the U-Net network, it is im- possible for the network to segment entire objects, and the segmentation mask consists of pieces. Such results can be explained by the inability of the model to take into account global information about the scene.

As for the DeepLabV3+ method, it provides significantly better results in terms of visual quality. The predicted seg- mentation mask is much more coherent with the ground truth mask. This result is due to the fact that the network uses atrous convolution along with the ASPP block, which makes it possible to extract multi-scale information from the scene. Overall, DeepLabV3+ shows higher visual performance, especially when dealing with complex objects.

TABLE I
PERFORMANCE COMPARISON OF U-NET AND DEEPLABV3+

Model	Accuracy (%)	IoU
U-Net	85.2	0.52
DeepLabV3+	91.8	0.81

The findings obtained by using the quantitative analysis method further confirm the qualitative findings. For instance, Deconvolution performs much better than Unet as evidenced by better accuracy scores and a far better IoU score.

However, despite getting relatively good results in terms of accuracy, U-Net fails to match the performance of Deconvo- lution in terms of IoU.

B. *Discussion*

The gap in performance in terms of accuracy between U- Net and DeepLabV3+ can be linked to the architecture of each network. While U-Net is efficient at encoding local features through skip connections, it fails in global contextual reasoning, thus affecting its performance in complex scenes. On the other hand, DeepLabV3+ uses atrous convolution and ASPP to encode both local and global features. Hence, DeepLabV3+ performs well on high-resolution imagery where objects differ in terms of sizes and shapes. Nonetheless, it is important to mention that DeepLabV3+ needs more computational power than U-Net.

V. CONCLUSION

The current research provided a comparative analysis of U- Net and DeepLabV3+ networks utilized for high-resolution semantic image segmentation. Both algorithms were tested in qualitative and quantitative ways, considering the accuracy and IoU indicators.

Based on the results of the experiment, it is possible to conclude that DeepLabV3+ demonstrates better performance, especially when used in complicated situations. Due to the employment of atrous convolution and ASPP layers, this network captures multi-

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

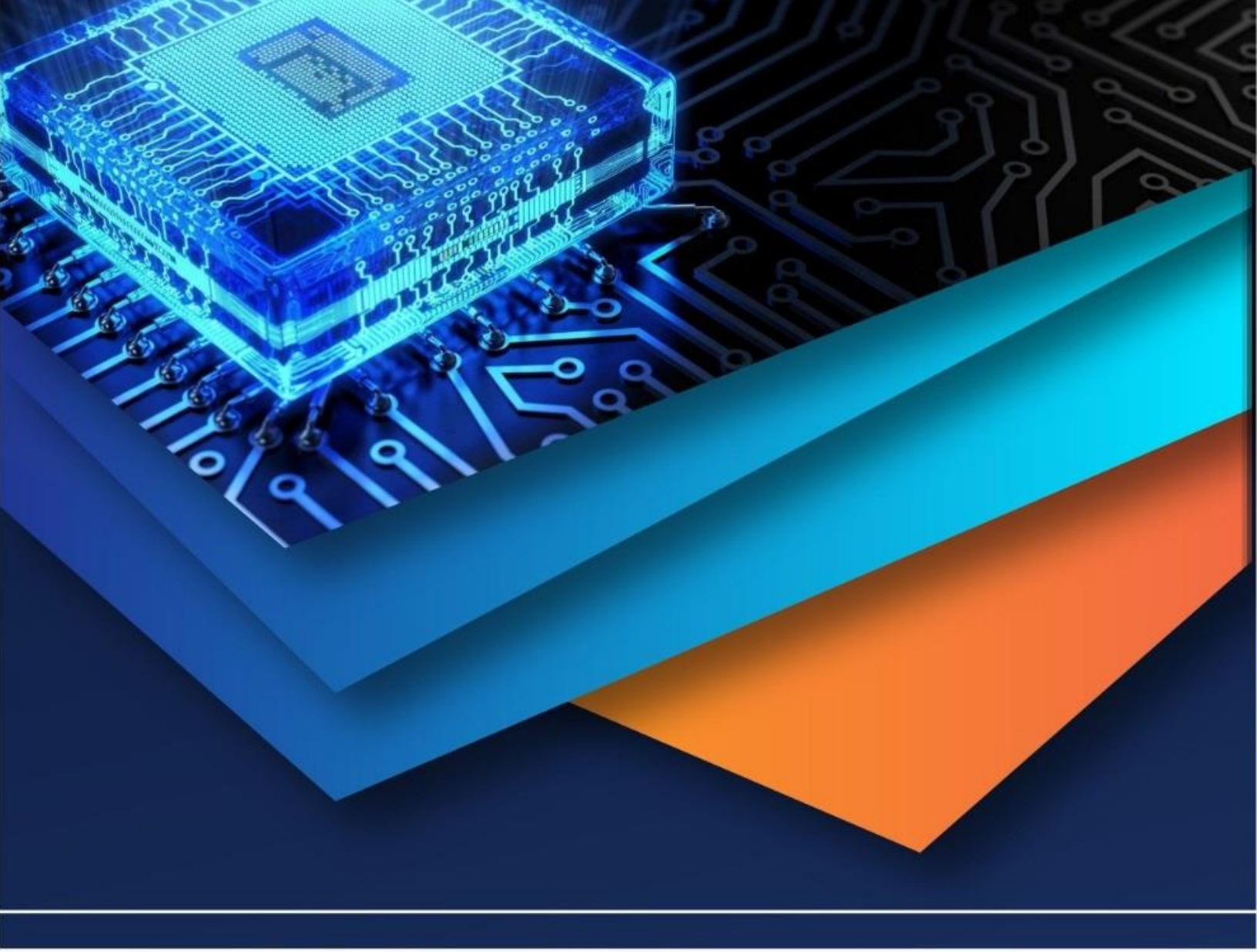
scale information, which improves the efficiency of object recognition and boundary detection. Meanwhile, U-Net successfully identifies fine-grained data but fails to grasp global information, which leads to fragmented results in difficult conditions.

Thus, DeepLabV3+ represents an optimal choice for com- plicated high-resolution tasks, whereas U-Net is more advan- tageous from the computational perspective. For future investigations, it would be reasonable to con- centrate on developing hybrid architectures based on the considered techniques. Also, one may try to optimize exist- ing approaches and implement innovative loss functions or attention modules.

REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, ”U-Net: Convolutional Networks for Biomedical Image Segmentation,” in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015, pp. 234–241.
[2] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ”Encoder- Decoder with Atrous Separable Convolution for Semantic Image Seg- mentation,” in Proc. European Conf. Computer Vision (ECCV), 2018, pp. 801–818.
[3] J. Long, E. Shelhamer, and T. Darrell, ”Fully Convolutional Networks for Semantic Segmentation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.
[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, ”Pyramid Scene Parsing Network,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
[5] I. Demir et al., ”DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2018, pp. 172–181.
[6] X. Liu, Y. Deng, and T. Li, ”Deep Learning for Image Segmentation: A Survey,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 42, no. 7, pp. 1737–1754, 2020.
[7] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos,, Image Segmentation Using Deep Learning: A Survey,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523–3542, 2021.
[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, ”SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.
[9] K. He, X. Zhang, S. Ren, and J. Sun, ”Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[10] X. Zhu et al., “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” IEEE Geoscience and Remote Sensing Magazine, vol. 5, no. 4, pp. 8–36, 2017.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)