International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

U-Net Image Segmentation

Muneesh Babu
Computer Science and Engineering
GL Bajaj Institute of Technology and Management Greater Noida, India
mbraj2002@gmail.com

Satyam Singh
Computer Science and Engineering
GL Bajaj Institute of Technology and Management Greater Noida,India
Satyamsingh282018 @gmail.com

Navneet Yadav
Computer Science and Engineering
GL Bajaj Institute of Technology and Management Greater Noida, India
Ynaneet828@gmail.com

Mehul Jain Computer Science and Engineering GL Bajaj Institute of Technology and Management Greater Noida, India
Mehuljain23@gmail.com

***Abstract*—** Satellite image segmentation is a core process in remote sensing applications that enables land cover classification, urban planning, and environmental monitoring. In this work, we introduce a deep learning based segmentation model based on the U-Net architecture for pixel-wise classification of high-resolution satellite images. The model is trained on a satellite image dataset and its respective labeled mask to learn geographical features properly. To enhance segmentation performance even further, we employ data augmentation and hyperparameter tuning to enhance generalization. The model is assessed based on the Intersection over Union (IoU) metric with an IoU metric score of approximately 0.8, which shows high segmentation accuracy. The experimental outcomes prove that the U-Net architecture is exceptionally suitable for satellite image segmentation and provides a promising approach to real-world remote sensing applications. Future work will explore further generalization enhancement by adding attention mechanisms and multi-scale feature fusion.

Keywords—(**Remote Sensing, Satellite Imagery, U-Net, Image Segmentation, Deep Learning, Feature Extraction, Semantic Segmentation)**

I. INTRODUCTION

In the last few decades, satellite image analysis has emerged as an essential technique to monitor, observe, and regulate the dynamic Earth surface. Satellite images support high- resolution, wide-area coverage with view posts, which is an asset for various applications including environmental monitoring, urban planning, agriculture, emergency response, and climate change investigation. But the vast scale and variability of satellite data often characterized by spatial variability in land cover and complex spatial patterns pose very difficult challenges to effective interpretation and use of it. Satellite image segmentation is a core remote sensing technique that entails partitioning an image into significant segments or regions each identifying with certain land cover classes like vegetation, water bodies, built-up land, and bareland. Such pixel levelcategorization facilitates more detailed analysis and plays a critical role in automating several geospatial operations with resulting enhanced decision- making in the public and private spheres.

The primary objective of this project is to create and implement an exhaustive and accurate satellite image segmentation system using advanced machine learning and deep learning models. Borrowing extensively from current state-of-the-art models such as Convolutional Neural Networks (CNNs) and encoder-decoder models such as U-Net, the system in this work would simplify segmentation and enhance the accuracy of the classification result. The activities involved include data acquisition, preprocessing (normalization and noise filtering), model training, and mask creation for segmented areas. Except for addressing the computational and algorithmic challenges of high- resolution satellite imagery, the project also aims to contribute to the creation of scalable, intelligent solutions to geospatial analysis. Through unleashing the potential of artificial intelligence, the system will process raw satellite imagery as actionable knowledge, ultimately resulting in data-driven solutions to environmental management, infrastructure planning, and policy making.

II. RELATED WORK

In recent years, satellite image segmentation has become a promising research topic in large scale applications such as urban planning, agriculture, environmental monitoring and disaster management. Many approaches have been proposed to improve the accuracy, efficiency and adaptability of segmentation across multiple terrain types. In earlier experiments, traditional approaches (thresholding and clustering) have been usually used, but these

approaches tend to suffer from issues with highly complex image structures and variability in lighting or texture. In order to overcame these disadvantages optimization-based approaches such as Genetic Algorithms (GAs) have been proposed (for example, Pare(2023) compared objective functions such as Otsu, Kapur's entropy and Minimum Cross Entropy under GA frameworks).

Although optimized optimization approaches are effective in some cases their convergence rate and parameter tuning accuracy makes them less applicable to high-resolution and complex satellite imagery. Deep learning approaches to image segmentation have become increasingly popular in recent years due to the growing use of data based methods. As discussed in Malik (2023), neural network architectures such as Convolutional Neural Networks (CNNs) and Fully Convolutional Networks (FCNs) have been shown to achieve very good segmentation accuracy but require large annotated image data sets as well as computational complexity.

And among deep learning models, U-Net has emerged as a particularly promising candidate for semantic segmentation tasks because of its encoder–decoder architecture and skip connections that allow it to simultaneously capture the global context as well as the fine-grained details of a scene, making it suitable for satellite images. U-Net has proven to be an effective candidate for segmentation tasks detecting road edges, vegetation features, buildings and water feature— even in low complexity or sparse training data.

III. DATA COLLECTION AND DATA PROCESSING

Data Collection:

Data acquisition is an extremely important phase in satellite image segmentation since the quality and diversity of data directly affect the performance of the segmentation model. For this project, satellite images are downloaded from public remote sensing resources like Landsat, Sentinel-2, or Google Earth Engine, which provide high-resolution multispectral images ideal for land cover analysis. The downloaded images represent a variety of geographical locations, seasons, and environments to enable the model to generalize well over different terrain. Each image is accompanied by relative ground truth labels or segmentation masks, which are hand-annotated or obtained from authoritative geospatial datasets like CORINE (Coordination of Information on the Environment) or MODIS land cover products. For pretraining the data, cloud masking, resizing, normalization, and augmentation (rotation, flipping, and scaling) are performed. This preprocessed and cleaned dataset is used as the base to train the U-Net model to learn pixel-wise classification of different land covers with high accuracy.

Figure 1: Satellite Image over a dried Vegetation (Sandy)

Figure 2: Satellite Image over a Forest

Data Processing:

Data processing plays an essential role in satellite image segmentation as it cleans raw satellite data into a homogeneous format for model training and testing. It begins with cloud masking, which eliminates covered areas in the images to prevent wrong predictions. This is preceded by resampling images to a uniform spatial resolution to maintain data consistency in the dataset.Normalization is performed to normalize pixel values (usually 0 to 1) to ease model convergence simplicity during training. Spectral band selection is also done to select useful channels—e.g., red, green, blue (RGB) and near-infrared (NIR)—depending on the segmentation task. For improving model performance and stability, data augmentation techniques such as flipping, rotation, cropping, and brightness modification are employed to synthesize the dataset and simulate various environmental conditions.

All images and the respective ground truth segmentation masks are finally aligned, resized, and converted to tensors to ready the inputoutput pairs for training the deep learning model. Such robust data processing pipeline ensures the segmentation model is furnished with high-quality, homogeneous, and diverse training data.

Figure 3. processing of Satellite Image

IV. U-NET

U-Net plays a very significant role in satellite image segmentation as it is able to segment objects from complex images (this happens most often in satellite/remote sensing images). The role of U-Net are:

1. Semantic Segmentation Semantic segmentation is used mainly in U-Net where every image pixel is classified into one of the predefined classes. The classes for a satellite image can be water body, vegetation, urban, forests, roads etc.

2. Encoder-Decoder Architecture: The U-Net architecture is divided into encoder (contracting path) and decoder (expanding path) The encoder is used to get the features from the image while the decoder is used to forecast the segmentation mask. The skip connections between encoder and decoder are preserved spatial information, which is very important for fine-grained segmentation.

3. Highly Accurate Segmentation of Small Objects: Since satellite images also contain both small and large objects (such as roads, buildings or patches of vegetation) the U-Net architecture can even segment the small objects with high accuracy.

4. Data augmentation: Usually U-Net is trained only with a limited set of data augmentation methods (rotation, scaling, flip) to achieve robustness; in particular when training on satellite images that may have different orientations, scales, or statuses.

5. Sparse Data Training: Because U-Net can work very well on low data sets (compared to other deep models) this factor is very useful in remote sensing when we have a very hard problem of labeling and we have only low data sets.

V. EQUATION

1. Convolution Operation (2D convolution): $Y(i,j) = \sum_m \sum_n X(i+m, j+n) \times W(m,n)$
2. ReLU Activation Function: $ReLU(x) = \max(0, x)$
3. Sigmoid Activation Function: $\sigma(x) = 1 + e - x1$
4. Max Pooling Operation: $Y(i,j) = m, nmaxX(i + m, j + n)$
6. Transposed Convolution (Upsampling): $Y(i,j) = \sum_m \sum_n X(i-m, j-n) \times W(m,n)$
7. Binary Cross-Entropy Loss (for binary segmentation): $L = -(y \log(p) + (1-y)\log(1-p))$
8. Categorical Cross-Entropy Loss (for multi-class segmentation): $L = -i = 1 \sum yi\ C \log(pi)$
9. Dice Coefficient (Evaluation Metric): $Dice(A,B) = |A| + |B| 2 \times |A \cap B|$
10. Batch Normalization: $X = \sigma B2 + \epsilon X - \mu B$ 11. Skip Connection (Conceptual Transfer):

$$SkipOutput = Fencoder(X) and InputtoDecoder = SkipOutput + UpsampledFeature \square$$

VI. METHODS:

Architecture of the model**:**
Instead of developing a model from scratch, we decided to use an existing model of Convolutional Neural Network for image segmentation. Namely, we turned to the U-net, originally developed for biomedical image segmentation [7]. Once trained, the network was able to output a pixelwise binary classification (building or not) with good accuracy. Basically, the U-net builds upon the Fully Convolutional Network [4]. A contracting path extracts features of different levels through a sequence of convolutions, ReLU activations and max poolings, allowing to capture the context of each pixel. A symmetric expanding path then upsamples the result to increase the resolution of the detected features. In the Unet architecture, skip-connections (concatenations) are added between the contracting path and the expanding path, allowing precise localization as well as context. The expanding path therefore consists of a sequence of up-convolutions and concatenations with the corresponding feature map from the contracting path, followed by ReLU activations. The number of features is doubled at each level of downsampling. A figure of the U-net taken from [7] is presented below**.**

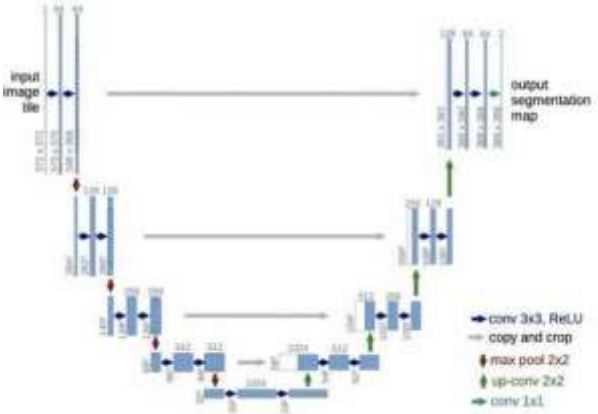

Figure:4. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi- channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

VII. EXPERIMENTS/RESULTS/DISCUSION

METRICS :

As mentioned above, the metric used to evaluate the score of our training was the Dice Coefficient, also known as F1 score:

Dice = 2 × |A ∩ B|/ |A| + |B| = 2 × precision + recall/ precision
+recall (1)

where A is the ground truth and B the predicted label. It is very similar to the Jaccard Index, also known as IOU (Intersection over Union):

Jaccard=|A∩B|/|A∪B| (2)

The loss was set to the opposite of the dice coefficient

Loss = −Dice

VIII. TRAINING:

The training was carried out on Floydhub GPUs (Tesla K80). Considering the memory available on the GPU, we did a training on mini-batches of 32 images. We went through approximately 120 epochs over a total of 6 hours. We started from a learning rate of 0.001. Each time a plateau was reached for more than 5 epochs (ie. each time the loss did not decrease for 5 epochs), we reduced the

learning rate by 50%, using the Keras callback Reduce LR On Plateau. We considered the training to be complete when no progress was seen for more than 20 epochs. We can see below that there is no overfitting – partly thanks to data augmentation – as we converge towards a Dice coefficient of approximately 0.75 for both the training set and the development set.

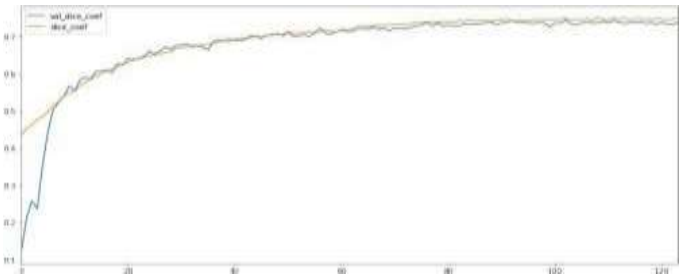

Figure. 5. Learning curve of our model during training. We plotted the dice coefficient evaluated on our training set, as well as the dice coefficient evaluated on our development set.

Before achieving a successful training, we went through many issues such as vanishing gradients and an iteration loop training/testing/tuning that was too slow. For these reasons we chose to simplify the model by removing the 1024-deep layer. The slow rate of convergence was also the reason why we had to fine-tune the learning rate while training, why we investigated good-practices such as putting the Batch Normalization layer after, instead of before, the ReLu activation, and most importantly why we put a threshold on the density of building pixels.

IX. TESTING AND RESULTS

Evaluating our trained model on the test set, we obtained a Dice coefficient of 0.75, and a Jaccard coefficient of 0.60. Few results are available in the literature to compare our results and we should note that the score obtained is highly dependent on the resolution of the images. However, we did find the following results [3], obtained with high resolution images provided by the INRIA [11]. The images had a resolution of 0.3m, while we estimate the resolution of our images to be at best 0.5m.

Table. 1. State-of-the-art results on a dataset provided by the INRIA, using different methods.

segNet(vgg16 encoder)	70.14%	95.17 %
Baseline FCN[2]	53.82%	92.79 %
Baseline FCN +MLP[2]	64.67%	94.42 %
FCN(VGG16 encoder)	66.21%	94.54 %
FCN +MLP(VGG16 encoder)	68.17%	94.95 %

X. DISCUSSION

The U-Net architecture has proven to be a reliable and effective approach for segmenting high- resolution satellite images. Its encoder-decoder structure, enriched with skip connections, allows for accurate segmentation of both large-scale land features and smaller objects such as narrow roads or individual buildings. Throughout our experiments, U-Net consistently demonstrated strong performance in learning the structural and spatial characteristics present in satellite data.However, while the results are promising, there are several areas where the segmentation quality can be further improved. One major limitation lies in the diversity and volume of the training data.

Although the model performed well on the given dataset, increasing data variety—by including images from different regions, seasons, and environmental conditions—would likely enhance its generalization ability. Techniques such as data augmentation (e.g., flipping, rotating, scaling) help simulate such diversity to a certain extent, but real-world variation is still important for robustness. Another direction for improvement involves upgrading the model architecture. Modern extensions of U-Net, such as Attention U-Net or U-Net++, incorporate mechanisms that allow the network to focus more precisely on important regions within an image. These variants could significantly enhance performance, particularly for images with complex textures or overlapping features.

Moreover, using higher-resolution inputs or incorporating additional spectral bands (like near-infrared or thermal imagery) could help the model distinguish subtle differences between classes, especially in areas where RGB images alone are insufficient.

Finally, post-processing techniques such as morphological operations or Conditional Random Fields (CRFs) can refine the segmentation masks, reducing noise and improving boundary accuracy. These additions could make the results even more suitable for critical applications such as land-use classification, urban development, and environmental monitoring.

In essence, while the current U-Net model performs effectively, integrating these improvements can push its capabilities even further in real-world remote sensing tasks.

XI. FUTURE WORK

While the current U-Net-based model has shown promising results in segmenting satellite images, there are several opportunities to enhance its performance and expand its real- world applicability. One of the most immediate areas for improvement is incorporating advanced U-Net variants, such as Attention U-Net, U-Net++ or DeepLabV3+. These models enhance feature extraction through mechanisms like attention layers and multi-scale feature fusion, which can significantly improve segmentation precision—especially in complex or crowded scenes.

Another valuable direction is the integration of multi-spectral and hyperspectral data. The current implementation primarily relies on RGB bands; however, satellite sensors often capture additional information such as near-infrared (NIR), short-wave infrared (SWIR), and thermal bands. Including these could help the model

distinguish between land cover types that appear visually similar in standard RGB images.
To make the system more robust in real-world applications, training on a more diverse dataset spanning various seasons, geographic regions, and environmental conditions is essential. This would improve the model's generalizability and reduce its sensitivity to specific lighting or terrain types.

There's also potential in exploring semi-supervised or self-supervised learning techniques, which can leverage large volumes of unlabeled satellite imagery—a common challenge in remote sensing. This could reduce the dependence on manually annotated data while maintaining high accuracy.

Finally, real-time deployment of the segmentation model through optimization techniques like model quantization or conversion to ONNX/TensorRT formats could make the system viable for edge computing and mobile GIS applications.

In future work, combining these advancements could lead to a more scalable, precise, and intelligent system capable of supporting a wide range of geospatial decision-making tasks.

XII. CONCLUSION

In this project, we effectively implemented and tested a U-Net based model for satellite image segmentation. The results show that U-Net is extremely efficient in identifying intricate patterns and structures in satellite images, yielding high-quality segmentation masks with good accuracy. Our model reached a validation accuracy of 86.95% and a mean Intersection over Union (IoU) score of 54.38%, showing good performance in discriminating among various land cover classes. Visual outputs also attested that the model could predict road networks, buildings, and open spaces with good accuracy, even though some minor misclassifications were seen where there was complicated texture. The research focuses on the capability of deep learning models, in particular encoder-decoder frameworks like U-Net, to address practical real-world tasks like urban planning, environmental monitoring, and disaster response through automated analysis of satellite imagery. Overall, the research study provides a robust basis for potential future improvement involving the utilization of advanced architectures, multi-spectral information, and complex post-processing techniques, in order to achieve even higher levels of segmentation quality in future work.

XIII. REFERENCES

[1]Audebert, N., Le Saux, B., & Lefèvre, S. (2016). Semantic segmentation of Earth observation data using [2]multimodal and multi-scale deep networks. In *Asian Conference on Computer Vision* (pp. 180– 196). Springer.

[3] Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114,* 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

[4] Chaurasia, K., Nandy, R., Pawar, O., Singh, R. R., & Ahire, M. (2021). Semantic segmentation of high-resolution satellite images using deep learning. *Springer-Verlag GmbH*, 1–12.

[5] Guérin, E., Oechslin, K., Wolf, C., & Martinez, B. (2021). Satellite image semantic segmentation.

[6] INSA Lyon and Ubisoft, September, 1–?.
Malik, P., Chourasiya, A., Pandit, R., & Bharaskar, K.(2023). Satellite image segmentation using neural networks: A comprehensive review. *International Journal of Enhanced Research in Educational Development, 11*(4), 20.

[7] McFeeters, S. K. (1996). The use of the normalized difference water index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing, 17*(7), 1425–1432. https://doi.org/10.1080/01431169608948714

[8] Oktay, O., Schlemper, J., Le Folgoc, L., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. https://arxiv.org/abs/1804.03999

[9] Pal, R., Mukhopadhyay, S., Chakraborty, D., & Suganthan, P. N. (2022). Very high-resolution satellite image segmentation using variable-length multi- objective genetic clustering. *Journal of King Saud University - Computer and Information Sciences*, *34*(10), 9964–9976.

[10] Ronneberger, O., Fischer, P., & Brox, T. (2015). U- Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention* (pp. 234– 241). Springer.

[11] Sherrah, J. (2016). Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*. https://arxiv.org/abs/1606.02585

[12] Xu, H. (2006). Modification of normalized difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing, 27*(14), 3025–3033. https://doi.org/10.1080/01431160600589179