RESEARCH ARTICLE OPEN ACCESS

Comparison of Semantic Segmentation for Land Cover using FCN and UNet Architecture

*Muna, Ali , **Saif . Fatooh,***Fatima Alzhra ,Mohammed
Computer Science
AlRibat University -Sudan

ABSTRACT

Semantic segmentation is a key task in computer vision that involves assigning a specific label to each pixel in an image. Effective semantic segmentation depends on understanding the broader context of a scene rather than relying solely on textures or colors.

In this research, we address the task of semantic image segmentation with Deep Learning architecture and made two main contributions that are experimentally shown to have substantial practical merit. Two public datasets were used to train a multi-layer of a fully convolutional neural network (FCN) and a deep learning architecture U-Net.

For the first dataset (Dubai Dataset), was trained only with UNet and different backbones—ResNet34, InceptionV3, and VGG16 were implemented. The MIoU for the backbone of ResNet34 is 74% and for the backbone Inception V3 is 70%. Moreover, the second dataset (the Bhuvan Satellite Dataset) was trained against both FCN and UNet architecture and the achieved performance was 80.3% ,89.4% respectively. For instance, the U-net architecture achieves high performance on a satellite imagery and it encounters the imbalance of classes in the dataset and achieved a very fast segmentation results and all the trained was done on Google Backend.

Keywords :- U-Net , FCN, CNN

I. INTRODUCTION

Semantic image segmentation is a key task in the field of computer vision that involves assigning a specific label to each pixel in an image. This process allows computers to understand and interpret images at a detailed and granular level, making it a fundamental step in many applications ranging from autonomous driving and medical imaging to augmented reality and landscape analysis. Semantic segmentation plays an important role in image understanding and essential for image analysis tasks. It has several applications in computer vision &artificial intelligence - autonomous driving [1], [2], robot navigation [3], remote sensing [4]; In medical sciences - medical imaging analysis [5] etc.

Semantic image segmentation, also called pixel-level classification which unlike object detection or image classification, semantic segmentation requires the classification of every pixel in an image. Each pixel is assigned a class label that represents what object or feature it belongs to, such as "car", "tree", "building", "road", etc. Two other main image tasks are image level classification and detection. according to its specific sectors or groups. Classification is easy.

Effective semantic segmentation relies not just on recognizing textures or colors, but also on understanding the larger context of scenes. For instance, pixels resembling the texture of a road might actually belong to a building roof. Thus, understanding the placement and context of various objects in relation to each other is crucial. A significant challenge in semantic segmentation is accurately defining the boundaries between different objects. Precisely determining these edges is critical for many applications, such as medical imaging, where the exact delineation of tissues can affect diagnostic outcomes. The recent works in deep learning dealing with semantic segmentation have been significantly improved by using neural networks. Neural networks have a long history since the 1940s and they did not get much of the attention of researchers until 1990s [1]. Neutral networks made huge progress because of large amount of data is available thanks to the rise of digital cameras, cell phone cameras, and the computing power, which is getting faster as GPUs become general purpose computing tools. Deep neural networks are very effective in semantic segmentation that is labelling each region or pixel with a class of objects/non D objects.

2. LITERATURE REVIEW

The researcher [6] segmented the images with a standard fully convolutional network (FCN) and initialize convolutional layers with Glorot uniform [initializers. Specifically, VGG-16 is taken as the backbone, and the outputs of the last two convolutional blocks are upsampled to the original resolution and fused with an elementwise addition. The fused feature maps are finally fed into a convolutional layer, where the number of filters is equivalent to the number of classes. In the training phase, all weights are trainable and updated with Nesterov Adam [19], using $\beta 1$ = 0.9, $\beta 2$ = 0.999, and _ = 1e−08 as recommended. To train the network, they defined the loss as (2), and λ is set experimentally to 0.1 and 0.01 for the Vaihingen and Zurich Summer data sets, respectively. Trade-off parameters, α, β, and γ, are set as 0.5, 1.5, and 1, to ensure that: 1) the regularizes governing feature and spatial relations are balanced and 2) neighbouring pixels in the image space receive more attention. The network, as well as FESTA, is implemented on TensorFlow and trained on one NVIDIA Tesla P100 16-GB GPU for 100k iterations. The size of the minibatch is set as five during the training procedure. In the training phase, they used a sliding window to crop training images into 256 × 256 patches, and its stride is set to 64 pixels. Besides, no class-dependent configurations are considered. In the test phase they employed dense CRF to refine predictions before calculating metrics. For instance, another work [7] involves an application of Fully Convolutional Networks (FCNs) to the task of semantic labelling on the Vaihingen dataset from the ISPRS 2D semantic labelling contest. The dataset includes 33 tiles of aerial orthophotomosaic images with three spectral bands (red, green, and near-infrared) and a digital surface model (DSM). It covers a total of approximately 170 million pixels particularly challenging and useful for remote sensing image analysis due to its high resolution and the inclusion of various land cover types. The dataset contains roughly 1.7 × 108 pixels in total, but ground truth is only released for half of the tiles, which are designated for training and validation. For the remainder, the ground truth is withheld by the organizers for objective evaluation of submitted results. The images are rich in detail, with a GSD of 9 cm. Categories to be classified are *Impervious Surfaces, Buildings, Low Vegetation, Trees, and Cars*. In order to keep our pipeline automated to the largest possible degree, they refrain from any pre-processing that would require human intervention or selection of data-specific hyper-parameters (such as DSM-to-DTM filtering, or radiometric adjustments), but rather feed the data provided by the benchmark directly into the network. They trained different models such as FCN-Pascal of [8] was pre-trained on Pascal VOC, FCN-ImageNet of [9] was pre-trained on the ImageNet data set, and FCN-Places of [10] was pre-trained on the Places data set. All models are fine-tuned on our aerial data without any changes to their network architectures. Label prediction on the four images of the hold-out data subset They reach 88.4% overall accuracy with the FCN ensemble alone, and 88.5% with FCRF post-processing, we reach the second best overall result, 0.6 percent points below the top-performing method. The research showcases how leveraging pre-trained networks on diverse datasets and fine-tuning them on specific tasks like aerial image segmentation can yield high-quality results. This approach also underscores the importance of a robust testing framework, as provided by the Vaihingen dataset, for developing and validating image segmentation algorithms in remote sensing. Moreover, their method works particularly well on the smaller *tree* and *car* classes and, with 86.9%, reaches the highest average F1- score, 1 percent point higher than the nearest competitor.

Another work was proposed by using the algorithm U-Net[8], is particularly effective due to its unique architecture that efficiently handles spatial hierarchies for pixel-wise classification, making it highly suitable for tasks like segmenting aerial images. It works to give different color to each category, and it is possible to assign a category to each pixel in the image, such as the label with the word car or plane, and this is called semantic. The process of adding the corresponding pixels is performed directly with the previous operations, which are very smart operations, so it differs from FCN. It has preserved its spatial information because it contains the copy & crop process. The result of a network can be trained end-to-end from very few images and outperforms the prior best method (fully convolutional network) .Moreover, the network is fast. Segmentation of a 572x572 images taken less than a second on a recent GPU. As a result, a matrix of the same dimensions for the input image was obtained, so U-Net was applied to the aerial images, and the process of prediction of pixels in the border region was accurate and fast through the results that was applied by the Pytorch library.

2.1 Fully Convolutional Network (FCN)-Based Approaches

CNNs were first introduced in 1998 by [9] who developed a model to recognize digits and zip codes, showing promising performance. It was the first step toward automatic feature extraction; this article showed CNN's ability to eliminate the need for hand-crafted feature extractors. However, memory and hardware limitations, as well as the inaccessibility of large training data sets, delayed progress on this approach for some, with progress in hardware and memory, [9] proposed a deeper net to learn multi-level hierarchies of features and set the path of deep learning research on computer vision tasks to find methods that can automatically learn good feature hierarchies. Since then, many models have been published. The number of variations is overwhelming; this section provides a description of the most influential and the challenges they face.A classic CNN consists of two components: Convolutional layers and fully connected layers located at a deeper level of the network. Convolutional layers operate as a floating window, are not bound to a fixed-size image, and can create feature maps of arbitrary-size. Fully linked layers, on the other hand, must have a fixed-size input. This requirement can reduce recognition accuracy for images and sub-images of arbitrary size. In this approach, the fully connected layer is removed and replaced by the fully convolutional layer, thus converting CNN to FCN. Thus, it is ensured that CNN takes images of arbitrary size as input and obtains an output of arbitrary size.

This study [10] is pioneering work in this area. In their work, they have adapted classification networks such as "VGGNet", "GoogleNet" and "AlexNet", which have been very popular in recent years, to fully convolutional networks. The backbone network involves the primary structure of the network, which is produced for the image classification task. These structures, essentially, perform feature extraction for the task of semantic segmentation. These classification networks are called backbone networks within our study.

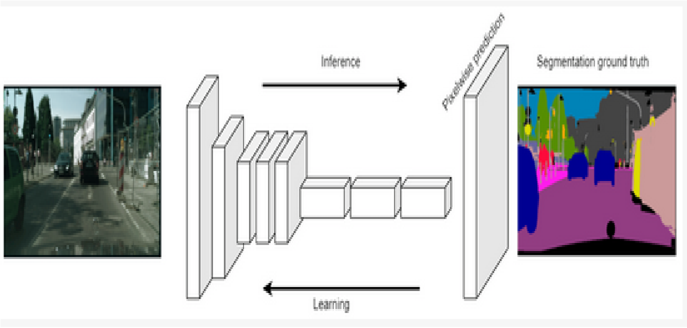


Figure 2-1 Framework of FCN.

The basic FCN-based method **[10]** has major limitations for semantic segmentation. Low-level features from shallow layers of the network contain more detailed information at higher resolution, i.e., they have richer spatial details. High-level features from the deeper part of the network have higher semantic information, but due to the pooling layer, the feature map resolution is lowered; that is, spatially detailed information is lost. For this reason, an encoder-decoder network has been developed to extract the features, which reduces the spatial size and then gradually recovers the spatial size of the features obtained through upsampling. Another restriction is that the FCN has a predefined fixed-size receptive field because of the convolution operation. This ignores the global information in the image when it encounters an object larger or smaller than the receiving field. To use this global information, that is, to include the semantic context, methods based on generating features with larger receptive fields without sacrificing spatial resolution have been developed. Dilated convolution methods use dilated/atrous convolutions in FCN to expand the receptive field of convolutions and enable dense predictions, feature fusion methods fuse high-level low-resolution and low-level high-resolution features, thereby visibly improving performance, multi-scale methods combine multi-scale/stage features by modelling local and global information from different layers and pyramid methods significantly increase performance by expanding the receptive field by multi-resolution pyramid-based representations and methods using “Recurrent Neural Networks (RNN)” [11] and “Long Short-Term Memory (LSTM)” [12] capture long-range semantic dependencies in images. A graphical model, the “Conditional Random Field (CRF)” [13] has also been used to introduce global context into an FCN and improve output accuracy. In these studies, segmentation performance is often improved by applying the CRF to the CNN as a post-processing step or by fully integrating the CRF into the CNN to train the entire network end-to-end [14-16]

2.2 THE U-NET ARCHITECTURE

U-Net is a widely used deep learning architecture that was first introduced in the “U-Net: Convolutional Networks for Biomedical Image Segmentation” paper. The primary purpose of this architecture was to address the challenge of limited annotated data in the medical field. This network was designed to effectively leverage a smaller amount of data while maintaining speed and accuracy.The architecture of U-Net is unique in that it consists of a contracting path and an expansive path. The contracting path contains encoder layers that capture contextual information and reduce the spatial resolution of the input, while the expansive path contains decoder layers that decode the encoded data and use the information from the contracting path via skip connections to generate a segmentation map. The contracting path in U-Net is responsible for identifying the relevant features in the input image. The encoder layers perform convolutional operations that reduce the spatial resolution of the feature maps while increasing their depth, thereby capturing increasingly abstract representations of the input. This contracting path is similar to the feedforward layers in other convolutional neural networks. On the other hand, the expansive path works on decoding the encoded data and locating the features while maintaining the spatial resolution of the input. The decoder layers in the expansive path upsample the feature maps, while also performing convolutional operations. The skip connections from the contracting path help to preserve the spatial information lost in the contracting path, which helps the decoder layers to locate the features more accurately.

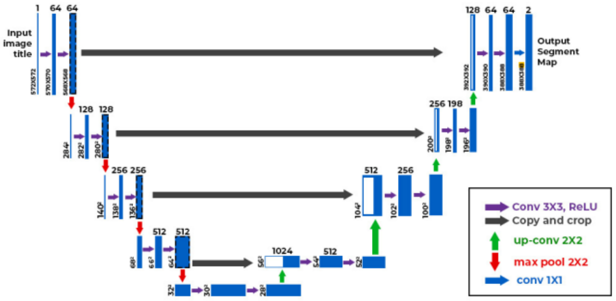


Figure 2-2 Framework of UNet

Figure 5 illustrates how the U-Net network converts a grayscale input image of size 572×572×1 into a binary segmented output map of size 388×388×2. We can notice that the output size is smaller than the input size because no padding is being used. However, if we use padding, we can maintain the input size. During the contracting path, the input image is progressively reduced in height and width but increased in the number of channels. This increase in channels allows the network to capture high-level features as it progresses down the path. At the bottleneck, a final convolution operation is performed to generate a 30×30×1024 shaped feature map. The expansive path then takes the feature map from the bottleneck and converts it back into an image of the same size as the original input. This is done using upsampling layers, which increase the spatial resolution of the feature map while reducing the number of channels. The skip connections from the contracting path are used to help the decoder layers locate and refine the features in the image. Finally, each pixel in the output image represents a label that corresponds to a particular object or class in the input image. In this case, the output map is a binary segmentation map where each pixel represents a foreground or background region.

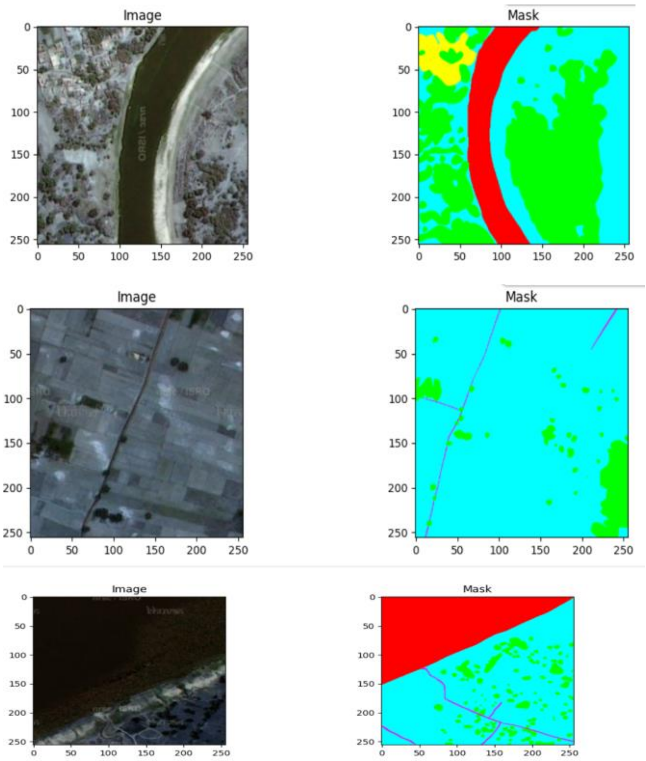
3. EXPERMINT AND RESULTS

The experiment datasets, the Bhuvan Satellite Dataset[17] which includes a collection of satellite images and corresponding segmentation masks. The segmentation masks provide a pixel-level classification for five distinct land cover classes: vegetation, urban areas, forest, water bodies, and roads. The dataset consists of satellite 2D images of Varanasi, a city located in the northern part of India, in the state of Uttar Pradesh, with coordinates ranging from 25.3° to 25.5° N latitude and 83° to 83.2° E longitude. It comprises a collection of high-resolution images capturing the Earth's surface. These images were obtained from the Indian Remote Sensing Satellite (IRS) and were processed and made available through the Bhuvan Geo Platform, which is managed by the Indian Space Research Organization (ISRO) {kaggle].The dataset includes pixel-based_mask which contains pixel-level segmentation masks that classify each pixel in the satellite images into one of the land cover classes, such as vegetation, urban areas, forest, water bodies, and roads. The test_mask and train_mask which contain the manually generated segmentation masks specifically for creating the pixel-based mask,test_image and train_image which contain the corresponding satellite images that are used for testing and training the land cover classification models. These high-resolution 2D images provide visual information about the Earth's surface in Varanasi. It consist Land Cover images and annotated with pixel-wise semantic segmentation of 5 classes. The total volume of the datasets is 42 images for training and the masks larger tiles each tile contain 9 images and corresponding masks and 14 for test images .The datasets were trained against the architecture FCN on Colab Google Backend for semantic segmentation to use the Free GPU T4 for processing and handle the data.

Table3-1 shows the class on the Bhuvan datasets

Class	Color RGB	HEX
Urban	0, 255, 255	#00FFFF
Agriculture	255, 255, 0	#FFFF00
Road	255, 0, 225	#FF00FF
Forest	0, 255, 0	#00FF00
Water	0, 0, 255	# 0000FF

The images are in different sizes and to make all the images be the same, they have to be cropped to the size divisible by 224 and extract the patches. The patches size was selected to be 224 and the figure[3-1] shows the images and the corresponding mask



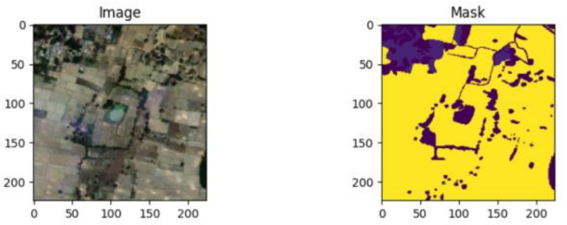


Figure 3-1 the original and the mask image for the dataset

In the preprocessing of the masks to convert them to RGB, we have detected the change of the colors and it found that a total number of colors are about 1812 .So, we the reduced the colors to class Id of [0 1 2 3 4] and then clustered the colors by K-means into 5 classes used the k-means for clustering and it found that the class imbalance is very high .Below in Figure [3-2]shown the original mask and the converted mask as Class IDs

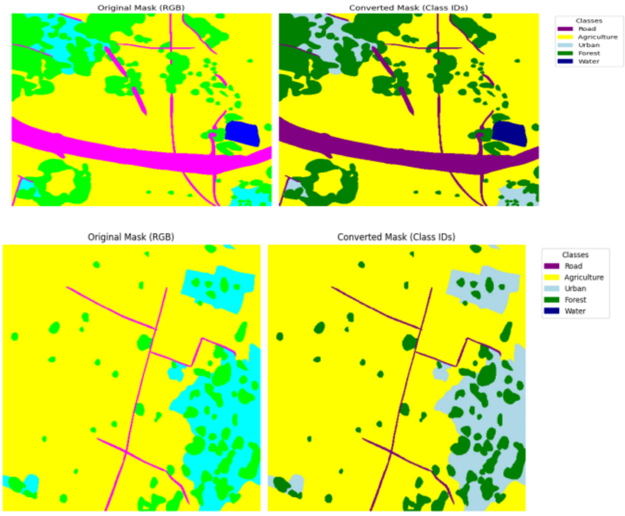


Figure 3-2 The original mask and the converted mask in IDs

Table 3-1shows the results for the losses for the Dice Loss

Loss	0.1683
Accuracy	0.803

The graph below [3-3][3-4] [3-5]demonstrates the Accuracy , the loss and the confusion matrix in which FCNs provide high accuracy in tasks like image segmentation.

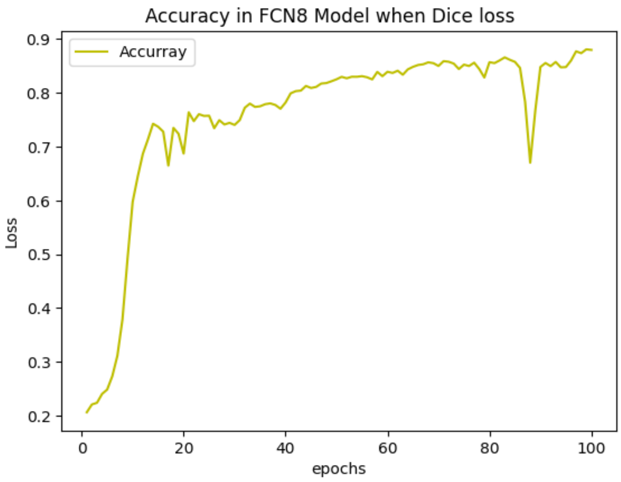


Figure 3- 3 Accuracy with Dice Loss for FCN

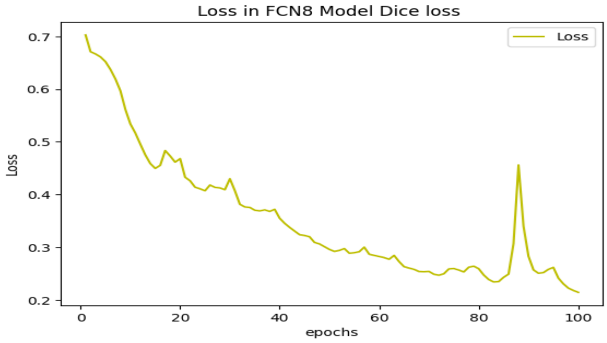


Figure 3- 4 Accuracy with Dice Loss for FCN

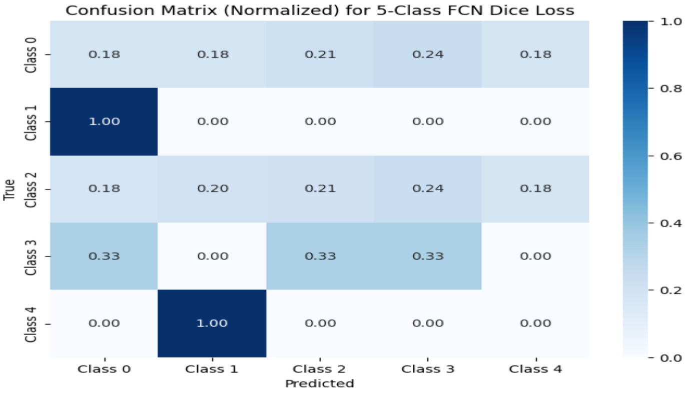


Figure 3- 5 Confusion Matrix for FCN

For the classes in the segmentation for our model FCN8 we have to compute IoU which measures how much the predicted segmentation overlaps with the ground truth mask so for each class as follows for each class c

$$IoUc = \frac{TPc}{TPc + FPc + FNc}$$

…..equation [1]

Where

TPc = cm [c,c](true positives for class c)

FPc = $\sum_j \mathrm{cm}\,[j, c] - TPc\ (False\ positives\,)$

$FNc=\sum_j \mathrm{cm}\,[j, c] - TPc\ (False\ negatives)$

And Mean IoU =

$\frac{1}{c}\sum_{c=1}^{c} IoUc$ …….equation[2]

And the achieved Mean IoU for the segmentation process is 71.6% and the IoU for forest and water was not detected as they occupy very few pixels compared to other classes, the model does not predict dominant ones instead (like urban or road). as shown in the figure[3-6]

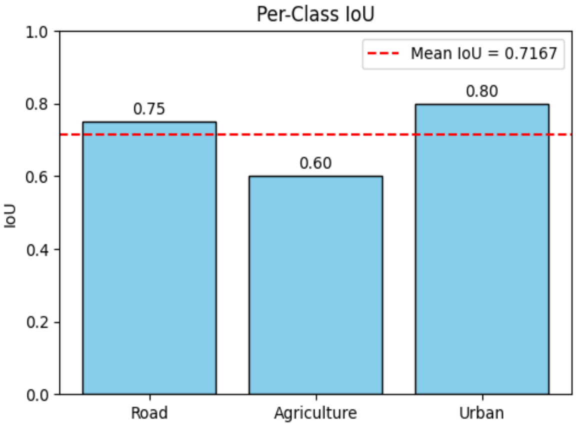
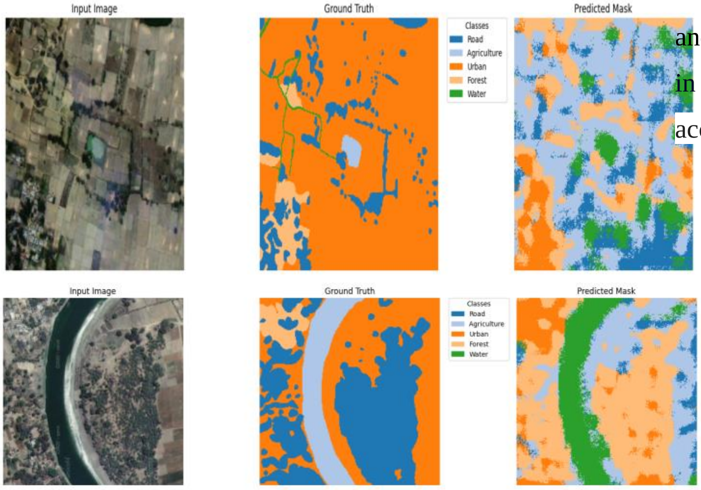


Figure 3-6 Shows the IoU for each class for FCN



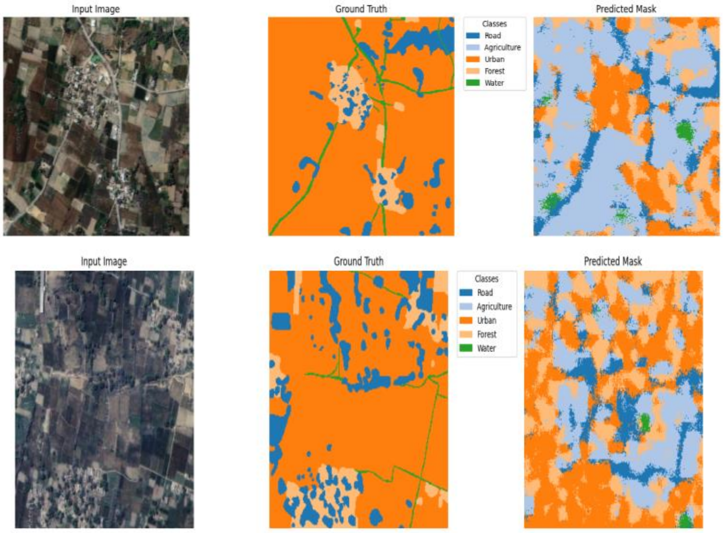


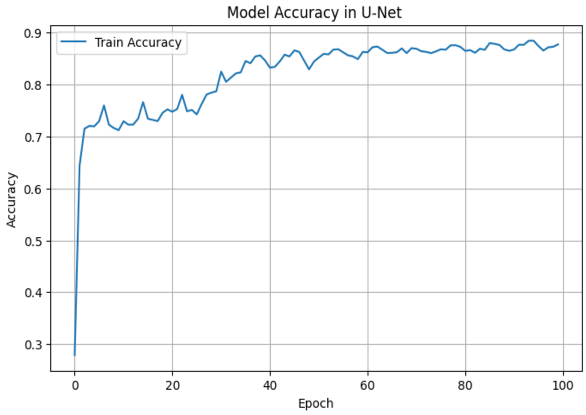
Figure 3-7 Shows the predicted mask Using FCN

The same data was trained against U-Net architecture with the size 256*256 and channel of 3 and the result as follows

Table 3-2: shows the results for the losses for the Dice Loss .

Loss	0.3127
Accuracy	0.8937

The graph below [3-7][3-8][3-9] shows the Accuracy, Loss and confusion matrix in which U-Net provide high accuracy in tasks like image segmentation and Loss. The result of the accuracy for the training of the model



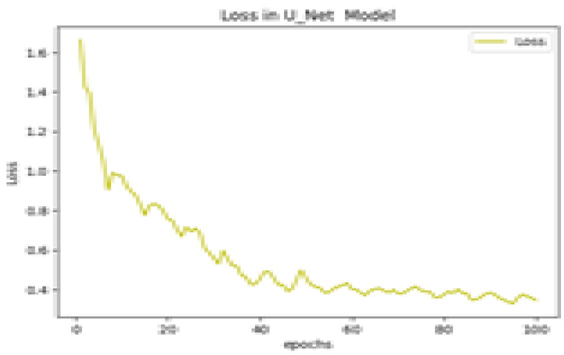


Figure 3-8 Loss in the U-Net Architecture

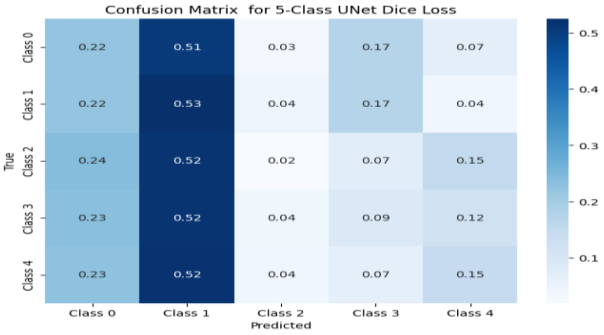
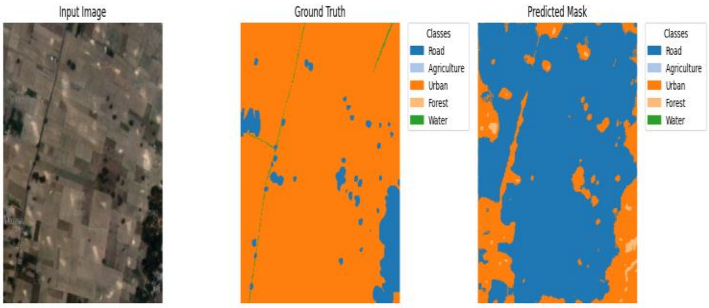


Figure 3-9 Confusion Matrix For UNet Architecture



And the following figure [3-10] demonstrate predicted mask for Unet Architecture where there is an blur in the prediction as follows

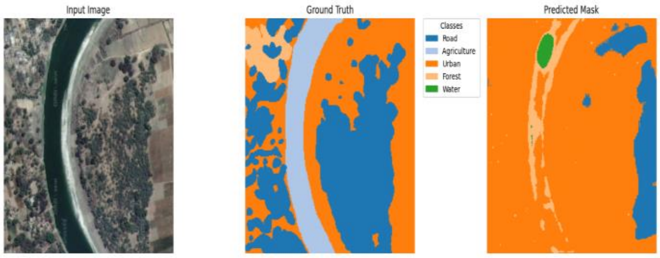
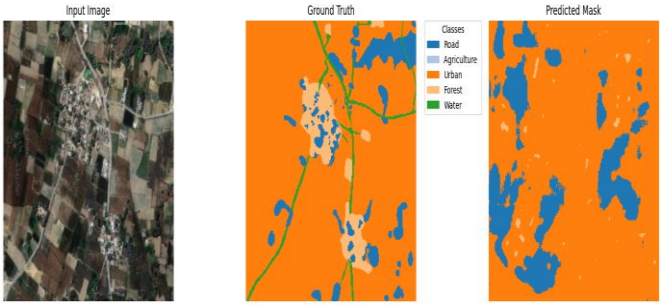


Figure 3-10 The Predicted mask In UNet -Architecture



4. CONCLUSION

This research presented the implementation and evaluation of Fully Convolutional Networks FCN-8 and the U-Net architecture for semantic segmentation of satellite imagery using TensorFlow and Keras. The Bhuvan dataset, obtained from Kaggle, was preprocessed and categorized into five major land cover classes: Road, Agriculture, Urban, Forest, and Water. The models were trained for end-to-end pixel-level prediction to extract spatial and contextual features directly from imagery data.

The experimental results demonstrate that both FCN and U-Net architectures are capable of achieving accurate semantic segmentation. U-Net outperformed FCN in terms of boundary precision and class differentiation, while FCN provided competitive accuracy with faster convergence. The study confirms that fully convolutional and encoder–decoder-based deep networks are effective for automated land cover classification, offering a powerful alternative to traditional machine learning methods that rely on handcrafted features.

Future research should consider extending these models by incorporating larger and more diverse datasets, exploring optimization strategies. These improvements will enhance generalization and further advance the use of deep learning in environmental and geospatial analysis.

5.RECOMMDATION AND FUTURE WORK

According to the results of the research here are some recommendations for future work

1.To enhance the model, it should explore some advanced segmentation architectures such as DeepLabV3+, SegNet, or transformer-based models to further improve segmentation accuracy and boundary detection.

2.To increase the size and diversity of training data, multispectral or multi-temporal images should be included from different regions can improve the robustness and generalization ability of the models.

3.To overcome class boundaries and reduce noise in the predicted segmentation maps, techniques like Conditional Random Fields (CRFs) or morphological filters should be used for post Preprocessing.

ACKNOWLEDGMENT

The author gratefully acknowledges the invaluable guidance, encouragement, and continuous support of Prof. Saif throughout the course of this research. The author also wishes to express sincere appreciation to the National Ribat University for providing the facilities and resources that made this research possible.

REFERENCES

[1] Gregor K, Danihelka I, Graves A, Rezende DJ, Wierstra D. DRAW: a recurrent neural network for image generation. Nature. 2015;521(7553):436–44. doi:10.1038/nature14236.

[2] Kamavisdar P, Saluja S, Agrawal S. A survey on image classification approaches and techniques. Int J Adv Res Comput Commun Eng. 2013;2(1):1005–9. doi:10.23883/IJRTER.2017.3033.XTS7Z.

[3] Rastafari M, Ordonez V, Redmon J, Farhadi A. XNOR-net: Imagenet classification using binary convolutional neural networks. In: Lecture Notes in Computer Science. Vol. 9908 LNCS. Springer; 2016. p. 525–42. doi:10.1007/978-3-319-46493-0_32.

[4] Sherrah J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. arXiv [Preprint]. 2016 Jun 8 [cited 2025 Aug 30]. Available from: http://arxiv.org/abs/1606.02585

[5] Hartigan JA. Clustering algorithms. New York: Wiley; 1975.

[6] Hua Y, Marcos DI, Mou L, Zhu XX, Tuia D. Semantic segmentation of remote sensing images with sparse annotations. IEEE Geosci Remote Sens Lett. 2022;19:1–5. doi:10.1109/LGRS.2021.3051053

[7] Marmanis D, Wegner JD, Galliani S, Schindler K, Datcu M, Stilla U. Semantic segmentation of aerial images with an ensemble of CNNs. ISPRS Ann Photogramm Remote Sens Spatial Inf Sci. 2016;III-3:473–480. doi:10.5194/isprsannals-III-3-473-2016.

[8] Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W. Wells and A. Frangi, eds. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, vol. 9351. Cham: Springer, pp. 234–241. doi:10.1007/978-3-319-24574-4_28

[9] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. IEEE; 2010. p. 253–6.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. IEEE; 2015. p. 3431–3440.

[11] Elman JL. Finding structure in time. Cogn Sci. 1990;14(2):179–211.

[12] Hochreiter S, Schmidhuber J. Long short-term memory. Neural
babilistic models for segmenting and labeling sequence data [Internet]. 2001 [cited 2023 Apr 5]. Available from: https://repository.upenn.edu/cis_papers/159/

[13] Chen LC, Papandreou G, Murphy K, Yuille AL. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv [Preprint]. 2016. Available from: http://arxiv.org/abs/1606.00915

[14] Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems; 2011 Dec 12–14; Granada, Spain. Vol. 24.

[15] Zheng S, Jayasumana S, Romera-Paredes B, Vineet V, Su Z, Du D, et al. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. p. 1529–1537.

[16] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015. p. 3431–3440. doi:10.1109/CVPR.2015.7298965

[17] **Bhuvan Satellite Dataset:** Indian Space Research Organisation. Bhuvan Satellite Data for Urban Classification. Indian Remote Sensing Programme, 2016. Available from: http://bhuvan-app1.nrsc.gov.in/bhuvan2d/bhuvan/bhuvan2d.php