ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume X-5/W2-2025
Unleashing the power of Geospatial & Frontier Technologies for a Sustainable Future – GIFTS Summit 2025,
1–3 September 2025, Symbiosis International University, Pune, India

Road extraction from satellite images using Deep Learning on HPC

Shankar Naik Rathod Karamtoth*, Soham Rangdal, Prakhar Verma, Kedar Nagnathrao Ghogale, Sajeevan G

Centre for Development of Advanced Computing (C-DAC), Pune, Maharashtra, India
*kshankar@cdac.in

Keywords: Satellite Imagery, Deep Learning, Semantic Segmentation, U-Net, Remote Sensing.

Abstract

Road extraction from satellite imagery is required for various applications, such as infrastructure planning, city development, transportation, and disaster response planning. Road extraction is treated as semantic segmentation problem where U-Net is proven to be highly effective. However, the standard U-Net model consists approximately 31 million parameters, which leads to high computational costs and resources. To address this challenge, we have modified U-Net architecture by integrating Depth-wise Separable Convolutions (DSC) that reduced the number of trainable parameters down to 3.5 million and improved the model ability to capture road features. For this work, we utilized a combination of open-source satellite imagery datasets and a custom dataset. This dataset contains 53877 high-resolution satellite image and road mask pairs. Comparative result shows that the U-Net with DSC achieves a mean accuracy of 0.956 and a mean Intersection over Union (mIoU) of 0.626, outperforming the standard U-Net, which recorded a mean accuracy of 0.949 and mIoU of 0.581. These models are trained on C-DAC PARAM Siddhi-AI high-performance computing (HPC) infrastructure. The prediction results improve the road continuity and clarity, making the model suitable for national initiatives such as PMGSY National GIS. This model integration can greatly enhance the efficiency of planning, monitoring, and execution of road infrastructure projects.

1. Introduction

Roads are crucial for economic development and social well-being. For effective planning and management, it is necessary to digitise and regularly update road networks within Geographic Information System (GIS) platforms. The traditional methods for road mapping such as manual surveys and visual interpretation are tedious job and time-consuming process which is error prone. Automating road extraction from satellite imagery offers the potential to reduce manual effort while providing accurate and up-to-date data to support planning and decision making. Despite advancements in satellite imaging and computational capabilities, automated road extraction remains challenging because of the complex appearance of roads, occlusions from trees and buildings, varying road widths and diverse surface materials. Traditional image processing methods often fail to generalise across different environments. The advancement of remote sensing technologies has been significantly driven by the integration of modern deep learning methods. Several Deep Learning (DL) and Machine Learning (ML) models have been developed in this area to get road information from images, and each one works very well on its own. The goal of this work is to use deep learning-based semantic segmentation methods to automatically find road networks in satellite images.

2. Related Work

We have explored various traditional and deep learning methods for automated road extraction from satellite imagery. Earlier approaches (Barzohar and Cooper, 1996) relied on edge detection, morphological operations and thresholding to extract linear features; however, these methods often struggled with occlusions, shadows and noise leads to limiting their ability to generalise diverse conditions. Later, knowledge-based approaches incorporated geometric constraints and rule-based systems to improve generalization across various environments (Huang and Zhang, 2009). Even these methods provided initial solutions, they often struggled with variations in road appearance, inconsistent structures, and diverse terrain characteristics. To address these limitations, traditional ML (e.g., Support Vector Machines (SVMs) and Random Forests) models using leveraged handcrafted spectral and texture features to improve detection (Mokhtarzade et al., 2007). However, these models were heavily dependent on manual feature design and often lacked robustness across large-scale datasets. With the development of cutting-edge Deep Learning technology and processing resources, Deep learning technology has been increasingly popular in computer vision, multimedia and natural language processing applications. For example, Convolutional Neural Networks (CNNs) have proven to be effective in extracting contextual information from images (Krizhevsky et al., 2017). Notable CNN architectures and algorithms in this regard include ImageNet, ResNet (K. He et al., 2016), Mask-RCNN (K. He et al., 2017), and FCNs (J. Long et al., 2015).

While CNNs excel at capturing spatial patterns and high-level features, they primarily focus on image-level classification, which limits their suitability for pixel-level tasks such as road extraction from satellite imagery. Additionally, the down sampling operations used for feature extraction often result in a loss of fine spatial details, which is critical for accurately delineating narrow road structures in images. To address these limitations, Fully Convolutional Neural Networks (FCNNs) extend CNNs by substituting fully connected layers with convolutional layers, allowing pixel-wise predictions across entire images (Long et al., 2015). This allows FCNNs to perform semantic segmentation tasks, making them applicable for road extraction in satellite images. However, FCNNs often produce coarse segmentation maps due to down sampling, which can affect the clarity of extracted road boundaries. To further improve fine detail preservation in segmentation tasks, the U-Net architecture builds upon the FCNN framework by incorporating an encoder-decoder structure with skip connections (Ronneberger et al., 2015). These skip connections transfer

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-5-W2-2025-301-2025 | © Author(s) 2025. CC BY 4.0 License.

feature information from the encoder to the decoder, allowing the model to retain fine spatial details while maintaining appropriate understanding, with this the U-Net is useful for precise, pixel-wise road extraction from high-resolution satellite images, while also performing well on limited labelled datasets typically available in remote sensing applications. While the U-Net architecture is effective for precise, pixel-level road extraction, its original implementation requires over 31 million parameters, leading to high computational costs during training and inference. This computational demand poses challenges for large-scale road extraction tasks using high-resolution satellite imagery, particularly when scalability and rapid deployment are necessary. To mitigate this challenge, depthwise separable convolutions (DSC) have been incorporated into CNN architectures to reduce computational complexity while preserving the model's capability to capture indispensable spatial features (Howard et al., 2017).

3. Dataset

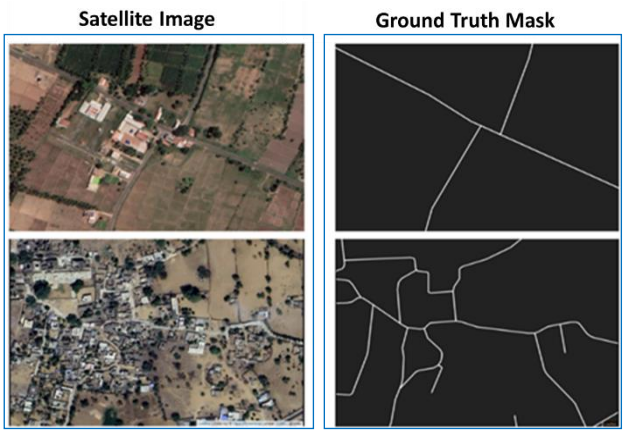

Figure 1. Satellite Image and Roads Feature Mask Image pairs

Applying deep learning methods for road extraction in Rural India faces challenges with publicly available DeepGlobe roads dataset (Demir I et al., 2018) and SpaceNet roads datasets (A. Van Etten, 2018) have limited Indian samples. We have used 53877 satellite images that were annotated to create ground truth road masks (Figure 1). For this study, we split the dataset into 70% for training, 10% for validation, and the remaining 20% for testing.

4. Methodology

The study we use a modified U-Net architecture, replacing Depthwise Separable Convolutions (DSC) in place of standard convolutions and built upon the classical encoder–decoder structure of the U-Net while incorporating DSC to significantly reduce computational cost while maintaining accuracy in extracting road networks from satellite imagery. The standard U-Net architecture contains 31 million parameters, leads to high computational demands during training and inference. By replacing standard convolutions with DSC, the parameter count can be significantly reduced to around 3.5 million, resulting in faster training and lower memory usage without compromising segmentation performance. The encoder path systematically reduces spatial dimensions and increases feature depth through four blocks, each have a 3×3 depth-wise convolution, a 1×1 pointwise convolution, batch normalisation, ReLU activation, and 2×2 max pooling for down-sampling. At the core, the bottleneck layer captures deeper background information using similar DSC operations, safeguarding required features are retained before up-sampling. The decoder path mirrors the encoder structure with four blocks, each beginning with a 2×2 transposed convolution for up-sampling, followed by concatenation with the corresponding encoder outputs through skip connections to save spatial details important for accurate segmentation. Each and every decoder block further refines the feature maps using depth-wise and point-wise convolutions with batch normalisation and ReLU activation, enabling clarity in the segmentation outputs. The network concludes with 1x1 convolution to minimise 3 channels to single channel output, followed by a sigmoid activation to produce a 256x256x1 binary segmentation road mask network.

The architecture is designed to retain high segmentation accuracy while reducing computational complexity, making it suitable for large-scale processing and deployment on low-resource field devices. Table 1 shows the configuration of the U-Net with DSC model used in this work.

Stage	Layer	Operation	Input Shape	Output Shape	Filters
Input	-	RGB Satellite Image	256×256×3	256×256×3	-
Encoder 1	DSC Block + MaxPool	Depthwise (3×3), Pointwise (1×1), BN, ReLU, MaxPool (2×2)	256×256×3	128×128×32	32
Encoder 2	DSC Block + MaxPool	Depthwise (3×3), Pointwise (1×1), BN, ReLU, MaxPool (2×2)	128×128×32	64×64×64	64
Encoder 3	DSC Block + MaxPool	Depthwise (3×3), Pointwise (1×1), BN, ReLU, MaxPool (2×2)	64×64×64	32×32×128	128
Encoder 4	DSC Block + MaxPool	Depthwise (3×3), Pointwise (1×1), BN, ReLU, MaxPool (2×2)	32×32×128	16×16×256	256
Bottleneck	DSC Block	Depthwise (3×3), Pointwise (1×1), BN, ReLU	16×16×256	16×16×256	256
Decoder 1	Upsample + DSC Block	Transposed Conv (2×2), Concat with Encoder 4, DSC (3×3), Pointwise (1×1), BN, ReLU	16×16×256	32×32×128	128
Decoder 2	Upsample + DSC Block	Transposed Conv (2×2), Concat with Encoder 3, DSC (3×3), Pointwise (1×1), BN, ReLU	32×32×128	64×64×64	64
Decoder 3	Upsample + DSC Block	Transposed Conv (2×2), Concat with Encoder 2, DSC (3×3), Pointwise (1×1), BN, ReLU	64×64×64	128×128×32	32
Decoder 4	Upsample + DSC Block	Transposed Conv (2×2), Concat with Encoder 1, DSC (3×3), Pointwise (1×1), BN, ReLU	128×128×32	256×256×16	16
Output	1×1 Conv + Sigmoid	Reduces to 1 channel for binary segmentation	256×256×16	256×256×1	2

Table 1. U-Net with DSC Model Configuration

The enhanced U-Net-integrated with DSC (shown in Figure 2) is trained and evaluated using a combination of open-source and custom-built datasets on the PARAM Siddhi-AI high-performance computing system.

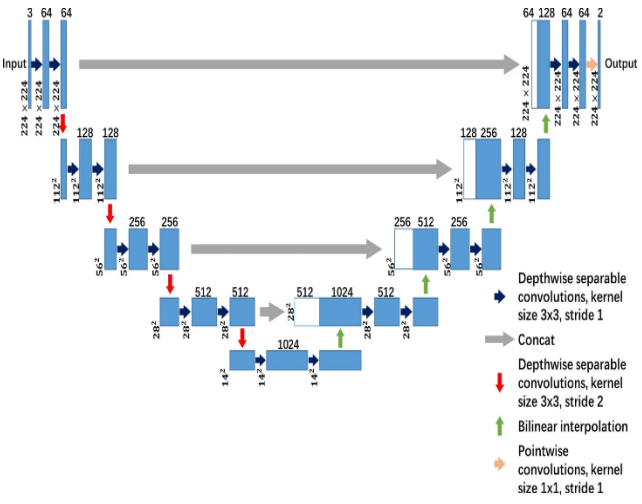

Figure 2: U-Net with DSCArchitecture (Zhang et al., 2020)

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-5-W2-2025-301-2025 | © Author(s) 2025. CC BY 4.0 License.

4.1 Evaluation Metrics

The performance of the DL models is evaluated using quantitative metrics shown in Table 2.

Quantitative Metrics	Formula
Accuracy	$\frac{TP + TN}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$
Precision	$\frac{TP}{(\mathrm{TP} + \mathrm{FP})}$
Recall	$\frac{TP}{(\mathrm{TP} + \mathrm{FN})}$
F_1-score	$\frac{2\ x\ (Precision\ x\ \ Recall)}{(Precision\ +\ \ Recall)}$
Intersection over Union (IoU)	$IoU\ =\ \frac{TP}{(TP\ + FP + FN)}$
Where TP - True Positives, FP - False Positives, TN - True Negatives and FN - False Negatives	

Table 2. Summary of the Quantitative Metrics

4.2 Training Setup

Model Configuration		
Dataset size 53877 Split		70% Train, 10% validation and 20% Test
Epochs		150
Batch Size		4
Learning Rate		1e – 4
Loss function		Binary Cross Entropy (BCE)
Activation function		Rectified Linear Unit (ReLU)
Metrics		Accuracy, IoU, Precision, Recall, F1 Score
Hardware and Software		PARAM Siddhi-AI HPC system with four NVIDIA A100 SXM4 GPUs with 40GB
U-Net-DSC	Parameters	Total: 3,576,476 Trainable: 3,564,700 Non-trainable: 11,776
	Training time	1 day, 23 hours and 31 minutes
U-Net	Parameters	Total: 31,402,497, Trainable: 31,390,721 Non-trainable: 11,776
	Training time	3 days, 23 hours and 33 minutes

Table 3. Model Configuration and Training Setup

4.3 Model Training

Both the U-Net and the U-Net with DSC models were trained for 150 epochs using the same set of hyperparameters. The Adam optimizer was chosen with a learning rate of 0.0001 and a batch size of 4 to ensure stable training. Since the task focused on binary segmentation, Binary Cross-Entropy (BCE) loss was used. ReLU activation functions were applied throughout the encoder and decoder layers to introduce non-linearity, while a sigmoid activation at the output layer generated pixel-wise probability maps. To improve training stability and speed up convergence, batch normalization was applied after each convolutional layer. A dropout rate of 10% was also added to reduce overfitting by randomly deactivating some neurons during training. Finally, early stopping was used to monitor the Intersection over Union (IoU) score, with training stopped if performance did not improve by at least 0.01 over 25 consecutive epochs. A summary of the training setup is provided in Table 3.

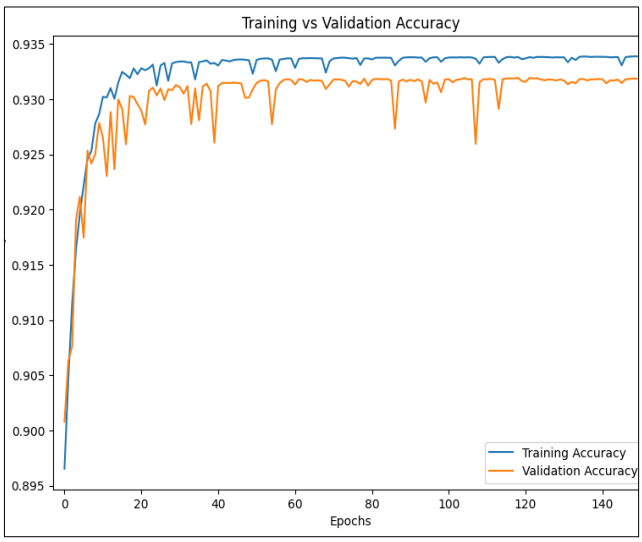

Figure 3. U-Net with DSC - Training vs. Validation Accuracy

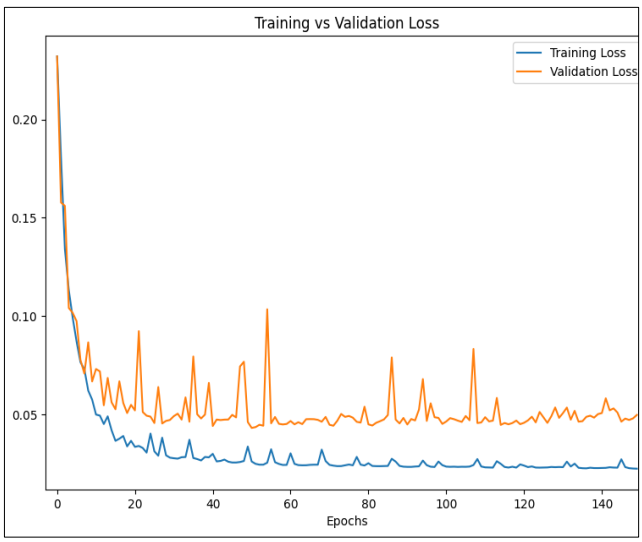

Figure 4. U-Net with DSC - Training vs. Validation Loss

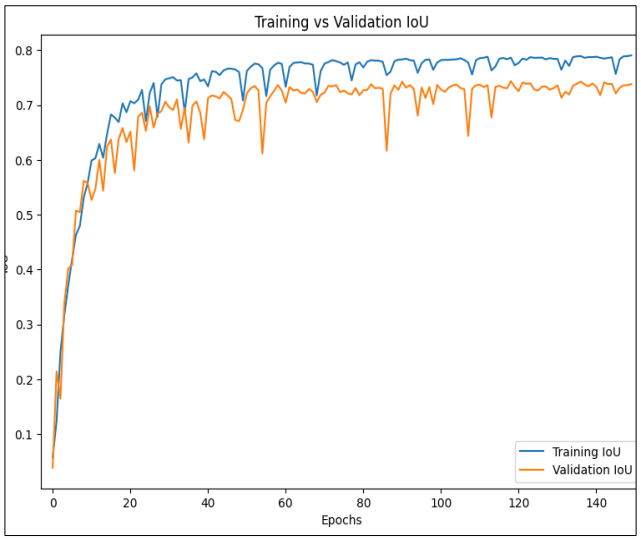

Figure 5. U-Net with DSC - Training vs. Validation IoU

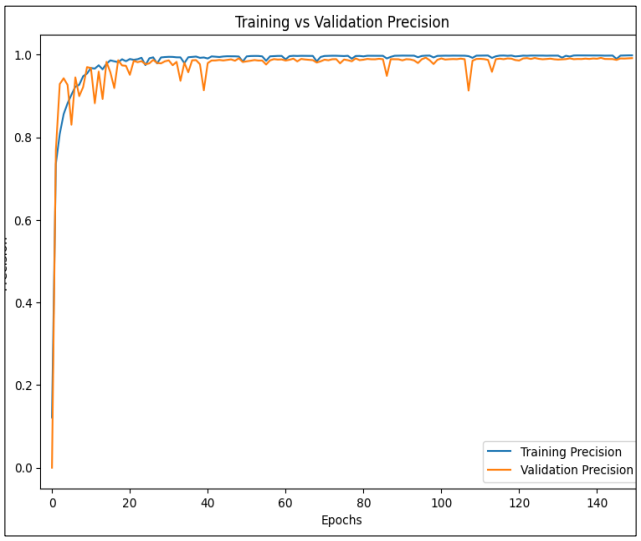

Figure 6. U-Net with DSC - Training vs. Validation Precision

Figure 7. U-Net with DSC - Training vs. Validation Recall

During training, the performance of both the standard U-Net and the U-Net with DSC was continuously monitored using metrics such as accuracy, loss, IoU, precision, recall, and F1-score. However, for clarity and conciseness, only the training and validation performance graphs of the U-Net with DSC are presented here (Figures 3 to 7). These plots show a consistent improvement across all metrics, indicating stable convergence and effective learning throughout the training process. The gradual increase in accuracy and IoU, along with the steady decline in loss, demonstrates the model's ability to accurately capture road structures. Moreover, the balanced trends in precision, recall, and F1-score highlight the robustness and reliability of the U-Net with DSC in achieving high-quality segmentation results.

5. Results And Discussion

5.1 Quantitative Results

A comparative evaluation of the employed U-Net with DSC and the standard U-Net was conducted using key performance metrics: Accuracy, Intersection over Union (IoU), Precision, Recall, and F1-Score. As shown in Figure 8. The U-Net with DSC achieved a mean accuracy of 0.956 and a mean IoU of 0.626, surpassing the standard U-Net, which recorded a mean accuracy of 0.949 and IoU of 0.581. Additionally, the U-NET with DSC exhibited higher precision (0.829 vs. 0.73) and F1-score (0.766 vs. 0.722), with a marginal decline in recall (0.733 vs. 0.761).

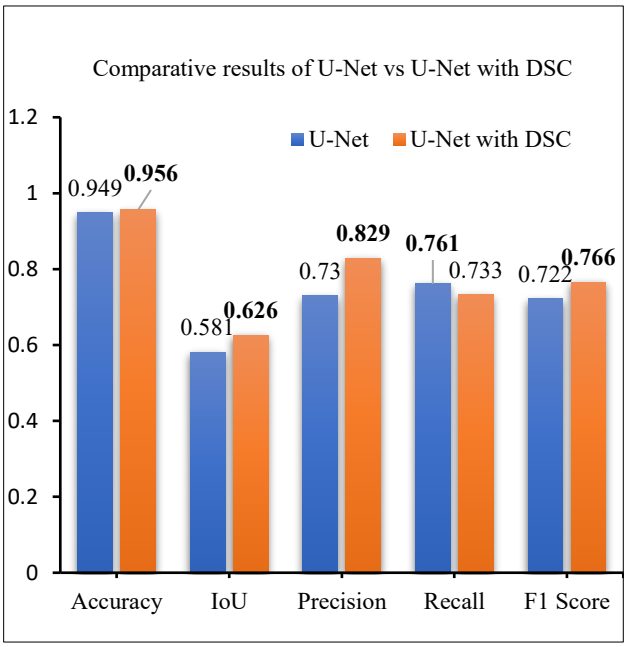

Figure 8: Comparative performance of U-Net and U-Net with DSC for road extraction.

5.2 Qualitative Results

Figures 9 and 10 represent the qualitative output results of road extraction from satellite imagery. The input satellite images (Panel A) are processed through both the standard U-Net and the used Modified U-Net with DSC. These models take the original satellite inputs and generate predicted road segmentation outputs (Panel C) through inference (Shown in Figure 9 and 10).

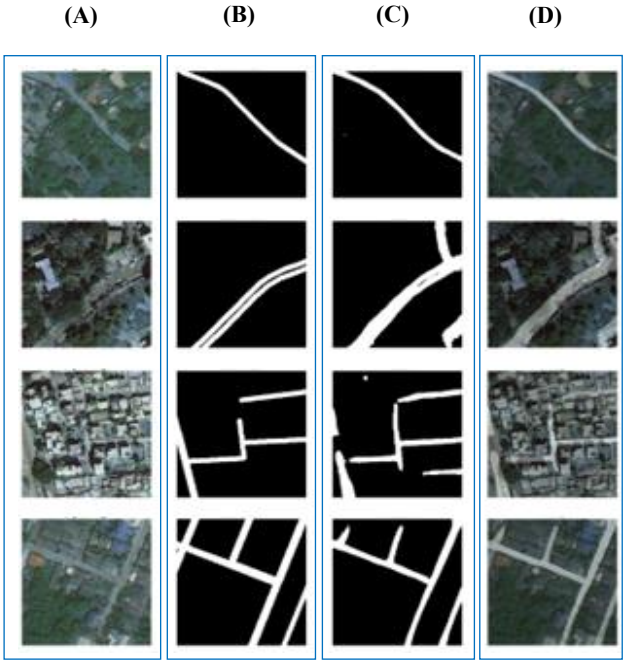

Figure 9. U-Net – Predicted Roads Output: Pannel A - Source Input Image, Pannel B - Ground Truth Image, Pannel C - Predicted Image and Pannel D - Overlay predicted on Source Input Image

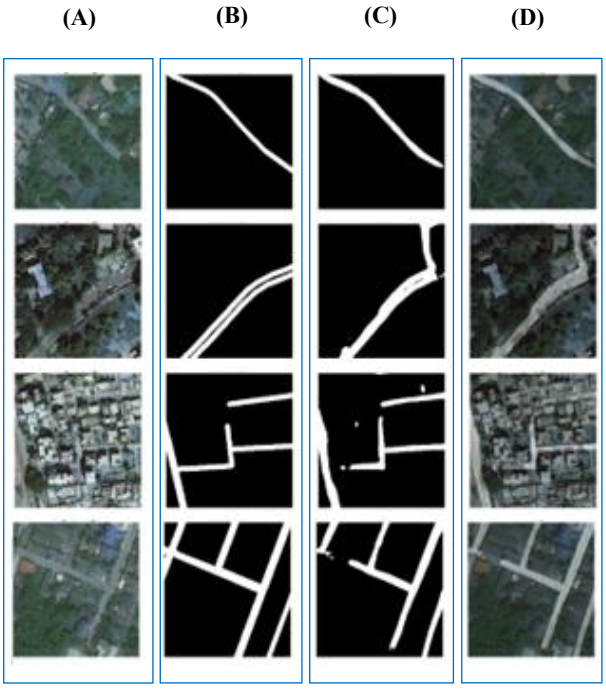

Figure 10: U-Net with DSC – Predicted Roads Output: Pannel A – Source Input Image, Pannel B - Ground Truth Image, Pannel C - Predicted Image and Pannel D - Overlay predicted on Source Input Image.

To make comparison easier, Panel B shows the actual road annotations, so you can visually check how closely the model's predictions match the real road layout. In Panel D, the predicted road mask is overlaid on the original satellite image, giving a clearer picture of how accurately the model identifies and aligns with the roads in the image. The results clearly demonstrate the effectiveness of integrating DSC into the U-Net architecture. The improved IoU indicates superior overlap between predicted and ground truth road masks, reflecting better spatial feature extraction. The substantial gain in precision suggests a reduction in false positives, which is critical for minimizing incorrect road segment predictions in satellite imagery. Although the recall for U-Net with DSC is slightly lower than that of the standard U-Net, the overall improvement in the F1-score confirms a more balanced and robust segmentation performance. These outcomes highlight the model's enhanced generalization capability in handling the complex and heterogeneous landscape characteristics typical of Indian regions.

6. Conclusion

In this study, we have used a modified version of the U-Net model with DSC on PARAM Siddhi-AI high-performance computing platform to improve the accuracy and computing time of road extraction from satellite imagery. U-Net with DSC architectural modification brought down the number of trainable parameters from 31 million in the original U-Net to just 3.5 million. As a result, the training time came down noticeable from almost 4 days to less than 2 days. Even with fewer parameters, the modified model performed better than the standard U-Net. The U-Net with DSC attained a mean accuracy of 0.956 and a mean IoU of 0.626. The predicted road networks can be integrated into platforms like the *Pradhan Mantri Gram Sadak Yojana (PMGSY)* National GIS, which can support automated and large-scale planning, monitoring, and evaluation of rural road development.

Acknowledgement

The authors express their sincere gratitude to Dr. Siva Sai Krishna Tirumani, C-DAC, Pune, for his valuable inputs during the preparation of this manuscript.

References

A. Van Etten, "SpaceNet: Road network detection in satellite images," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018.*

Barzohar, Meir, and David B. Cooper. "Automatic finding of main roads in aerial images by using geometric-stochastic models and estimation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18, no. 7 (2002): 707-721.

Demir, Ilke, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. "Deepglobe 2018: A challenge to parse the earth through satellite images." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 172-181. 2018.

Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

Huang, Xin, and Liangpei Zhang. "Road centreline extraction from high-resolution imagery based on multiscale structural

features and support vector machines." *International Journal of Remote Sensing* 30, no. 8 (2009): 1977-1987.

K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks." *Communications of the ACM* 60, no. 6 (2017): 84-90.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

Mokhtarzade, Mehdi, and MJ Valadan Zoej. "Road detection from high-resolution satellite images using artificial neural networks." *International journal of applied earth observation and geoinformation* 9, no. 1 (2007): 32-40.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241. Cham: Springer international publishing, 2015.

Zhang, Jiaqiang, Xiaoyan Li, Liyuan Li, Pengcheng Sun, Xiaofeng Su, Tingliang Hu, and Fansheng Chen. "Lightweight U-Net for cloud detection of visible and thermal infrared remote sensing images." *Optical and Quantum Electronics* 52, no. 9 (2020): 397.