



A LIGHTWEIGHT MULTILEVEL FEATURE FUSION APPROACH FOR ACCURATE SEMANTIC SEGMENTATION OF REMOTE SENSING DATA

Anoop Patel^{*1}, Dr. Safvan Vahora², Dr. Kuntesh Jani³ and Dr. Narendrasinh Chauhan⁴

¹Ph.D. Scholar, Gujarat Technological University, Ahmedabad, 382424, Gujarat, India.

²Associate Professor, Government Engineering College, Modasa, 383315, Gujarat, India.

³Associate Professor, L. D. College of Engineering, Ahmedabad, 380015, Gujarat, India.

⁴Professor, A. D. Patel Institute of Technology, Anand, 388121, Gujarat, India.

¹<http://orcid.org/0009-0001-9986-7688>, ²<http://orcid.org/0000-0002-8051-8429>,

³<http://orcid.org/0000-0001-7225-1878> ⁴<http://orcid.org/0000-0001-7225-1878>

Email: *anoop.p.patel@gmail.com, safvan.vahora@gecmodasa.ac.in, kuntesh.jani@ldce.ac.in, it.nccchauhan@adit.ac.in

ARTICLE INFO

Article History

Received: February 20, 2026

Reviewed: March 20, 2026

Accepted: May 10, 2026

Published: June 30, 2026

Keywords:

Landcover mapping,
Semantic segmentation,
Remote sensing,
Urban development, Feature
fusion

ABSTRACT

Remote sensing has become essential for land cover classification, leveraging deep learning-based segmentation to improve accuracy in environmental monitoring, agriculture, and urban planning. This study introduces Disc Net, a streamlined DeepLabv3+ adaptation optimized for semantic segmentation on the OpenEarthMap dataset. The model employs a structured pipeline, integrating InceptionV3 with squeeze-excitation (SE) attention in the encoder to enhance feature extraction, while multilayer feature fusion combines information from both shallow a deep layers for richer representations. The decoder, built on DeepLabv3+ with depthwise separable convolutions, reduces computational complexity without compromising segmentation quality. Preprocessing with stratified sampling ensures balanced class representation, further refining model performance. DiscNet outperforms existing baselines, achieving an mIoU of 79.6%, significantly surpassing U-Net (52.7%), simpleCNN (51.3%), and ResUNet (42.5%). This strong performance, validated through qualitative and quantitative assessments, demonstrates DiscNet's ability to deliver high-precision segmentation while maintaining efficiency, making it a promising solution for large scale land cover mapping.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Semantic segmentation plays a critical role in land cover mapping, enabling the pixel-wise classification of high-resolution remote sensing imagery to identify distinct land cover types such as vegetation, water bodies, urban areas, and roads [1-4]. Traditional approaches such as SVM [5], relied on manual feature extraction, which was labor-intensive and lacked the ability to capture complex spatial relationships. The evolution of traditional classification techniques such as Random Forest (RF), Decision Tree (DT), Maximum Likelihood Estimation (MLE), and K-Nearest Neighbor (k-NN) has significantly impacted land object classification and land cover mapping [6], [7]. Traditional land cover segmentation techniques, such as Random Forest, Decision Trees, and Maximum Likelihood Estimation, face significant challenges, including difficulties in handling high-resolution and complex spatial data, limited ability to resolve ambiguous or overlapping boundaries, and reliance on large, balanced labeled datasets, which are often scarce and imbalanced [8]. Additionally, their performance is hindered by variability in remote sensing data due to seasonal changes, lighting conditions, and complex land cover patterns, limiting their scalability and adaptability to modern remote sensing needs [9]. The advent of deep learning, specifically convolutional neural networks (CNNs), has revolutionized semantic segmentation, leading to significant improvements in both accuracy and scalability [10-12]. Early methods, like those described by [13], used Fully Convolutional Networks (FCNs), which replaced fully connected layers with convolutional layers to handle inputs of arbitrary size and produce segmentation maps of the same size as the input images. Since then, more sophisticated architectures such as U-Net [14] and its variants have set the standard for high-resolution remote sensing image segmentation.

U-Net implementation over sentinel-2 dataset as described in [15], achieved accurate LULC maps in output. These models and others like them have significantly improved segmentation performance, particularly in land cover mapping tasks. Modification over U-Net such as U-Net++ [16] achieved 97% accuracy in land cover classification using Gaofen and Sentinel-2 dataset, while Attention Dense U-Net (ADU-Net) [17] measured mIoU of 77.66% on the GID dataset. LULC-SegNet [18], a semantic segmentation network for land use and land cover (LULC) on privately sourced LULC dataset achieved mIoU value 73.76%. are among the most widely used architectures due to their effective use of skip connections and multi-scale feature extraction. Combining the PCA, U-Net and attention mechanism, PSE-UNet model as defined by [19], reported accuracy of 88% mIoU on Salinas dataset. One more variant TransUNet [20], identified building rooftops with significant accuracy. Thus, traditional encoder-decoder model U-Net have been improved with techniques like residual connections, attention mechanisms, and nested architectures to better capture spatial and contextual information. The integration of transformers, as seen in networks like TransUNet and UNetFormer, introduces global attention mechanisms, enabling the capture of long-range dependencies crucial for fine-grained semantic segmentation in remote sensing and provide solutions across diverse applications. Advanced encoder-decoder architectures and transformers for semantic segmentation have emerged through iterative enhancements of deep learning networks. DeepLab [8], a pioneering encoder-decoder model, achieved a notable mIoU of 79.7% on the Pascal-VOC dataset. Subsequently, researchers [21-25] have extensively explored DeepLab and its variants for segmentation tasks in remote sensing imagery, demonstrating superior performance over existing traditional encoder-decoder models in various aspects. One such network CFAMNet as described in [26] achieved mIoU value 77.22% while segmenting images in Gaofen dataset. A multimodal feature fusion based encoder-decoder network MFFNet [27] achieved significant improvement in land cover classification.

Few additional networks, such as PSPNet, HRNet, RAANet, DeaNet, and MARE [28], [29], have also been introduced in this direction, showcasing significant performance improvements. Another important development in semantic segmentation has been the evolution of transformer-based architectures, which have been extensively utilized for remote sensing tasks [30]. Authors in [31], [32] explored possibility of Vision Transformer in classifying land covers in PolSAR data. The modified ViT structure based swin transformer introduced in [30] achieved better results when compared to existing transformer framework. A multi-level fusionbased transformer network SoftFormer [33] outperformed numerous state-of-the-art techniques across four distinct LULC datasets. The primary performance metric in semantic segmentation is Mean Intersection over Union (mIoU), which measures the overlap between predicted and ground truth segments [28], [29]. In addition, overall accuracy, precision, recall, and F1 score are often used to evaluate model performance [28], [29]. While accuracy is a useful metric, it can be misleading in cases where class imbalances exist, as is often the case in land cover mapping tasks. For instance, a model that performs well on dominant classes (e.g., vegetation) but poorly on minority classes (e.g., roads or water bodies) may still achieve high overall accuracy but low mIoU. This has led many researchers to prioritize mIoU as the key evaluation metric in land cover segmentation tasks [1], [34], [35]. A range of datasets has been used in land cover mapping studies. Some of the most popular datasets include ISPRS Vaihingen and ISPRS Potsdam, which provide high-resolution aerial images for urban segmentation [28]. Sentinel-1 and Sentinel-2, part of the Copernicus program, offer multispectral and radar imagery and have been widely used in land cover studies [36], [37]. SpaceNet [38], FLAIR [39] and LandCover.ai [40] have also been widely used for segmentation and classification tasks in remote sensing. Large-scale datasets like OpenEarthMap [41], DeepGlobe [42], BigEarthNet [43], offer satellite imagery with diverse geographical contexts, enabling models to generalize across different environments. Each of these datasets presents unique challenges, such as varying resolutions, class imbalances, and different environmental conditions, which test the generalization capability of segmentation models.

Despite extensive advancements in semantic segmentation for land cover classification, many studies focus on dataset-specific scenarios, models, and techniques that fail to address the challenges posed by unique conditions, such as differentiating visually similar classes in RGB satellite images. Existing approaches often rely on large datasets or complex architectures requiring high computational resources, limiting their practicality in lightweight and resource-constrained environments. These challenges, particularly data scarcity and limited computational power, emphasize the need for carefully tuned, data-efficient, and computationally efficient models tailored for specific applications. Our study distinguishes itself by proposing a tailored approach that utilizes RGB data enhanced with innovative architectural components, such as the squeeze excitation block and depthwise separable convolutions, to achieve high segmentation accuracy. By leveraging the OpenEarthMap dataset [41] and integrating a robust class-balanced sampling strategy, our work aims to improve the classification of underrepresented classes and enhance feature refinement through an encoder-decoder based mechanism. This focus on efficient, resource-friendly methods makes our research particularly relevant for applications where computational efficiency and data limitations are critical. The contributions of our research are significant in the context of land cover classification, offering a robust solution for generating precise segmentation maps from satellite images. To achieve the desired outcomes, major objectives were systematically outlined, implemented in sequence, thoroughly analyzed, and iteratively refined.

1. Through rigorous statistical analysis and visualization techniques, we establish a class-balanced strategy that ensures effective representation of all target classes during training, enabling comprehensive model learning.
2. The adoption of the DeepLabV3+ architecture, which incorporates depthwise separable convolutions in place of standard convolutions. This modification significantly enhances the model's computational efficiency, reduces trainable parameters, and decreases complexity without compromising performance.
3. The squeeze excitation block in our model plays a pivotal role in improving feature representation by adaptively recalibrating channel-wise feature responses. This mechanism enhances the model's ability to focus on relevant features, thereby improving segmentation accuracy.
4. The decoder of our proposed model plays a pivotal role in refining segmentation outputs. By leveraging contextual information from neighboring pixels, the decoder enhances the coherence and accuracy of the class map, effectively handling fine edge refinements as part of the decoding process.

5. The integration of feature fusion during the training phase played a pivotal role in enhancing model performance. By combining detailed spatial features from skip connections with global semantic information from the SE block, the model achieved a more comprehensive feature representation, resulting in significant improvements in segmentation accuracy.

Highlighting phase-wise, we begin with data sampling techniques as a critical part of the preprocessing phase, balancing training samples to improve model robustness during training. Following this, we propose a lightweight DeepLabV3+ model with an InceptionV3 backbone, integrated with a squeeze excitation attention mechanism to enhance segmentation accuracy. The OpenEarthMap dataset, known for its variety, diversity, and accurately labeled land cover types, serves as an essential resource for training and evaluating our model. The mean Intersection over Union (mIoU) metric serves as a reliable benchmark for evaluating the performance of our semantic segmentation model.

II. MATERIALS AND METHODS

II.1 OVERVIEW OF DATASET

The OpenEarthMap dataset [41] is a high-resolution satellite imagery dataset designed for land cover and land use classification, making it well suited for semantic segmentation tasks. With detailed annotations across diverse classes such as urban areas, water bodies, forests, and agriculture, it provides the rich, varied data necessary for training deep learning models. The dataset's high spatial resolution (30 cm to 50 cm ground sampling distance) allows for accurate distinction of fine-grained features, such as water bodies and roads, which are crucial for semantic segmentation in remote sensing. Its diversity of scenes, including rural and urban environments, ensures generalization across different geographies, making it an ideal choice for improving model performance in real-world applications like environmental monitoring and urban planning.

II.2 DATASET PREPARATION FOR EXPERIMENT

As a preprocessing step, we implement a stratified sampling technique that iterates over a dataset of 5,000 sample images, each with a size of 1000x1000x3. The goal of this process is to create an optimized subset of images that balances the distribution of classes within the dataset. By ensuring that each class is adequately represented, this approach helps mitigate class imbalance issues and guarantees sufficient training samples for each class. Additionally, to reduce the complexity of the training process, we introduce image patches by downsampling the original 1000x1000 images to smaller 256x256 patches. This modification decreases computational load while maintaining sufficient detail for training, enabling more efficient model learning. The resulting balanced and optimized subset of 256x256 patches is then used to train our proposed network, improving its ability to generalize and enhancing performance during training.

II.3 PROPOSED ARCHITECTURE

In this work, as per Figure 1 we propose architecture DiscNet, which combines both the encoder (InceptionV3 backbone [44]) and the decoder (DeepLabV3+ [8]) components. We specifically focus on integrating two key modifications into the DeepLabv3+ model: the incorporation of Squeeze-and-Excitation (SE) and Depthwise Separable Convolutions (DWConv). These modifications aim to enhance the model's performance in semantic segmentation tasks, leveraging the strengths of both the SE block and DWConv. The architecture is carefully designed to optimize segmentation accuracy while maintaining computational efficiency.

II.3.1 Encoder: InceptionV3 Backbone

The encoder part is the feature extraction part, which is based on the InceptionV3 model. InceptionV3 is a pre-trained model that extracts multi-scale features using its inception modules

Input Image: Let the input image be represented as

$$X_{in} \in R^{H_{in} \times W_{in} \times C_{in}}, \text{ where} \quad (1)$$

H_{in} =Height of input image, W_{in} =Width of input image, C_{in} =Number of channels in input image. InceptionV3 Backbone:

The InceptionV3 model uses ImageNet pre-trained weights for its convolutional layers. These weights are loaded from the pre-trained model, allowing the network to start with a set of feature extractors already tuned for general image understanding. For simplicity, let's define the output after passing the input through the InceptionV3 Backbone as

$$X_{enc} \in R^{H_{enc} \times W_{enc} \times C_{enc}}, \text{ where} \quad (2)$$

H_{enc} =reduced height of the feature map after encoding, W_{enc} =reduced width of the feature map after encoding, C_{enc} = the depth of the feature map, which corresponds to the number of filters in the final layers.

Squeeze-Excitation (SE) Block: The Squeeze-Excitation (SE) block recalibrates channel-wise feature responses by applying attention mechanisms [45]. Global Average Pooling (GAP): Apply Global Average Pooling to the encoder output X_{enc} to capture the global context of each channel.

$$X_{gap} = GAP(X_{enc}) \quad (3)$$

Fully Connected Layers: Apply two fully connected layers with ReLU and Sigmoid activations to generate the channel-wise attention. This produces a scaling factor for each channel.

$$SE(X_{enc}) = \sigma(W_2 \cdot ReLU(W_1 \cdot X_{gap})) \quad (4)$$

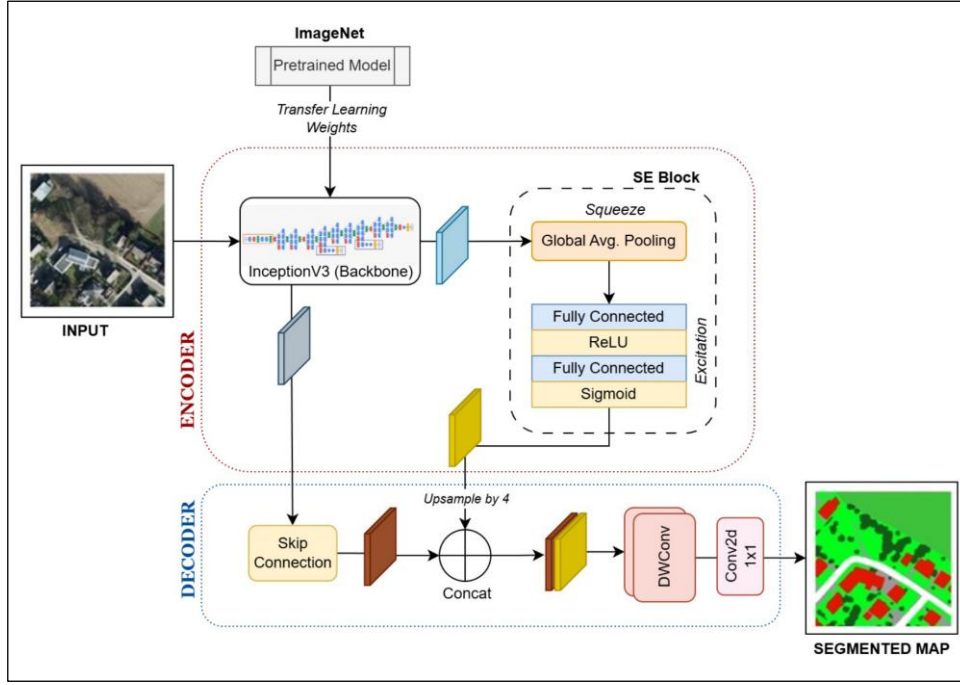


Figure 1: Architecture of proposed model (DiscNet).
Source: Authors, (2026).

Where $W_1 \in R^{\frac{C_{enc} \times C_{enc}}{r}}$ is the weight matrix of the first fully connected layer (with r as the reduction ratio), $W_2 \in R^{\frac{C_{enc} \times C_{enc}}{r}}$ is the weight matrix of the second fully connected layer, σ is the sigmoid activation function.

Recalibration: The SE block generates attention weights $A_{enc} \in R^{1 \times 1 \times C_{enc}}$, which are used to recalibrate the feature map X_{enc} by element-wise multiplication.

$$X_{enc_att} = X_{enc} \times A_{enc}, \text{ where } X_{enc_att} \in R^{H_{enc} \times W_{enc} \times C_{enc}} \quad (5)$$

II.3.2 Decoder: DeeplabV3+

The decoder part upsamples the feature maps and merges the features from both the encoder and the SE block to produce the output. The decoder is designed to reconstruct the segmentation map or output image. Up-sampling:

The feature maps are upsampled (typically using bilinear interpolation or transposed convolutions) to the spatial resolution of the input image.

$$X_{dec_upsampled} = \text{Upsampling}(X_{enc_att}) \quad (6)$$

Skip Connections: The skip connection allows the decoder to use both high-level and low-level features by concatenating the upsampled feature maps with the skip connection from earlier layers of the encoder ("mixed3" in InceptionV3). Let the skip connection feature map be

$$X_{skip} \in R^{H_{enc} \times W_{enc} \times C_{skip}} \quad (7)$$

Feature Fusion: The skip connection is concatenated with the upsampled output. Concatenating the outputs of the skip connection and the SE block enables the model to combine detailed spatial features from shallow layers with global semantic information from deeper layers. This process facilitates feature fusion, where both local and global contexts are integrated, yielding a more comprehensive and robust feature representation for precise and accurate segmentation.

$$X_{concat} = \text{Concatenate}(X_{dec_upsampled}, X_{skip}) \quad (8)$$

Decoder Convolutions: After concatenation, apply convolutional layers (typically using depthwise separable convolutions) to refine the feature maps.

$$X_{dec_conv} = \text{DWConv}(X_{concat}) \quad (9)$$

Final Output Layer: The final layer consists of a 1x1 convolution that outputs class scores for each pixel in the segmentation task. Let the output image be

$$X_{out} \in R^{H_{out} \times W_{out} \times C_{out}}, \text{ where } H_{out} = H_{in}, W_{out} = W_{in}, C_{out} = \text{Number of classes} \quad (10)$$

$$X_{out} = Conv2D(X_{dec_{conv}}, filters = C_{out}, kernel = (1,1)) \quad (11)$$

The modified DeepLabv3+ model, with the InceptionV3 backbone, introduces two key optimizations:

1. Depthwise Separable Convolutions (DWConv): These convolutions are applied in both the encoder and decoder to reduce computational complexity and improve processing efficiency.
2. Squeeze-and-Excitation (SE): Following each convolutional block, SE is used to recalibrate the feature map channels, allowing the model to focus more effectively on the most important features.

These optimizations ensure that both feature extraction and segmentation are carried out more efficiently, with a significant reduction in computational cost due to the use of depthwise separable convolutions. Meanwhile, the SE mechanism improves the model's capacity to adaptively emphasize the most relevant features, enhancing overall segmentation performance. Additionally, the feature fusion through concatenation of skip connection and SE block outputs ensures the integration of local spatial details with global semantic context, resulting in a more enriched and robust feature representation for segmentation.

III. RESULTS AND DISCUSSIONS

III.1. EXPERIMENTAL SETUP AND MODEL HYPERPARAMETERS

III.1.1 Hardware and Software Requirements

The experiments were conducted on the OpenEarthMap dataset, which includes high-resolution RGB images with corresponding pixel-wise annotated ground truth maps, using a system running Ubuntu 20.04. The hardware configuration included an Intel Xeon CPU, an NVIDIA Tesla T4 GPU with 16GB VRAM, and 32GB of RAM, ensuring efficient deep learning model training and inference. For software, Keras 3.0 was used as the primary deep learning framework, with TensorFlow 2.16.1 serving as the backend. The implementation was carried out in Python 3.7, utilizing various additional libraries such as NumPy, Matplotlib, and Pandas for numerical computations, data visualization, and data manipulation, respectively. This environment provided a robust foundation for developing and evaluating the model effectively.

III.1.2 Hyperparameters and Training Settings

All subset images were resized to a uniform resolution of 256×256 pixels. The models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 16. The categorical cross-entropy loss function was used to handle multi-class segmentation tasks. The training process was conducted for 50 epochs with early stopping based on training loss. Stratified sampling ensured balanced representation of all classes during training. The models were evaluated using Mean Intersection over Union (mIoU) metric. The proposed model was compared with U-Net, ResUNet, and a simple CNN (simple and custom made CNN architecture) to benchmark performance. All baseline models were implemented with identical preprocessing steps and trained under the same conditions for fair comparison.

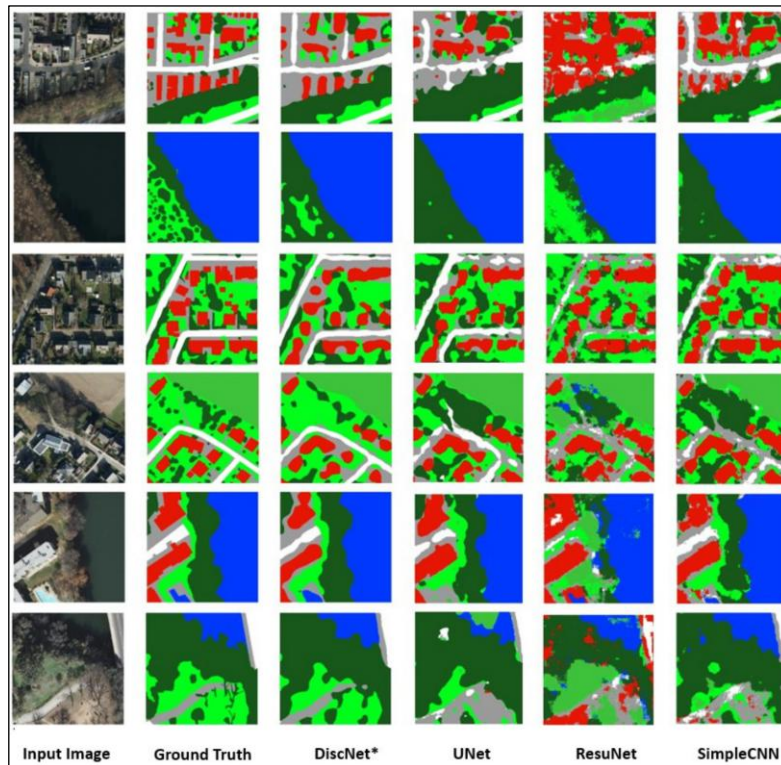


Figure 2: Final segmented map with various models.

Source: Authors, (2026).

III.2 VISUAL REPRESENTATION AND MODEL PERFORMANCE

The qualitative assessment of segmentation outputs reveals notable differences in the performance of the four models. As per Figure 2 DiscNet, the proposed model, demonstrates the highest alignment with ground truth, producing visually refined segmentation maps with well-defined boundaries and accurate class delineation, even in regions with overlapping spectral properties such as roads, water bodies, and bareland.

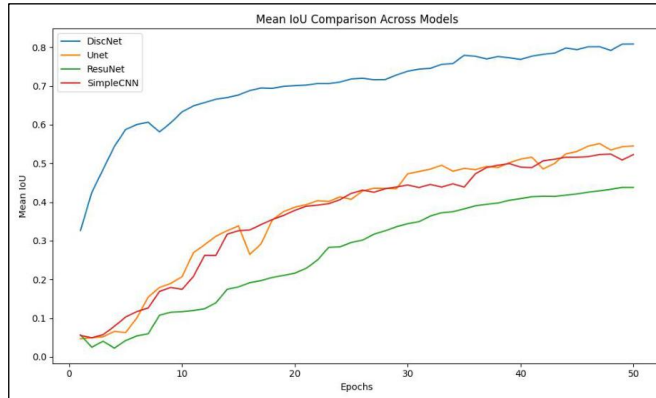


Figure 3: Avg. MioU vs Epochs.
Source: Authors, (2026).

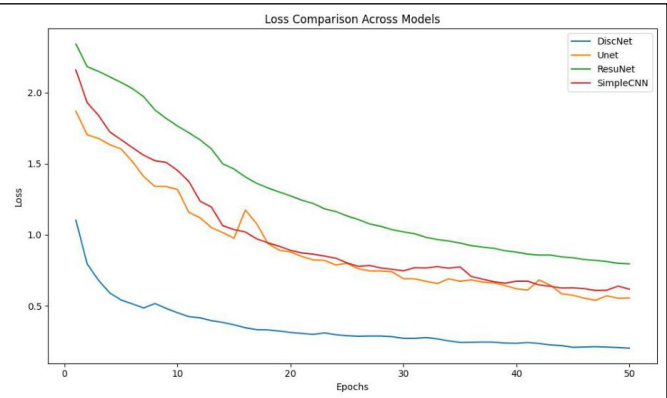


Figure 4: Avg. Loss vs Epochs.
Source: Authors, (2026).

The set of segmented images produced by DiscNet closely approximates the ground truth, demonstrating model's effectiveness in accurate land cover classification. U-Net delivers reasonably accurate maps but struggles with finer class boundaries, leading to minor misclassifications. The Simple CNN provides moderate performance, with evident class merging and a lack of detail in areas requiring complex segmentation. In contrast, ResUNet, despite its complex architecture featuring residual connections, produces poor segmentation quality characterized by frequent misclassifications and blurred boundaries, indicating inefficiencies in adapting to the specific challenges posed by this dataset.

III.3 QUANTITATIVE ANALYSIS OF MIOU AND LOSS

The quantitative evaluation as per values considered in plotting Figure 3 and 4 corroborates the visual findings, with DiscNet achieving the highest average mIoU of 0.7947 and the lowest average loss of 0.2183, reflecting its superior learning efficiency and ability to generalize across all 8 classes. U-Net ranks second with an average mIoU of 0.5275 and an average loss of 0.5876, showing reasonable performance but limited refinement in complex regions. The Simple CNN, with an average mIoU of 0.5134 and an average loss of 0.6310, achieves moderate results but is constrained by its simplistic architecture. Despite its sophisticated residual connections, ResUNet performs the worst, with the lowest average mIoU of 0.4245 and the highest average loss of 0.8318, likely due to poor optimization and unstable convergence in handling the dataset's intricate spatial and spectral complexities.

III.4 COMPARATIVE ANALYSIS OF AVERAGE CLASS-WISE MIOU

The class-wise evaluation of segmentation performance as per Table 1 highlights the superior generalization ability of DiscNet across all eight land cover classes. DiscNet consistently achieves the highest mean Intersection over Union (mIoU) values, particularly excelling in water bodies, agricultural land, and buildings, where precise boundaries are crucial. U-Net follows as a distant second, showing moderate performance with better results in developed space and building segmentation but struggles in classes with higher spectral overlap. Simple CNN outperforms ResUNet in several classes, demonstrating its moderate capability despite its simpler architecture. ResUNet, despite its architectural complexity, underperforms across all classes, indicating its inefficiency in adapting to the spectral and spatial intricacies of the dataset. Overall, DiscNet's superior segmentation accuracy reinforces its robustness and adaptability for multi-class land cover segmentation tasks.

Table 1: Average Class-wise mIoU for Different Models.

	Bareland	Rangeland	Developed Space	Road	Tree	Water	Agri Land	Building
DiscNet	0.5991	0.6187	0.6967	0.6949	0.6589	0.8747	0.8885	0.865
Unet	0.3142	0.3826	0.5334	0.5141	0.5948	0.6751	0.5835	0.7777
ResuNet	0.2122	0.2523	0.3245	0.2601	0.3752	0.5531	0.356	0.5702
SimpleCNN	0.2856	0.3972	0.4969	0.449	0.6312	0.7775	0.6525	0.736

Source: Authors, (2026).

IV. CONCLUSIONS

This study presents DiscNet, a novel lightweight deep learning model for semantic segmentation of high-resolution RGB satellite images. Our experiments on the OpenEarthMap dataset demonstrate the effectiveness of DiscNet in segmenting diverse land cover classes with high accuracy.

Through a detailed comparative analysis with established models, including U-Net, ResUNet, and a simple CNN, we show that DiscNet outperforms these models in terms of both visual representation and quantitative metrics, particularly in challenging regions such as water bodies, vegetation, and agricultural land. The results of the evaluation, based on metrics such as Mean Intersection over Union (mIoU) and loss, clearly indicate DiscNet's superior performance across all classes. The inclusion of advanced architectural features, such as squeeze-excitation attention mechanisms and depthwise separable convolutions, contributes to enhanced segmentation accuracy. These findings suggest that DiscNet is a promising model for large-scale semantic segmentation tasks, offering a balance between computational efficiency and segmentation performance. Future work will focus on extending the model to include multilevel data and feature fusion techniques, integrating additional bands to input and or information rich feature set that could improve segmentation task.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Anoop Patel, Safvan Vahora.

Methodology: Anoop Patel, Safvan Vahora.

Investigation: Anoop Patel, Safvan Vahora.

Discussion of results: Anoop Patel, Safvan Vahora.

Writing – Original Draft: Anoop Patel, Safvan Vahora.

Writing – Review and Editing: Anoop Patel, Safvan Vahora, Kuntesh Jani, Narendrasinh Chauhan.

Approval of the final text: Anoop Patel, Safvan Vahora, Kuntesh Jani, Narendrasinh Chauhan.

VI. ACKNOWLEDGMENTS

Data acknowledgement and code availability

This study utilizes the OpenEarthMap dataset [41], a publicly available benchmark dataset for semantic segmentation. Upon acceptance of this manuscript, we will publicly release the source code, including preprocessing, data preparation, pre-trained models, and training scripts in a dedicated GitHub repository. This will ensure transparency and facilitate reproducibility for future research.

VII. REFERENCES

- [1] A. Vali, S. Comai, M. Matteucci, Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review, *Remote Sensing* 12 (15) (2020) 2495. doi:10.3390/rs12152495.
- [2] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, L. Jiao, Building extraction of aerial images by a global and multi-scale encoder-decoder network, *Remote Sensing* 12 (15) (2020) 2350. doi:10.3390/rs12152350.
- [3] P. Pourmohammadi, D. A. Adjero, M. P. Strager, Y. Farid, Pre-dicting developed land expansion using deep convolutional neural networks, *Environmental Modelling & Software* 134 (2020) 104751. doi:10.1016/j.envsoft.2020.104751.
- [4] A. Di Pilato, N. Taggio, A. Pompili, M. Iacobellis, A. Di Flo-rio, D. Passarelli, S. Samarelli, Deep learning approaches to earth observation change detection, *Remote Sensing* 13 (20) (2021) 4083. doi:10.3390/rs13204083.
- [5] G. Mutreja, S. Kumar, D. Jha, A. Singh, R. Singh, Identifying settle-ments using svm and u-net, in: *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2020, pp. 1217–1220. doi:10.1109/IGARSS39084.2020.9324040.
- [6] L. S. Macarringue, E. L. Bolfe, P. R. M. Pereira, Developments in land use and land cover classification techniques in remote sensing: A re-view, *Journal of Geographic Information System* 14 (1) (2022) 1–28. doi:10.4236/jgis.2022.141001.
- [7] L. Qu, Z. Chen, M. Li, J. Zhi, H. Wang, Accuracy improve-ments to pixel-based and object-based lulc classification with auxiliary datasets from google earth engine, *Remote Sensing* 13 (3) (2021) 453. doi:10.3390/rs13030453.
- [8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4) (2018) 834–848. doi:10.1109/TPAMI.2017.2699184.
- [9] W. Boonpook, Y. Tan, A. Nardkulpat, K. Torsri, P. Torteeka, P. Kam-sing, U. Sawangwit, J. Pena, M. Jainaan, Deep learning semantic segmentation for land use and land cover types using landsat 8 im-agery, *ISPRS International Journal of Geo-Information* 12 (1) (2023) 14. doi:10.3390/ijgi12010014.
- [10] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, *IEEE geoscience and remote sensing magazine* 5 (4) (2017) 8–36. doi:10.1109/MGRS.2017.2762307.
- [11] W. Sun, R. Wang, Fully convolutional networks for semantic segmen-tation of very high resolution remotely sensed images combined with dsm, *IEEE Geoscience and Remote Sensing Letters* 15 (3) (2018) 474–478. doi:10.1109/LGRS.2018.2802944.
- [12] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in: *International conference on learning representations*, 2015. doi:arXiv:1412.7062.
- [13] J. Song, S. Gao, Y. Zhu, C. Ma, A survey of remote sensing im-age classification based on cnns, *Big earth data* 3 (3) (2019) 232–254. doi:10.1080/20964471.2019.1657720.
- [14] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional net-works for biomedical image segmentation, in: *Medical Image Comput-ing and Computer-Assisted Intervention MICCAI*, 2015, pp. 234–241. doi:10.1007/978-3-319-24574-428.
- [15] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, S. P. Brumby, Global land use/land cover with sentinel 2 and deep learning, in: *2021 IEEE international geoscience and re-mote sensing symposium IGARSS*, IEEE, 2021, pp. 4704–4707. doi:10.1109/IGARSS47720.2021.9554653.

- [16] L. Wang, J. Wang, Z. Liu, J. Zhu, F. Qin, Evaluation of a deep-learning model for multispectral remote sensing of land use and crop classification, *The Crop Journal* 10 (5) (2022) 1435–1451. doi:10.1016/j.cj.2022.01.009.
- [17] M. Talha, F. A. Bhatti, S. Ghuffar, H. Zafar, Adu-net: semantic seg-mentation of satellite imagery for land cover classification, *Advances in Space Research* 72 (5) (2023) 1780–1788. doi:10.1016/j.asr.2023.06.015.
- [18] Z. Shi, J. Fan, Y. Du, Y. Zhou, Y. Zhang, Lulc-segnet: Enhanc-ing land use and land cover semantic segmentation with denois-ing diffusion feature fusion, *Remote Sensing* 16 (23) (2024) 4573. doi:10.3390/rs16234573.
- [19] J. Li, H. Wang, A. Zhang, Y. Liu, Semantic segmentation of hyperspec-tral remote sensing images based on pse-unet model, *Sensors* 22 (24) (2022) 9678. doi:10.3390/s22249678.
- [20] Z. Chen, D. Li, W. Fan, H. Guan, C. Wang, J. Li, Self-attention in reconstruction bias u-net for semantic segmentation of building rooftops in optical remote sensing images, *Remote sensing* 13 (13) (2021) 2524. doi:10.3390/rs13132524.
- [21] S. He, B. Zhou, B. Wang, Z. Ding, J. Zhang, Land classi-fication research based on the improved deeplabv3+ model and the ai earth remote sensing cloud platform, *IEEE Access* (2024). doi:10.1109/ACCESS.2024.1234567.
- [22] Z. Qian, Y. Cao, Z. Shi, L. Qiu, C. Shi, A semantic segmen-tation method for remote sensing images based on deeplab v3, in: 2021 2nd International Conference on Big Data & Artificial Intelli-gence & Software Engineering (ICBASE), IEEE, 2021, pp. 396–400. doi:10.1109/ICBASE53648.2021.00088.
- [23] K. Zheng, H. Wang, F. Qin, C. Miao, Z. Han, An improved land use classification method based on deeplab v3+ under gau-gan data enhancement, *IEEE Journal of Selected Topics in Ap-plied Earth Observations and Remote Sensing* 16 (2023) 5526–5537. doi:10.1109/JSTARS.2023.3281234.
- [24] K. Qiu, L. E. Budde, D. Bulatov, D. Iwaszczuk, Exploring fusion techniques in u-net and deeplab v3 architectures for multi-modal land cover classification, in: *Earth Resources and Environmental Remote Sensing/GIS Applications XIII*, Vol. 12268, SPIE, 2022, pp. 190–200. doi:10.1117/12.2630010.
- [25] R. T. H. Herlawati, Land cover segmentation of multispectral images using u-net and deeplabv3+ architecture, *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)* 17 (1) (2024). doi:10.21609/jiki.v17i1.1234.
- [26] Z. Wang, J. Wang, K. Yang, L. Wang, F. Su, X. Chen, Semantic segmen-tation of high-resolution remote sensing images based on a class feature attention mechanism fused with deeplabv3+, *Computers & Geosciences* 158 (2022) 104969. doi:10.1016/j.cageo.2021.104969.
- [27] Y. Wang, W. Zhang, W. Chen, C. Chen, Z. Liang, Mffnet: Multi-modal feature fusion network for synthetic aperture radar and optical image land cover classification, *Remote Sensing* 16 (13) (2024) 2459. doi:10.3390/rs16132459.
- [28] S. Ajibola, P. Cabral, A systematic literature review and bibliometric analysis of semantic segmentation models in land cover mapping, *Remote Sensing* 16 (12) (2024) 2222. doi:https://doi.org/10.3390/rs16122222.
- [29] A. Shreehari, P. R. Chandan, P. V. Vinod, R. Hebbar, Land use land cover classification using deep learning techniques: A com-parative study, in: 2024 11th International Conference on Sig-nal Processing and Integrated Networks (SPIN), 2024, pp. 7–11. doi:10.1109/SPIN60856.2024.10511460.
- [30] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, Transformers in vision: A survey, *ACM computing surveys (CSUR)* 54 (10s) (2022) 1–41. doi:10.1145/3505244.
- [31] H. Dong, L. Zhang, B. Zou, Exploring vision transformers for polari-metric sar image classification, *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021) 1–15. doi:10.1109/TGRS.2021.3137383.
- [32] H. Wang, C. Xing, J. Yin, J. Yang, Land cover classification for polari-metric sar images based on vision transformer, *Remote Sensing* 14 (18) (2022) 4656. doi:10.3390/rs14184656.
- [33] R. Liu, J. Ling, H. Zhang, Softformer: Sar-optical fusion trans-former for urban land use and land cover classification, *ISPRS Jour-nal of Photogrammetry and Remote Sensing* 218 (2024) 277–293. doi:10.1016/j.isprsjprs.2024.01.015.
- [34] X. Yuan, J. Shi, L. Gu, A review of deep learning methods for semantic segmentation of remote sensing imagery, *Expert Systems with Applica-tions* 169 (2021) 114417. doi:10.1016/j.eswa.2020.114417.
- [35] J. Li, Y. Cai, Q. Li, M. Kou, T. Zhang, A review of remote sensing image segmentation by deep learning methods, *International Journal of Digital Earth* 17 (1) (2024) 2328827. doi:10.1080/17538947.2024.2328827.
- [36] J. V. Sol´orzano, J. F. Mas, Y. Gao, J. A. Gallardo-Cruz, Land use land cover classification with u-net: Advantages of combining sentinel-1 and sentinel-2 imagery, *Remote Sensing* 13 (18) (2021) 3600. doi:10.3390/rs13183600.
- [37] T. Zhang, J. Su, Z. Xu, Y. Luo, J. Li, Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier, *Applied Sciences* 11 (2) (2021) 543. doi:10.3390/app11020543.
- [38] A. Van Etten, D. Lindenbaum, T. M. Bacastow, Spacenet: A remote sensing dataset and challenge series, *arXiv preprint arXiv:1807.01232* (2018). doi:10.48550/arXiv.1807.01232.
- [39] A. Garioud, N. Gonthier, L. Landrieu, A. De Wit, M. Valette, M. Poup´ee, S. Giordano, et al., Flair: a country-scale land cover semantic segmentation dataset from multi-source optical im-agery, *Advances in Neural Information Processing Systems* 36 (2024). doi:10.48550/arXiv.2306.15494.
- [40] A. Boguszewski, D. Batorski, N. Ziemia-Jankowska, T. Dziedzic, A. Zambrzycka, Landcover. ai: Dataset for automatic mapping of build-ings, woodlands, water and roads from aerial imagery, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog-nition*, 2021, pp. 1102–1110. doi:10.1109/CVPRW53098.2021.00121.
- [41] J. Xia, N. Yokoya, B. Adriano, C. Broni-Bediako, Open-EarthMap: A benchmark dataset for global high-resolution land cover mapping, in: *Proceedings of the IEEE/CVF Winter Con-ference on Applications of Computer Vision*, 2023, pp. 6254–6264. doi:10.1109/WACV56688.2023.00630.

- [42] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, R. 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, 2018, pp. 172–181. doi:10.1109/CVPRW.2018.00030.
- [43] G. Sumbul, M. Charfuelan, B. Demir, V. Markl, Bigearthnet: A large-scale benchmark archive for remote sensing image under-standing, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2019, pp. 5901–5904. doi:10.1109/IGARSS.2019.8898854.
- [44] M. A. Morid, A. Borjali, G. Del Fiol, A scoping review of transfer learning research on medical image analysis using ima-genet, Computers in biology and medicine 128 (2021) 104115. doi:10.1016/j.combiomed.2020.104115.
- [45] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceed-ings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141. doi:10.1109/CVPR.2018.00745.