SkyNetAI: Deep Learning for Multi-Source Semantic Segmentation of High-Resolution Aerial Imagery

Hardik Soni 20CS30023 Gmail GROUP 1 Rahul Mandal 20CS30039 Gmail

Abstract

Land-use classification and semantic segmentation of high-resolution aerial imagery are crucial for various applications like urban planning, environmental monitoring, and resource management. This study investigates the potential of deep learning models for semantic segmentation of aerial imagery, leveraging the rich information from a variety of geospatial datasets. Building upon existing research in deep learning for land cover classification, we explore the integration of contextual information and task-specific attention mechanisms to improve segmentation accuracy. This research aims to contribute to the advancement of Spatio-temporal Data Mining and Knowledge Discovery by extracting valuable land-use insights from high-resolution aerial imagery.

I. INTRODUCTION

The advancements in remote sensing technology have enabled the capture of high-resolution aerial images, offering detailed insights for land cover and land use (LCLU) classification. These images, with ground sampling distances as small as 5 to 10 cm, allow for the identification of fine details in both urban and natural environments. However, traditional image analysis methods, such as sliding windows and candidate regions, face limitations due to inefficiencies and high computational demands. To overcome these challenges, deep learning (DL) methods, particularly convolutional neural networks (CNNs), have gained traction in semantic segmentation tasks, significantly improving both the accuracy and speed of image analysis. Among these, the U-Net model is especially effective, with its encoder-decoder structure and skip connections helping to preserve critical spatial information.

Building upon the U-Net foundation, recent innovations have integrated self-attention mechanisms and separable convolutions to further enhance segmentation accuracy and efficiency. Self-attention enables the model to focus on key parts of the image while reducing the impact of irrelevant data, and separable convolutions optimize computational speed. These advancements allow for more precise classification of land features such as buildings, roads, and water bodies, supporting essential applications in urban planning, environmental monitoring, and resource management. This study demonstrates how these deep learning techniques are shaping the future of aerial imagery analysis, offering practical solutions to contemporary challenges in urban development and environmental sustainability.

II. OBJECTIVES

Here are few research objectives that we wish to achieve during this project:

1) **Develop a Hybrid Deep Learning Architecture for Semantic Segmentation:** The goal is to design a novel deep learning model, integrating U-Net-based architectures with self-attention mechanisms and separable convolutions. This model will aim to enhance the accuracy and efficiency of segmenting high-resolution aerial imagery, focusing on multi-class land use and land cover (LULC) classification.
2) **Incorporate Multi-Source Imagery for Enhanced Land Feature Detection:** To leverage various remote sensing data sources, such as multispectral, LiDAR, and SAR imagery, in a deep learning framework to improve the precision and robustness of land feature classification. The objective is to analyze how multi-source data fusion enhances segmentation accuracy and mitigates challenges posed by high-resolution data.
3) **Evaluate the Impact of Attention Mechanisms on Segmentation Accuracy:** Investigate how attention mechanisms, including self-attention and channel/spatial attention, contribute to refining deep learning models for more effective segmentation of aerial images. This objective will compare attention-based models with traditional convolutional methods in terms of accuracy, efficiency, and computational load.

4) **Test and Validate the Model in Real-World Applications:** Apply the proposed model to practical applications such as urban planning, disaster management, and environmental monitoring, using real-world high-resolution aerial datasets. The objective is to assess the practical impact of advanced deep learning methods in enhancing decision-making processes related to land use and environmental sustainability.
5) **Optimize Computational Efficiency for Large-Scale Aerial Image Processing:** Focus on optimizing the proposed deep learning model to process large-scale, high-resolution aerial imagery efficiently. This objective aims to address the computational challenges associated with high-resolution data by exploring innovations like separable convolutions and patch-wise image processing.

III. Literature Review

Recent advancements in deep learning for remote sensing and semantic segmentation of high-resolution aerial imagery have focused on improving accuracy, efficiency, and the integration of multi-source data. In the work by Latif et al. (2023), a hybrid approach combining CNNs and attention mechanisms is presented, which significantly enhances the segmentation of complex land cover structures by focusing on key areas while filtering irrelevant data. Similarly, Deng et al. (2019) explore the application of deep neural networks, emphasizing the importance of high-resolution data from sources like Sentinel-2 for land-use classification, while proposing a novel U-Net-based model. These approaches highlight the ability of advanced neural architectures to leverage high-resolution data, achieving unprecedented levels of accuracy in aerial image segmentation.

Attention mechanisms, such as those explored in the study by Lyu et al. (2020) , further enhance deep learning models by improving focus on relevant image regions, thus increasing segmentation precision without the need for manual feature engineering. Moreover, innovations like self-supervised learning in remote sensing, as discussed by Khan et al. (2023), provide new methods for reducing the need for large labeled datasets, further advancing the application of deep learning in multi-source imagery analysis. The integration of these methodologies into high-resolution remote sensing tasks provides a foundation for more accurate and computationally efficient semantic segmentation of aerial imagery, significantly contributing to practical applications in urban planning, agriculture, and environmental monitoring.

IV. Dataset Description

The **LandCover.ai** dataset represents a valuable resource for high-resolution and detailed land cover analysis utilizing deep learning methods and remote sensing data. Designed to facilitate semantic segmentation tasks for both environmental and administrative applications, this dataset focuses on capturing rural land cover categories through annotated aerial imagery. Specifically targeting rural regions in Poland, the dataset encompasses diverse environmental conditions, including seasonal variations, lighting changes, and weather effects, thus enhancing its applicability across a wide range of use cases.

A. Significance and Applications

The dataset is particularly relevant for critical applications in:

- **Natural Resource Management:** Mapping forests, water bodies, and agricultural areas for sustainable resource utilization.
- **Urban Planning:** Identifying and monitoring built-up areas to support population distribution studies and infrastructure planning.
- **Environmental Protection:** Monitoring ecological changes and addressing environmental concerns through accurate land cover mapping.

B. Dataset Composition and Characteristics

The LandCover.ai dataset spans a total area of **216.27 square kilometers** across rural Poland. The aerial imagery was acquired at two resolutions: **25 cm per pixel** and **50 cm per pixel**. This high spatial resolution enables fine-grained semantic segmentation, which is challenging to achieve using lower-resolution satellite imagery.

C. Folder Structure

The dataset is organized into two primary folders:

1) `images`: Contains the aerial images of the dataset.
2) `masks`: Contains the corresponding annotated masks for each image.

D. Land Cover Classes

The dataset is annotated into the following primary classes, each representing a distinct type of land cover:

- **Buildings:** Includes structures such as houses and buildings with visible roofs and walls. These annotations are crucial for urban planning and population distribution studies.
- **Woodlands:** Represents areas densely covered by trees, excluding isolated trees and orchards. This class is vital for forest health monitoring and biodiversity analysis.
- **Water:** Includes both flowing and stagnant water bodies, such as ponds and lakes, but excludes minor channels. These annotations support water resource management initiatives.
- **Roads:** Encompasses roadways and railways, with larger roads annotated as polygons and narrower paths as polylines. These are important for transportation planning and infrastructure development.

This figure here shows some images from the dataset:

E. Annotation Process

The annotations were created manually using the **VGG Image Annotator (VIA)**. Each image underwent a meticulous two-stage verification process to ensure the accuracy and reliability of the labels. This manual approach addresses the limitations of automatic labeling techniques, enhancing the dataset's suitability for high-stakes environmental and administrative applications.

F. Class Distribution

The class distribution of the dataset, including their respective coverage areas and labels, is presented in Table 1.

Class	Coverage (km^2)	Label
Background	125.75	0
Building	1.85	1
Woodland	72.02	2
Water	13.15	3
Road	3.50	4

Fig. 1. Class coverage and corresponding labels in the LandCover.ai dataset.

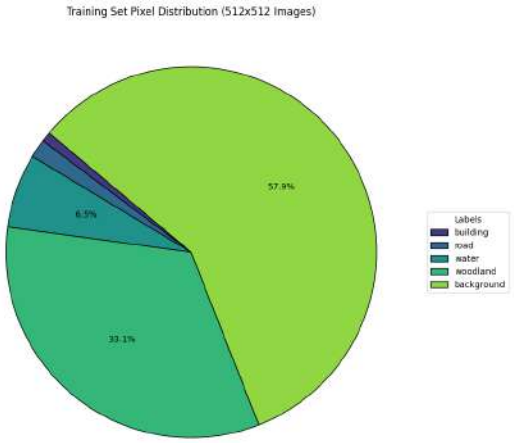


Fig. 2. Pie chart representation of class coverage.

G. Visualization

To facilitate dataset exploration, a utility functions named `visualize_tif`, `visualize_dataset`, `split_images` was created, containing all necessary. The custom function `visualize_tif` enables visualization of RGB images along with their corresponding masks. For consistency, the same color scheme used by the dataset creators is adopted:

- **Dark Gray:** Represents buildings.
- **Middle Gray:** Represents woodlands.
- **Light Gray:** Represents water.
- **White:** Represents roads.
- **Black:** Represents the background (other land uses).

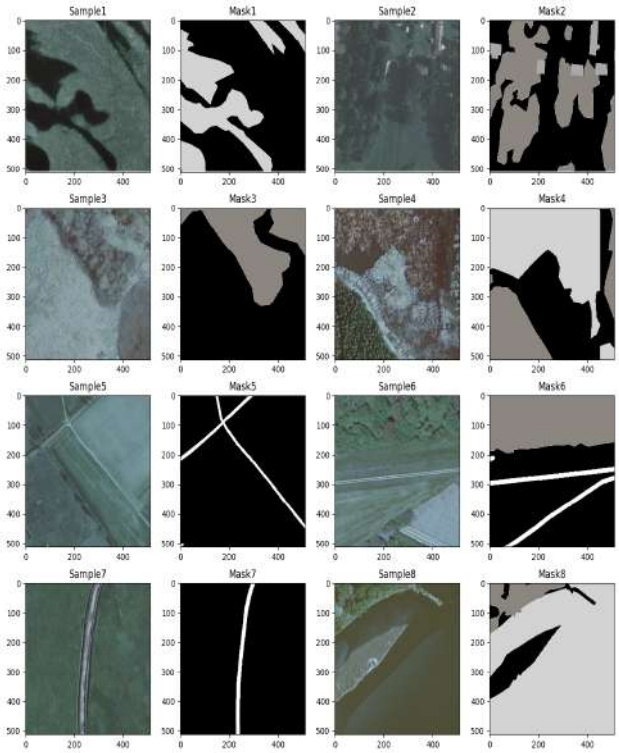
This visualization helps to accurately assess the dataset's quality and provides insights into the distribution of different land use classes.

H. Summary

The LandCover.ai dataset is a robust tool for advancing research in semantic segmentation and land cover analysis. Its high-resolution imagery, meticulous annotations, and detailed class representations make it an ideal choice for addressing challenges in environmental monitoring, urban planning, and natural resource management.

1) Data Visualization: Effective data visualization is crucial for understanding the characteristics and quality of the LandCover.ai dataset. Initial visualization functions are employed to display high-resolution aerial images alongside their corresponding segmentation masks. This process involves showcasing sample images to verify the alignment and accuracy of the masks with the actual land cover elements. By visualizing both the RGB images and the labeled masks, researchers can assess the diversity and complexity of the dataset, ensuring that the labels accurately represent the various land cover classes. These visualizations also aid in identifying any potential issues with the data, such as misalignments or inconsistencies, before proceeding to model training.

2) Image Cropping: Given the large size of the original aerial images, efficient handling and processing require splitting these images into smaller, manageable patches. The image cropping process involves dividing each high-resolution image and its corresponding mask into square tiles of a standardized size (e.g., 512x512 pixels). This approach not only facilitates easier storage and faster processing but also ensures uniform input dimensions for deep learning models. Each cropped image and mask pair is saved in an organized output directory, maintaining a clear correspondence between images and their annotations. This systematic cropping enhances the dataset's usability,

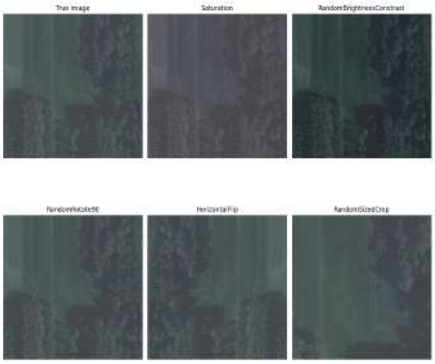


allowing for efficient batch processing during training and enabling models to focus on localized land cover features without the overhead of handling excessively large images.

3) Data Augmentation: To improve the robustness and generalization capabilities of segmentation models, various data augmentation techniques are applied to the dataset. Data augmentation introduces variability in the training data, helping models learn to recognize land cover elements under different conditions and transformations. The augmentation strategies include:

- **Color Adjustments**: Modifying hue, saturation, and brightness to simulate different lighting conditions.
- **Geometric Transformations**: Applying random rotations, flips, and crops to mimic different orientations and perspectives.
- **Random Cropping**: Extracting random patches from images to increase the diversity of training samples.

These augmentations are implemented using the Albumentations library, ensuring that both images and their corresponding masks undergo identical transformations to maintain label integrity. By augmenting the dataset, the models are exposed to a wider range of scenarios, enhancing their ability to accurately segment land cover classes in real-world applications where conditions may vary significantly.



4) Applications and Model Performance: The LandCover.ai dataset addresses gaps in publicly available high-resolution aerial datasets, particularly in Europe. The **benchmark performance** of segmentation models on this dataset has achieved **mean Intersection over Union (IoU) scores of 85.56%** and above, underscoring its quality for supervised deep learning tasks. Researchers and practitioners can use LandCover.ai for training models that support **automatic land cover mapping**, which is fundamental to numerous fields:

- **Agriculture**: Mapping and monitoring agricultural fields to enhance crop management and detect land-use changes over time.
- **Urban and Regional Planning**: Facilitating the development of rural areas by accurately mapping human-made structures and infrastructure.
- **Forestry and Conservation**: Assessing forest cover, health, and changes, especially important in the context of climate change and biodiversity conservation.
- **Hydrology**: Understanding the distribution of water resources and identifying changes in water bodies, which are critical for water management policies.

5) Comparison to Other Datasets: The LandCover.ai dataset is particularly notable when compared to similar datasets, such as the **Inria Aerial Image Labeling Dataset** (focused on urban areas) and the **Massachusetts Roads Dataset** (limited to roads in Massachusetts). Unlike these datasets, LandCover.ai offers **balanced representation across four key classes in rural settings** and features manual annotations that improve accuracy. Most competing datasets focus on specific land cover types or are limited to lower-resolution imagery, whereas LandCover.ai provides comprehensive rural coverage with high resolution and fine-grained annotations.

In conclusion, the LandCover.ai dataset fills a unique niche within remote sensing and geospatial analysis. By providing an openly accessible, high-resolution dataset with multi-class annotations in a rural European context, it supports the development of accurate and scalable models for a variety of land cover mapping tasks essential to both scientific research and public policy.

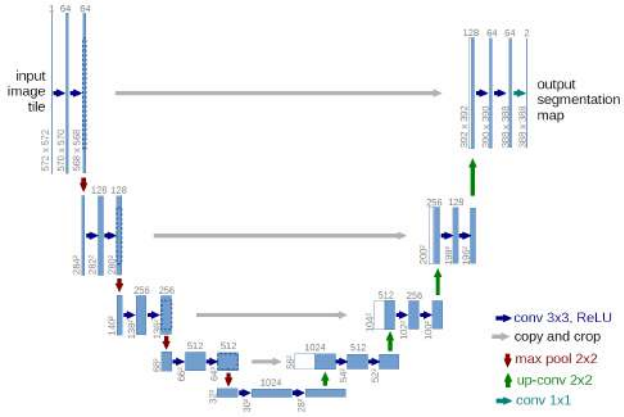
V. Experiments

This research focuses on developing and evaluating deep learning architectures—U-Net, U-Net with ResNet encoder, and ***DeepLabV3+*** with ***ResNet*** encoder—for multi-source semantic segmentation of high-resolution aerial imagery. The primary objective is to classify **land use and land cover (LULC)** features, including buildings, roads, water bodies, vegetation, and other natural structures. The study leverages high-resolution datasets such as Sentinel-2 imagery and publicly available aerial datasets like LandCover.ai, pre-processed to ensure compatibility for training and evaluation.

A. Baseline U-Net Model

The baseline U-Net model served as the foundational architecture for the segmentation experiments. U-Net employs a symmetric encoder-decoder design with skip connections, which are instrumental in retaining spatial

information essential for high-resolution image segmentation. The encoder progressively reduces the spatial dimensions, while the decoder reconstructs the segmented image through upsampling, incorporating information from corresponding encoder layers via skip connections. U-Net architecture as proposed in [15]:-



The implemented U-Net architecture includes the following refinements:

- **Architecture:** Each encoder block comprises two convolutional layers followed by batch normalization and ReLU activation. Max pooling is used for downsampling, while the decoder uses transposed convolutional layers for upsampling and concatenation from encoder outputs.
- **Loss Function and Optimization:** The model employs the Jaccard loss function (IoU) to address class imbalance, alongside the Adam optimizer for efficient convergence.

B. U-Net with ResNet Encoder

The U-Net architecture was enhanced by replacing the encoder with a ResNet-50 backbone pre-trained on ImageNet. The integration of ResNet's residual connections enables deeper feature extraction, significantly improving the ability to segment complex and intricate structures.

- **Architecture:** The ResNet backbone extracts hierarchical features, which are passed to a symmetric decoder similar to the baseline U-Net.
- **Advantages:** Residual learning enhances the model's capacity to detect fine-grained patterns, such as road networks and water boundaries, in high-resolution imagery.

C. DeepLabV3+ with ResNet Encoder

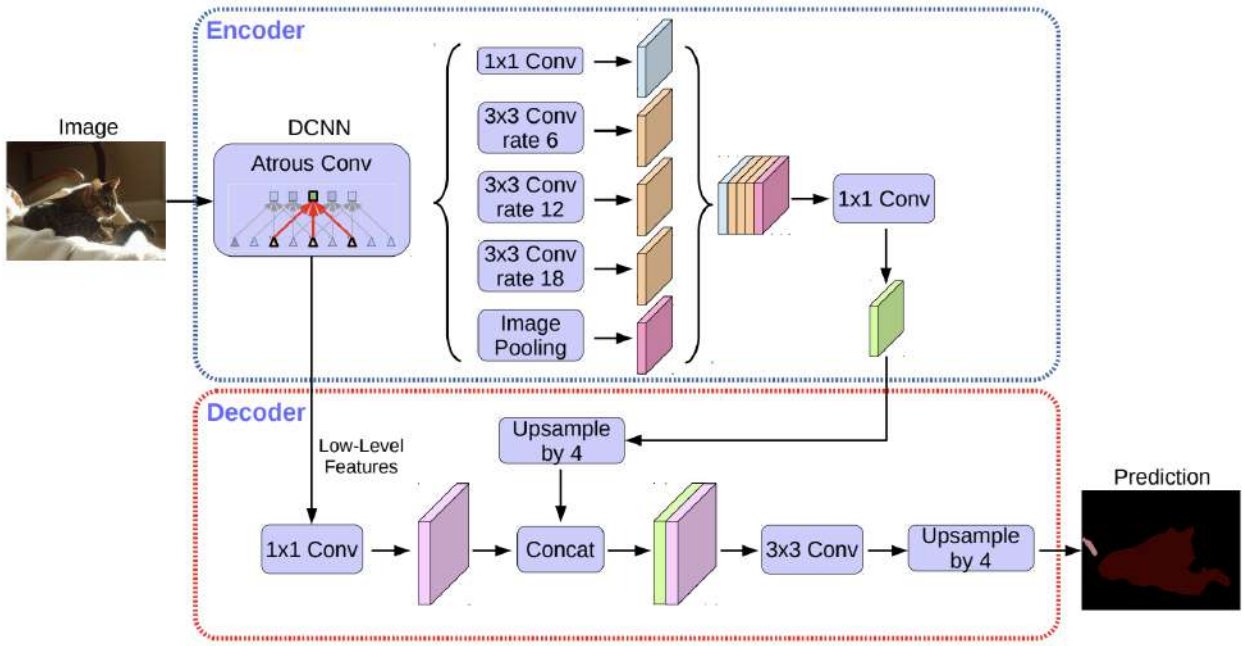
DeepLabV3+, a state-of-the-art model for semantic segmentation, was utilized for further experimentation. Its encoder incorporates ResNet-50, while the decoder employs **Atrous Spatial Pyramid Pooling (ASPP)** to capture multi-scale contextual information critical for effective segmentation. DeepLabV3+ architecture with ResNet encoder:-

- **Architecture:** The encoder uses atrous convolutions to capture contextual information at multiple scales, while the decoder refines spatial details for high-resolution segmentation.
- **Benefits:** DeepLabV3+ excels in handling both large-scale features, such as vegetation, and fine-grained elements, such as building edges.

D. Training and Hyper-Parameter Tuning

All models underwent rigorous training using a consistent pipeline to ensure comparability. Hyper-parameters were optimized through random search.

- **Training Settings:** Models were trained for 30 epochs with a batch size of 8. The learning rate was initialized to 5×10^{-5}, and early stopping was applied after three consecutive epochs without improvement in validation performance.
- **Metrics:** Performance was evaluated using Intersection-over-Union (IoU), precision, recall, and F1-score. Pixel-wise accuracy for each class was also reported.



E. Post-Processing

Post-processing techniques, such as Conditional Random Fields (CRFs), were applied to refine segmentation outputs, particularly at object boundaries.

- **Qualitative Evaluation:** Visual inspections of segmentation outputs were conducted to assess alignment with ground truth.
- **Quantitative Evaluation:** Metrics such as IoU and overall accuracy were computed on the test dataset to assess model robustness.

F. Data Augmentation

To enhance model generalization, the following data augmentation techniques were employed:

- **Color Adjustments:** Hue-saturation-value shifts and brightness-contrast normalization.
- **Geometric Transformations:** Random rotations, horizontal flips, and random cropping.

These augmentations improved the models' robustness to variations in aerial imagery conditions.

G. Summary of Results

- The baseline U-Net performed well for simpler classes, such as water bodies, but struggled with complex boundaries.
- U-Net with ResNet demonstrated superior feature extraction capabilities, excelling in fine-grained segmentation tasks.
- DeepLabV3+ achieved the highest IoU and F1-scores, effectively capturing both large-scale and intricate features.

The experiments underscore the critical role of advanced architectures and pre-processing strategies in achieving high-quality semantic segmentation for geospatial applications.

VI. Results and Observations

Results for Vanilla U-Net Model

Below are the training and validation loss plots, as well as the precision, recall, F1-score, and class probabilities obtained during the evaluation of the U-Net model. These results provide insights into the performance of the U-Net architecture for semantic segmentation of high-resolution aerial imagery.

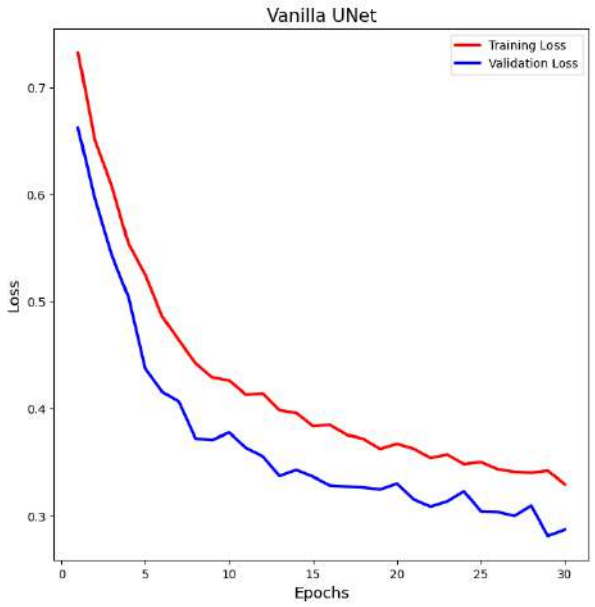


Fig. 3. Training and Validation Loss Plots for U-Net Model

Class	Precision	Recall	F1-score	Support
Background	0.95	0.93	0.94	239926712
Building	0.85	0.79	0.82	3956554
Woodland	0.90	0.95	0.92	144170112
Water	0.93	0.87	0.90	24388122
Road	0.76	0.61	0.68	7513188

Fig. 4. Class-wise Metrics

Discussion of Results: The U-Net model demonstrates a high overall accuracy of **92.39%** across all classes, highlighting its effectiveness for high-resolution segmentation tasks. The **Mean Intersection over Union (IoU)** value of **0.7510** further validates its robustness in handling both dominant and minority classes.

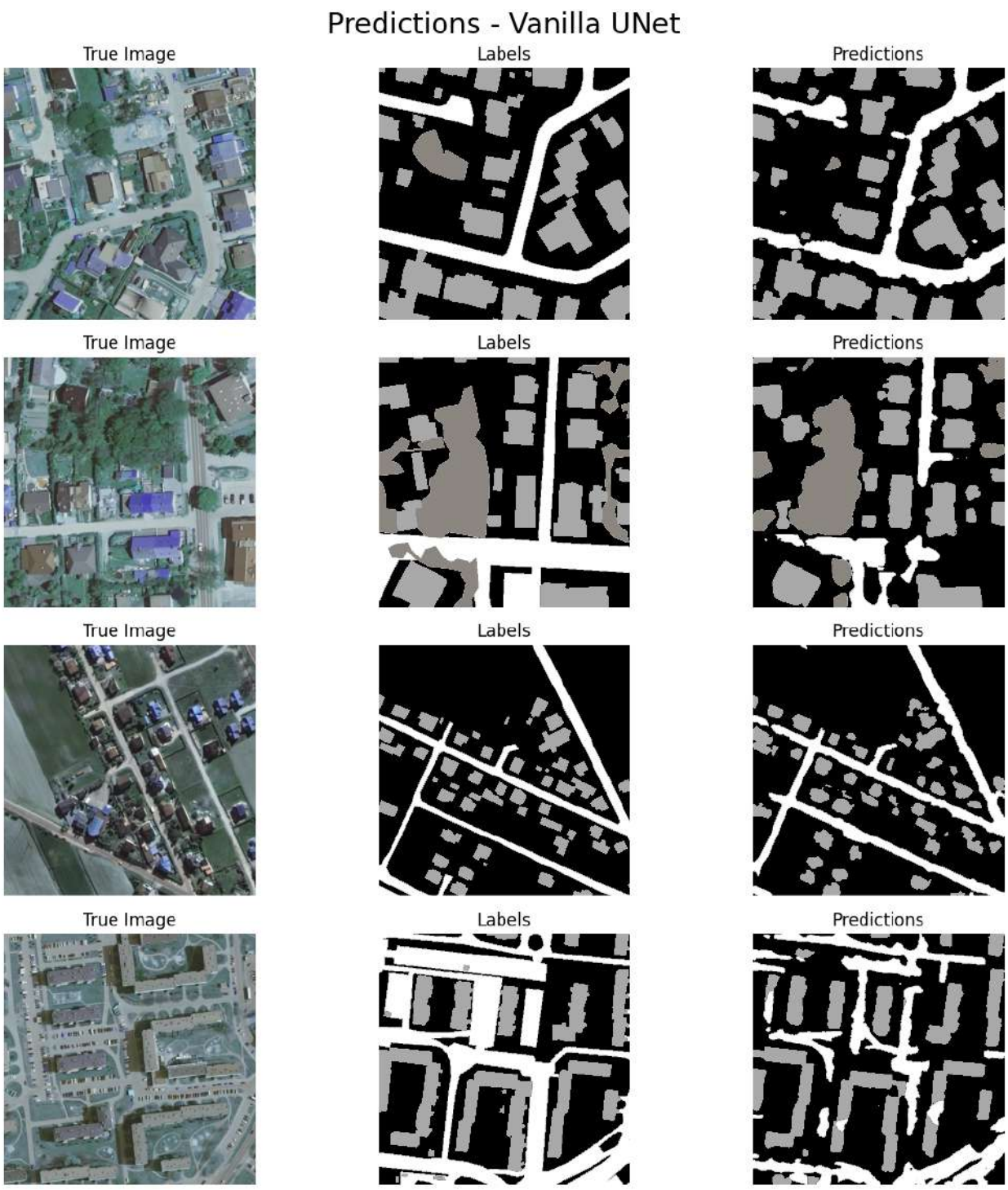
- **Background:** Achieves the highest performance with a precision of **95%**, recall of **93%**, and an F1-score of **94%**. This is expected as the background typically represents the majority class in the dataset.
- **Building:** Precision and recall values of **85%** and **79%**, respectively, indicate some challenges in capturing fine details or small structures, leading to an F1-score of **82%**.
- **Woodland:** A strong F1-score of **92%**, supported by high precision (**90%**) and recall (**95%**), highlights the model's capability to accurately segment vegetation areas.
- **Water:** With an F1-score of **90%**, the model effectively distinguishes water bodies but occasionally confuses boundaries, as shown by the slightly lower recall (**87%**).
- **Road:** Performance is lower for roads with a precision of **76%**, recall of **61%**, and an F1-score of **68%**. This suggests difficulty in segmenting narrow, elongated structures in high-resolution imagery.

Class Probabilities: The predicted class probabilities further validate the model's consistency:

Metric	Value
Total Accuracy	0.9239
Mean IoU	0.7510

Class Probabilities	
Background	0.998
Building	0.992
Woodland	0.998
Water	0.998
Road	0.990

These values indicate that the model is highly confident in predicting dominant classes like background and woodland, while slightly less confident for minority or challenging classes like roads and buildings.



Summary: The U-Net model performs well in segmenting high-resolution aerial imagery, particularly for dominant classes such as background and woodland. However, further refinements, such as architectural enhancements or post-processing techniques, could improve segmentation of challenging classes like roads and buildings.

U-Net with Resnet50 Encoder

Below are the training and validation loss plots, as well as the precision, recall, F1-score, and class probabilities obtained during the evaluation of the ResNet-U-Net model. These results provide insights into the performance of the ResNet-U-Net architecture for semantic segmentation of high-resolution aerial imagery.

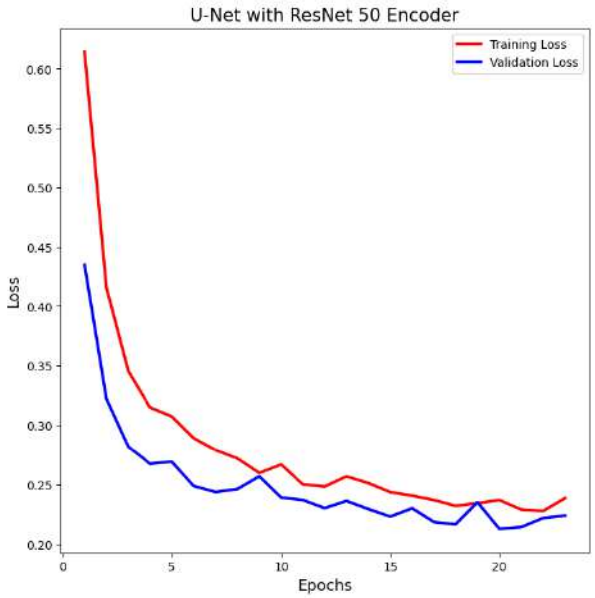


Fig. 5. Training and Validation Loss Plots for ResNet-U-Net Model

Class	Precision	Recall	F1-score	Support
Background	0.96	0.95	0.96	239926712
Building	0.90	0.83	0.86	3956554
Woodland	0.93	0.96	0.94	144170112
Water	0.97	0.95	0.96	24388122
Road	0.83	0.70	0.76	7513188

Fig. 6. Class-wise Metrics for ResNet-U-Net Model

Discussion of Results: The ResNet-U-Net model demonstrates a high overall accuracy of **94.78%** across all classes, indicating significant improvements over the Vanilla U-Net model. The **Mean Intersection over Union (IoU)** value of **0.8205** highlights its enhanced robustness, especially for handling complex and minority classes.

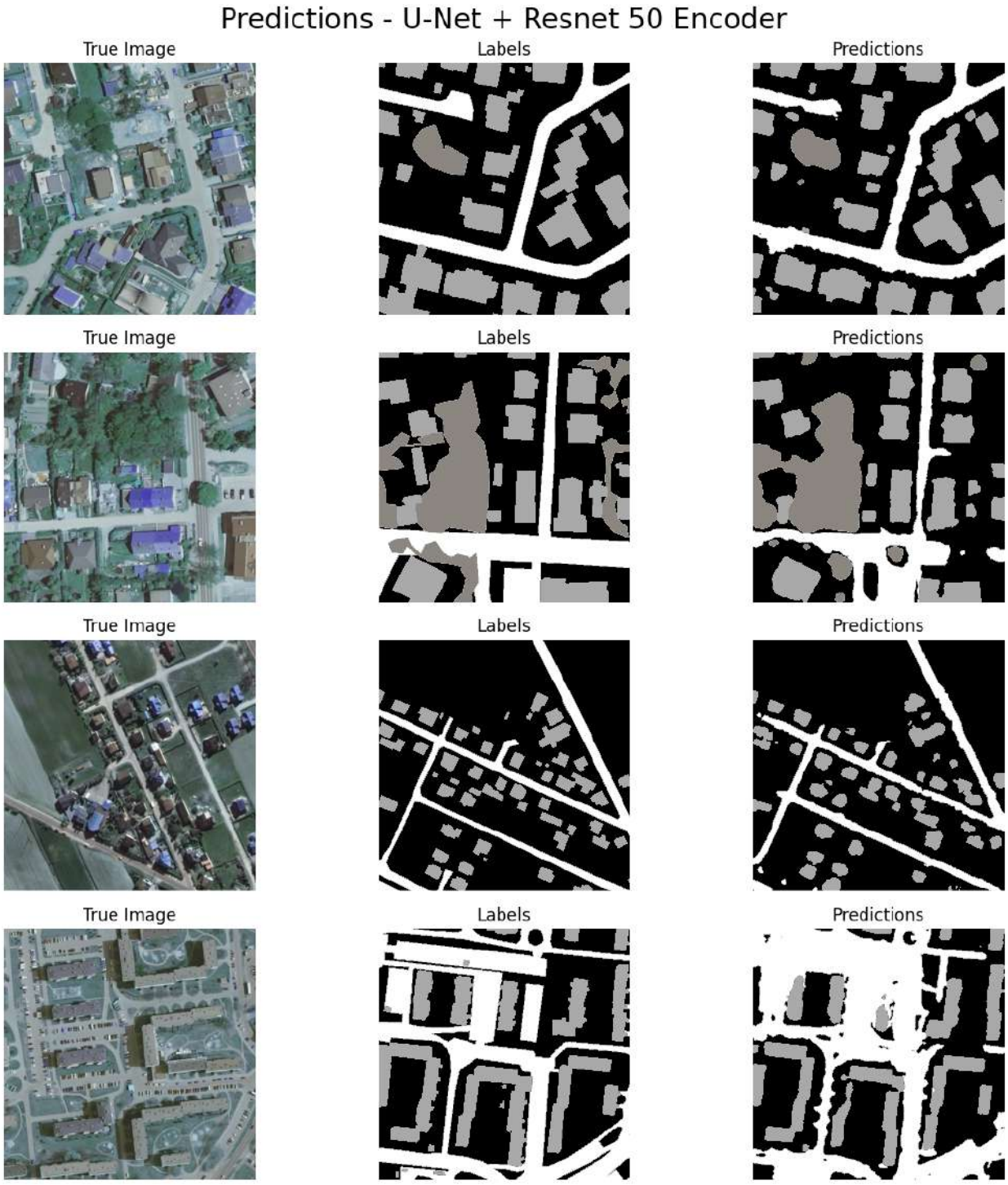
- **Background:** Exhibits exceptional performance with a precision of **96%**, recall of **95%**, and an F1-score of **96%**, showcasing its ability to effectively handle the dominant class.
- **Building:** Achieves a precision of **90%** and recall of **83%**, resulting in an F1-score of **86%**. While improved from the Vanilla U-Net, some challenges persist in capturing fine details.
- **Woodland:** With a precision of **93%**, recall of **96%**, and an F1-score of **94%**, the model effectively segments vegetation areas with high confidence.
- **Water:** Achieves a near-perfect performance with an F1-score of **96%**, supported by precision and recall values of **97%** and **95%**, respectively.
- **Road:** While showing improvement over the Vanilla U-Net, the precision of **83%**, recall of **70%**, and F1-score of **76%** highlight the complexity of segmenting narrow and elongated structures.

Class Probabilities: The predicted class probabilities confirm the model's strong confidence:

Metric	Value
Total Accuracy	0.9478
Mean IoU	0.8205

Class Probabilities	
Background	0.999
Building	0.994
Woodland	0.999
Water	0.999
Road	0.991

These probabilities suggest a high level of confidence in segmenting dominant classes like background and woodland while maintaining reasonable confidence for challenging classes such as roads and buildings.



Summary: The U-Net ResNet-50 model achieves superior performance in segmenting high-resolution aerial imagery compared to the Vanilla U-Net. With a higher Mean IoU and total accuracy, it effectively handles both dominant and minority classes. However, further architectural optimizations or post-processing techniques may be required to improve the segmentation of roads and small structures.

DeepLabV3+ with ResNet50 Encoder

Below are the training and validation loss plots, as well as the precision, recall, F1-score, and class probabilities obtained during the evaluation of the DeepLabV3+ model. These results provide insights into the performance of the DeepLabV3+ architecture with a ResNet50 encoder for semantic segmentation of high-resolution aerial imagery.

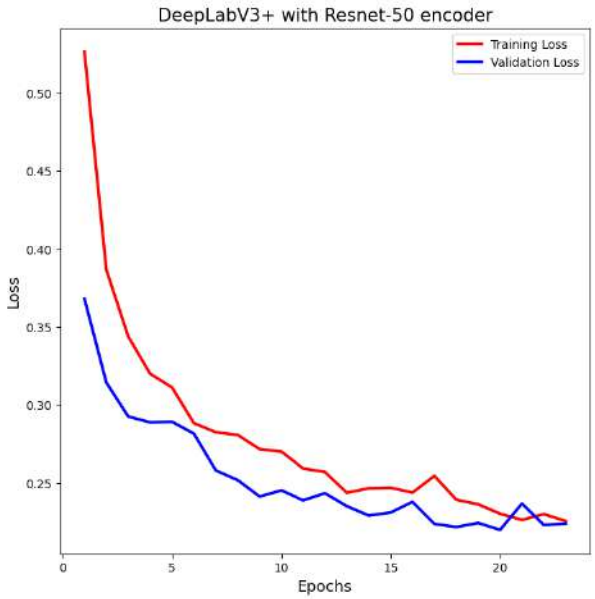


Fig. 7. Training and Validation Loss Plots for DeepLabV3+ with ResNet50 Encoder

Class	Precision	Recall	F1-score	Support
Background	0.95	0.96	0.96	239926712
Building	0.88	0.82	0.85	3956554
Woodland	0.94	0.94	0.94	144170112
Water	0.97	0.95	0.96	24388122
Road	0.81	0.72	0.76	7513188

Fig. 8. Class-wise Metrics for DeepLabV3+ with ResNet50 Encoder

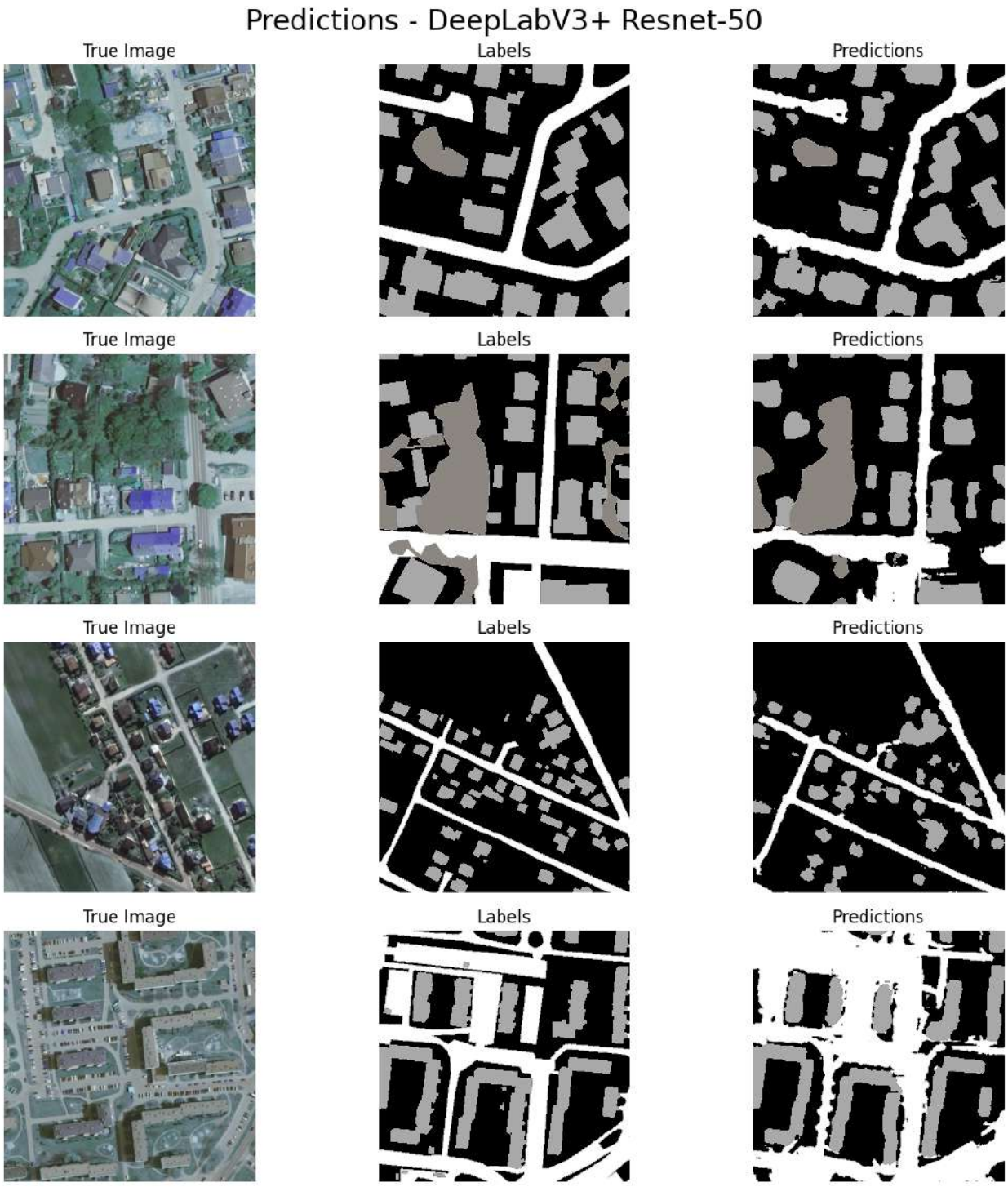
Discussion of Results: The DeepLabV3+ with ResNet50 encoder model demonstrates a high overall accuracy of **94.59%** across all classes, showing solid performance for high-resolution segmentation tasks. The **Mean Intersection over Union (IoU)** value of **0.8155** further indicates its robustness in handling both dominant and minority classes.

- **Background:** The model achieves a precision of **95%**, recall of **96%**, and F1-score of **96%**, indicating high accuracy in segmenting background areas.
- **Building:** Precision and recall values of **88%** and **82%**, respectively, result in an F1-score of **85%**, showing some challenges in capturing fine structures but performing better than the Vanilla U-Net.
- **Woodland:** With a precision of **94%**, recall of **94%**, and an F1-score of **94%**, the model effectively handles vegetation areas with high accuracy.
- **Water:** Precision and recall values of **97%** and **95%**, respectively, lead to an F1-score of **96%**, demonstrating the model's strong performance in identifying water bodies.
- **Road:** Despite an improvement over the previous models, the precision of **81%**, recall of **72%**, and F1-score of **76%** indicate challenges in segmenting narrow and elongated structures such as roads.

Class Probabilities: The predicted class probabilities indicate the model's high confidence in its predictions:

Metric	Value
Total Accuracy	0.9459
Mean IoU	0.8155

Class Probabilities	
Background	0.996
Building	0.982
Woodland	0.995
Water	0.995
Road	0.978



Summary: The DeepLabV3+ with ResNet50 encoder model demonstrates excellent overall performance, with high precision, recall, and F1-scores across all classes. The model's performance is particularly strong for dominant classes like background and woodland. While segmentation of roads and buildings shows improvement compared to previous models, further optimizations or post-processing techniques may still be necessary to enhance performance for challenging classes.

REFERENCES

[1] Priyanka, N. S., Lal, S., et al. *DIResUNet: Architecture for multiclass semantic segmentation of high-resolution remote sensing imagery data.* Applied Intelligence **52**, 15462–15482 (2022). https://doi.org/10.1007/s10489-022-03310-z.

[2] Khan, S. D., Alarabi, L., & Basalamah, S. *Segmentation of farmlands in aerial images by deep learning framework with feature fusion and context aggregation modules.* Multimedia Tools and Applications **82**, 42353–42372 (2023). https://doi.org/10.1007/s11042-023-14962-5.

[3] Khan, Bakht Alam, & Jung, J.-W. *Semantic Segmentation of Aerial Imagery Using U-Net with Self-Attention and Separable Convolutions.* Applied Sciences **14**(9), 3712 (2024). https://doi.org/10.3390/app14093712.

[4] Tong, X.-Y., Xia, G.-S., Lu, Q., et al. *Land-cover classification with high-resolution remote sensing images using transferable deep models.* Remote Sensing of Environment **237**, 111322 (2020). https://doi.org/10.1016/j.rse.2019.111322.

[5] Alem, A., & Kumar, S. *Deep Learning Methods for Land Cover and Land Use Classification in Remote Sensing: A Review.* Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), pp. 903-908 (2020). https://doi.org/10.1109/ICRITO48877.2020.9197824.

[6] Fu, X., Zhang, Y., & Fan, S. *Real-Time Semantic Segmentation of Aerial Images Based on Dual-Feature Attention Networks.* Proceedings of the 2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE), pp. 399-402 (2024). https://doi.org/10.1109/NNICE61279.2024.10498691.

[7] *LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery.* https://arxiv.org/abs/2005.02264 (2020).

[8] Dutta, A., & Zisserman, A. *The VIA Annotation Software for Images, Audio and Video.* https://arxiv.org/abs/1904.10699 (2019).

[9] Chen, L.-C., Zhu, Y., Papandreou, G., et al. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.* https://arxiv.org/abs/1802.02611 (2018).

[10] Chollet, F. *Xception: Deep Learning with Depthwise Separable Convolutions.* https://arxiv.org/abs/1610.02357 (2016).

[11] Dai, J., Qi, H., Xiong, Y., et al. *Deformable Convolutional Networks.* https://arxiv.org/abs/1703.06211 (2017).

[12] Chen, L.-C., Yang, Y., Wang, J., et al. *Searching for Efficient Multi-Scale Architectures for Dense Image Prediction.* https://arxiv.org/abs/1809.04184 (2018).

[13] Stewart, G., & Korzeniowski, M. *TorchGeo: Deep Learning With Geospatial Data.* https://arxiv.org/abs/2111.08872 (2021).

[14] Li, W., & Dong, L. *Supervising Remote Sensing Change Detection Models with 3D Surface Semantics.* https://arxiv.org/abs/2202.13251 (2022).

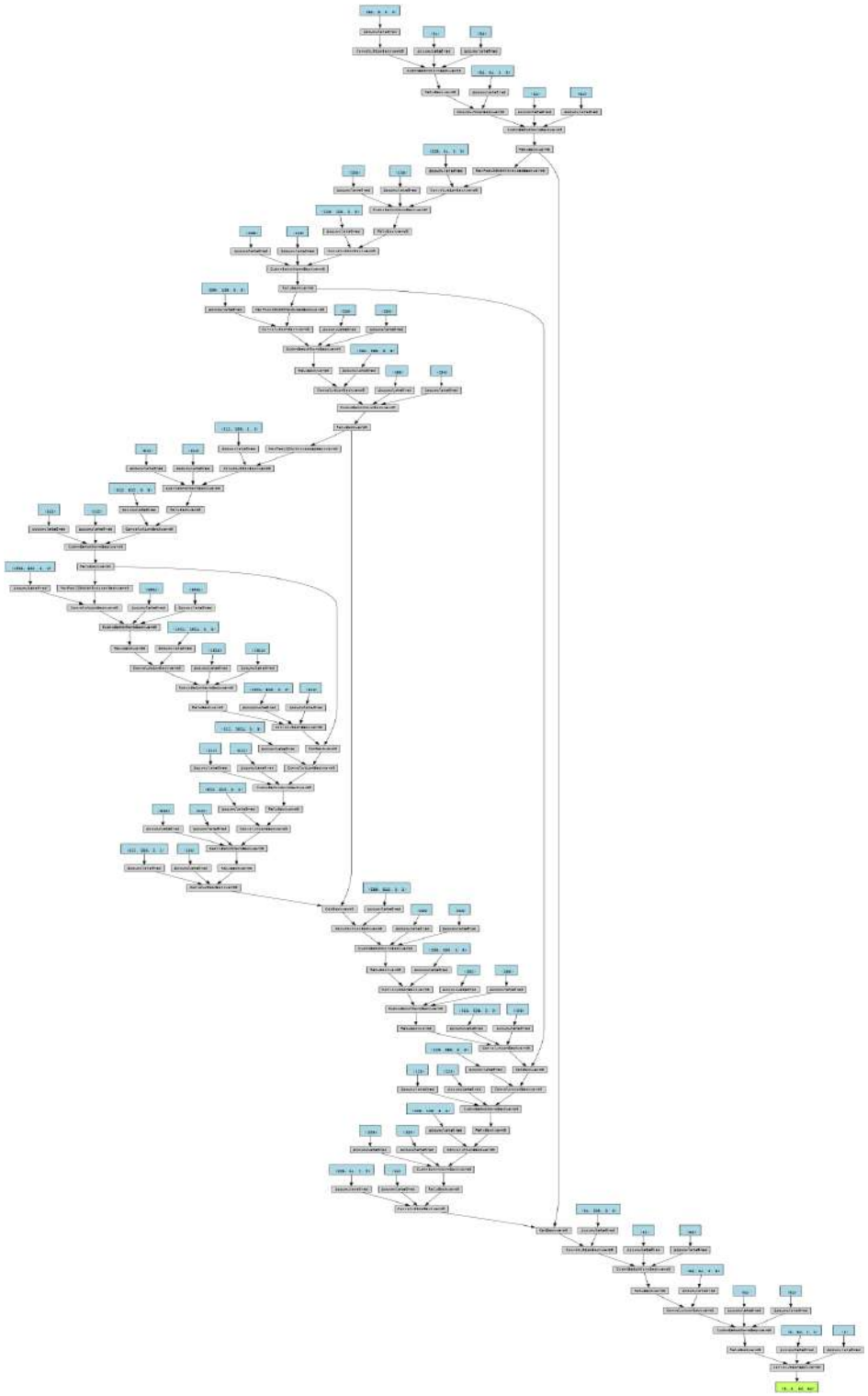
[15] Ronneberger, O., Fischer, P., & Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation.* https://arxiv.org/abs/1505.04597 (2015).

APPENDIX

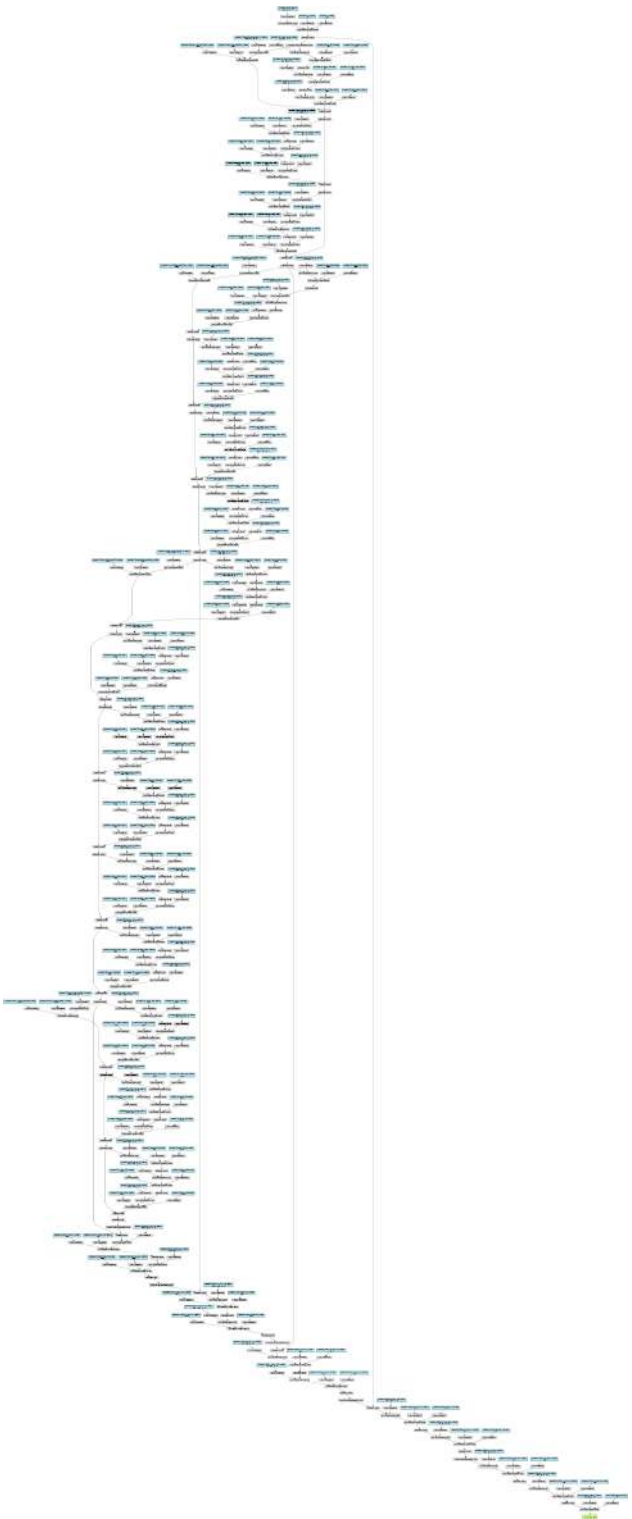
The following table presents the training and validation loss for all the models trained. The training was carried out for a total of 30 epochs, with early stopping triggered at epoch 23 for both the U-Net with ResNet-50 and DeepLabV3+ models. Early stopping occurred when no improvement was observed in the validation loss for 3 consecutive epochs, indicating that the models had converged to their optimal performance.

Epoch	**U-Net**		**U-Net with ResNet-50**		**DeepLabV3+ with Resnet-50**	
	Train Loss	**Eval Loss**	**Train Loss**	**Eval Loss**	**Train Loss**	**Eval Loss**
1	0.7326	0.6621	0.6141	0.4350	0.5263	0.3681
2	0.6507	0.5960	0.4157	0.3220	0.3866	0.3143
3	0.6084	0.5439	0.3455	0.2819	0.3438	0.2925
4	0.5539	0.5040	0.3150	0.2675	0.3199	0.2888
5	0.5253	0.4375	0.3074	0.2692	0.3112	0.2891
6	0.4866	0.4160	0.2891	0.2487	0.2882	0.2815
7	0.4639	0.4068	0.2791	0.2437	0.2824	0.2579
8	0.4425	0.3721	0.2720	0.2461	0.2807	0.2517
9	0.4291	0.3708	0.2598	0.2569	0.2716	0.2412
10	0.4263	0.3778	0.2668	0.2390	0.2700	0.2451
11	0.4133	0.3636	0.2502	0.2370	0.2592	0.2387
12	0.4144	0.3554	0.2483	0.2301	0.2569	0.2433
13	0.3986	0.3370	0.2568	0.2362	0.2436	0.2350
14	0.3961	0.3426	0.2512	0.2293	0.2465	0.2290
15	0.3836	0.3364	0.2436	0.2230	0.2469	0.2309
16	0.3850	0.3284	0.2406	0.2300	0.2437	0.2378
17	0.3756	0.3273	0.2369	0.2184	0.2544	0.2236
18	0.3718	0.3265	0.2320	0.2165	0.2391	0.2216
19	0.3625	0.3247	0.2343	0.2351	0.2362	0.2242
20	0.3672	0.3302	0.2369	0.2128	0.2302	0.2198
21	0.3627	0.3154	0.2289	0.2142	0.2262	0.2367
22	0.3540	0.3085	0.2277	0.2218	0.2299	0.2229
23	0.3571	0.3134	0.2385	0.2239	0.2254	0.2239
24	0.3482	0.3228	-	-	-	-
25	0.3501	0.3040	-	-	-	-
26	0.3434	0.3035	-	-	-	-
27	0.3409	0.2997	-	-	-	-
28	0.3401	0.3094	-	-	-	-
29	0.3419	0.2812	-	-	-	-
30	0.3294	0.2872	-	-	-	-

Architechture of Vanilla U-Net



Architechture of U-Net with Resnet-50 Encoder



Architechture of DeepLabV3+ with Resnet-50 Encoder