Semantic Segmentation of Aerial Images:
A Comparison of U-Net and SegNet Approaches

Cveevska Marija
marija.cveevska@studenti.unipd.it

Karakus Isikay
isikay.karakus@studenti.unipd.it

Abstract

This study compares the performance of two semantic segmentation models, UNet and SegNet, on aerial imagery. We describe the implementation, training, and testing of both models. The results show that UNet outperforms SegNet, particularly in this challenging task. UNet achieved Dice coefficients of 0.87 for training and 0.81 for validation, while SegNet scored 0.69 for training and 0.67 for validation. These metrics highlight UNet's superior accuracy and robustness.

GitHub Repository: Semantic Segmentation of Aerial Images

1. Introduction

Semantic segmentation is a crucial task in computer vision that involves the process of partitioning an image into coherent parts and labelling each part with a corresponding class. Unlike traditional image classification, which assigns a single label to an entire image, semantic segmentation provides a detailed, pixel-level understanding of the image. Each pixel is classified into a predefined category, such as road, building, vegetation, water, and other relevant classes. This pixel-level labelling enables a richer and more informative analysis of visual data.

The importance of semantic segmentation in the context of aerial imagery of Earth cannot be overstated. Aerial images, captured by satellites and drones, offer a comprehensive and large-scale view of the Earth's surface. This data is instrumental in numerous applications, including urban planning, agriculture monitoring, environmental monitoring, disaster response, and resource management. For instance, accurately segmented aerial images can help urban planners in identifying land use patterns, assist farmers in monitoring crop health, aid in tracking deforestation, and support emergency services in disaster-stricken areas.

Despite its significant potential, semantic segmentation of aerial images presents several challenges. Firstly, the variability in scale, orientation, and appearance of objects in aerial images complicates the segmentation task. Buildings, roads, and natural features can vary greatly in size and shape, making it difficult for a model to generalize across different images. Secondly, aerial images often contain high levels of detail and clutter, which can lead to confusion between similar-looking objects. For example, shadows cast by buildings might be mistaken for roads, or dense vegetation might obscure man-made structures.

Furthermore, aerial images are frequently affected by environmental factors such as varying lighting conditions, weather changes, and seasonal variations, which introduce additional complexity. Another challenge lies in the annotation process; labelling aerial images at the pixel level is labour-intensive and requires significant expertise, making it difficult to obtain large, accurately labelled datasets.

In this paper, we present a comparative analysis of two prominent deep learning models, U-Net and SegNet, for semantic segmentation of aerial imagery. We describe the implementation of a complete pipeline for training these models, including data preprocessing, model training, and evaluation. Our analysis includes a detailed examination of the results, highlighting the strengths and weaknesses of each model in this specific application context. Through this comparative study, we aim to provide insights into the effectiveness of U-Net and SegNet in handling the challenges posed by aerial image segmentation and contribute to the ongoing advancement of semantic segmentation techniques in remote sensing applications.

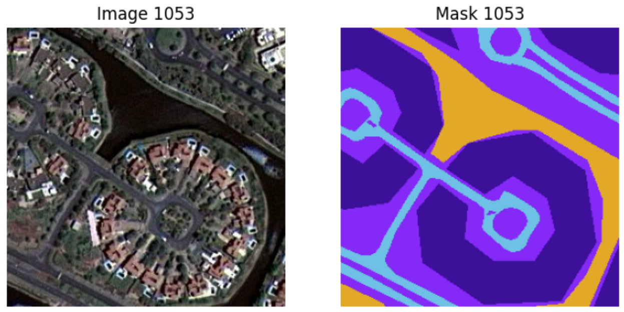

Figure 1. Image and Mask

2. Related Work

Semantic segmentation, a key task in scene understanding, involves partitioning an image into regions corresponding to different predefined classes. Recently, deep neural networks have been extensively used to address the problem of semantic segmentation, particularly for aerial and satellite imagery.

One notable approach is the Fully Convolutional Network (FCN) proposed by Long et al. [1], which modified the VGGNet by replacing fully connected layers with convolutional layers to preserve spatial information crucial for segmentation tasks. Following this, Badrinarayanan et al. [2] introduced SegNet, which extends VGG-16 with a deep deconvolutional part consisting of deconvolution and unpooling layers to predict pixel-wise class labels using the PASCAL VOC 2012 dataset.

The UNet architecture, proposed by Ronneberger et al. [3], has also been influential, especially in biomedical image segmentation. Its encoder-decoder structure with skip connections helps retain spatial information during the downsampling process, making it highly effective for the semantic segmentation of high-resolution images.

For aerial imagery, a CNN-based system for building and road extraction was proposed by Maggiori et al. [4]. This architecture processes RGB patches of aerial images using model averaging to prevent overfitting. The challenge of high-resolution satellite images versus low-resolution CNN inputs has led to the development of various approaches to improve segmentation accuracy, such as multi-scale active contour techniques and the JSEG algorithm used in the early 2000s.

More recent advancements include the application of UNet, Mask R-CNN, and graph neural networks for polygon segmentation in high-resolution imagery. These methods have shown significant improvements in delineating building boundaries and extracting features like roof polygons and faces [5, 6].

In addition to architectural innovations, feature extraction techniques have evolved. Classical texture analysis using statistical, spectral, structural, and model-based approaches has been complemented by modern methods like bag-of-visual-words and spatio-spectral networks. The combination of spatial and spectral features enhances the network's ability to capture deep color-texture properties [7, 8].

The evolution of semantic segmentation techniques for aerial and satellite imagery has advanced significantly with the advent of deep learning models. Architectures such as Fully Convolutional Networks (FCNs), UNet, and SegNet, along with advanced feature extraction methods, have markedly improved the accuracy and robustness of segmenting complex scenes. This paper presents our findings on the application of UNet and SegNet models to the semantic segmentation of aerial imagery of Dubai, obtained from MBRSC satellites [11].

3. Dataset

The dataset used in this study comprises aerial imagery of Dubai, obtained by satellites operated by the Mohammed Bin Rashid Space Centre (MBRSC). This dataset has been meticulously annotated with pixel-wise semantic segmentation, categorizing the imagery into six distinct classes:

- Building: `#3C1098`
- Land (unpaved area): `#8429F6`
- Road: `#6EC1E4`
- Vegetation: `#FEDD3A`
- Water: `#E2A929`
- Unlabeled: `#9B9B9B`

Each class is associated with a specific colour in hexadecimal format for clarity and consistency in the annotation.

The dataset is relatively small, consisting of 72 images in total. These images are grouped into 8 larger tiles, each representing a different section of Dubai. Each tile contains a folder with the original images and corresponding mask images that highlight the segmentation classes. The masks are encoded in RGB format, where each pixel's colour represents a class label according to the specified hexadecimal values. One of the main challenges presented by this dataset is its limited size, with only 72 images available for training and evaluation. This small sample size can impact the generalization capabilities of the trained models. Additionally, the images within each tile vary in height and width, which requires careful preprocessing to ensure consistency and compatibility with the input requirements of the deep learning models. Another challenge is the need to accurately convert the class colours from hexadecimal to RGB, as the masks are provided in RGB format, necessitating precise handling to maintain the integrity of the annotations.

Despite these challenges, this dataset provides a valuable resource for evaluating and comparing semantic segmentation models in the context of urban aerial imagery, offering a diverse set of scenes and features representative of a bustling metropolitan area like Dubai.

4. Segmentation Models

Semantic segmentation models are designed to classify each pixel in an image into one of several predefined categories, enabling detailed analysis of visual data. These models are essential for tasks that require a precise understanding of the spatial and structural properties of objects within an image. Two prominent deep learning models used for semantic segmentation are U-Net and SegNet.

4.1. U-Net

UNet is a convolutional neural network architecture designed for semantic segmentation tasks, characterized by its contracting and expansive paths. The network architecture comprises an encoder-decoder structure with skip connections that ensure the retention of spatial information.

The contracting path (encoder) on the left follows the standard convolutional network architecture. It consists of repeated application of two 3x3 convolutions (unpadded), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with a stride of 2 for downsampling. At each downsampling step, the number of feature channels doubles, capturing more detailed contextual information.

The expansive path (decoder) on the right involves upsampling the feature map followed by a concatenation with the corresponding feature map from the contracting path. This concatenation is performed to retain high-resolution features. Each upsampling step includes a 2x2 convolution (up-convolution) that halves the number of feature channels, followed by two 3x3 convolutions, each with a ReLU activation.

Boundary pixels are inevitably lost during convolutions, necessitating cropping. The final layer employs a 1x1 convolution to map the 64-component feature vector to the required number of classes. Overall, the network comprises 23 convolutional layers.

Our implementation of the UNet model includes batch normalization and dropout layers for regularization. The model is trained using the Adam optimizer with a learning rate and utilizes the Dice loss function, which is effective for segmentation tasks by focusing on the overlap between predicted and true masks. This implementation aims to leverage the original architecture's strengths while incorporating modern techniques for improved performance.

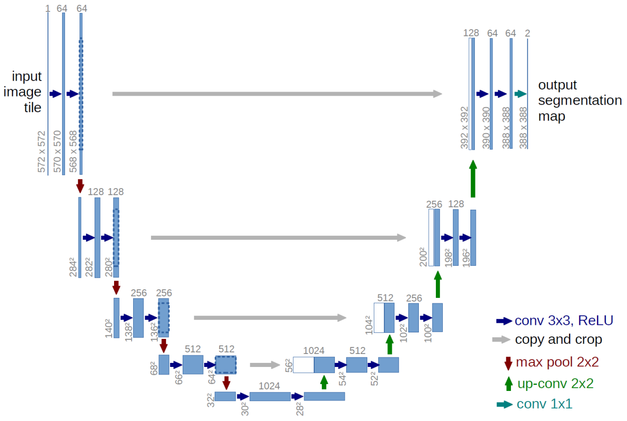

Figure 2. U-Net architecture

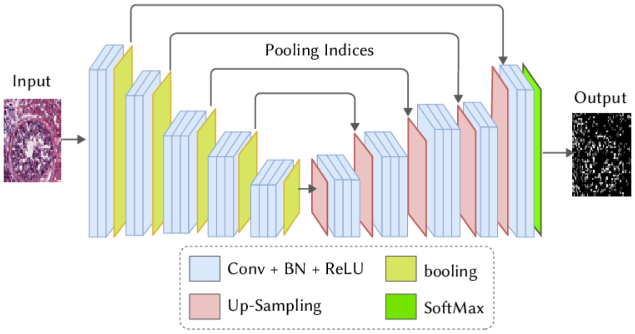

Figure 3. SegNet architecture

4.2. Seg-Net

SegNet is a deep, fully convolutional neural network architecture designed for semantic pixel-wise segmentation. It consists of an encoder network, a corresponding decoder network, and a pixel-wise classification layer. The novelty of SegNet lies in the manner in which the decoder upsamples its lower resolution input feature maps using pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. This eliminates the need for learning to upsample and allows for efficient memory and computation during inference.

SegNet differs from UNet in that only the pooling indices are transferred to the expansion path from the compression path, using less memory. In contrast, UNet transfers entire feature maps from the compression path to the expansion path, consuming more memory. The network takes the aerial imagery dataset as input and produces segmentation maps with a specified number of output channels corresponding to the number of classes. Batch normalization is used to stabilize and accelerate the training process.

The encoding part consists of multiple stages, each including convolutional layers, batch normalization, and max pooling. Max pooling is used for down-sampling, and the pooling indices are stored for use in the decoding layers. The number of filters in the convolutional layers increases with each stage, capturing more detailed features.

The decoding part mirrors the encoding stages, with each stage including upsampling (using max unpooling with stored indices), convolutional layers, and batch normalization. The goal of the decoding layers is to reconstruct the segmented output from the features obtained in the encoding layers. The number of filters in the convolutional layers decreases with each stage.

Our implementation of SegNet includes batch normalization and dropout layers for regularization. The model is trained using the Adam optimizer with a learning rate and the Dice loss function, focusing on the overlap between predicted and true masks. Unlike UNet, SegNet's architecture efficiently balances memory usage and computational requirements, making it a robust choice for semantic segmentation tasks.

5. Data Preprocessing

In the data preprocessing stage, we focus on preparing the aerial imagery dataset for semantic segmentation using both the UNet and SegNet models. The dataset, consisting of images and corresponding masks, is processed to ensure uniformity and manageability. Each image is loaded, cropped, and resized to dimensions divisible by a predefined patch size (256x256). This resizing ensures the images can be uniformly processed.

Images are transformed into arrays and divided into smaller patches for efficient neural network processing. Pixel values of these patches are scaled to a normalized range using MinMaxScaler to stabilize and expedite training. Corresponding masks, representing different classes through distinct RGB values, are converted into a categorical format by mapping specific RGB colors to class labels.

The transformed images and masks are assembled into arrays and visualized to verify preprocessing accuracy. The dataset is then divided into training and validation sets, ensuring an efficient model training process. This comprehensive preprocessing ensures the dataset is prepared for accurate semantic segmentation by both the UNet and SegNet models.

6. Experiments

We conducted our experiments using Google Colab Pro in a GPU environment to accelerate model training and ensure stability. The codebase, developed in Python 3, utilized PyTorch for both U-Net and SegNet models. Each model was trained for different epochs (10, 30, and 50) to observe the impact on learning and convergence. Additionally, we experimented with various learning rates (1e-4, 1e-5, and 1e-6) to determine the optimal step size for updating the model weights. Batch sizes of 2, 4, 8, and 16 were tested to study their effects on gradient estimates and training stability.

To enhance model generalization and robustness in the SegNet model, we applied data augmentation techniques, specifically colour shift and blurriness transformations. The ColorJitter transformation randomly adjusted the brightness, contrast, saturation, and hue of the images, helping the model learn under varying lighting conditions. The GaussianBlur transformation applied a Gaussian blur to reduce noise and focus on significant features. The Compose function (transforms.Compose) allowed us to chain these transformations together so they were applied sequentially to the images. In this case, the color shift was applied first, followed by the Gaussian blur. This sequential application ensured comprehensive augmentation, enhancing the diversity of the training dataset and thereby contributing to the robustness and accuracy of the SegNet model.

We employed both a standard train-validation split and a stratified k-fold cross-validation to ensure robust evaluation and mitigate the risk of biased performance metrics. To regularize the models and enhance their generalization to unseen data, we tested various dropout rates. These comprehensive experiments provided insights into the influence of each hyperparameter on model performance, guiding us towards optimal settings for accurate semantic segmentation.

We chose the Dice loss function for its effectiveness in segmentation tasks, as it measures the overlap between predicted and ground truth masks. The Adam optimizer was employed, and dropout was applied to prevent overfitting.

6.1. Performance Metrics

Sørensen–Dice Coefficient (DSC): The Sørensen-Dice coefficient is a widely used metric in semantic segmentation tasks. It measures the similarity between two sets by calculating the ratio of twice the intersection of the sets to the sum of their cardinalities.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|} \tag{1}$$

Recall: Recall measures the ability of a model to correctly identify all relevant instances of a specific class within an image. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Accuracy: Accuracy measures the proportion of correct predictions out of the total predictions made. It is a general metric used for classification tasks. In the context of segmentation, it is calculated as the ratio of the sum of true positives and true negatives to the sum of true positives, true negatives, false positives, and false negatives.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Intersection over Union (IoU): Also known as the Jaccard Index, IoU is a crucial metric in image processing. It evaluates the similarity between two sets by dividing the size of the intersection by the size of the union of the sets.We used this metric in the final image Prediction

$$IoU(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{4}$$

These performance metrics provide a comprehensive evaluation of the models' segmentation accuracy, highlighting their effectiveness in capturing specific classes within the aerial imagery of Dubai dataset. This detailed evaluation underscores the robustness and precision of the models in accurately segmenting complex scenes from aerial imagery.

7. Results

The results highlight the performance of the U-Net and SegNet models on the semantic segmentation of aerial imagery from the Dubai dataset. Our evaluation metrics included Dice Loss, Recall, and Accuracy for both training and validation datasets. Additionally, we assessed the models' predictive capabilities through visual comparison of real and predicted masks, as well as quantitative metrics such as Intersection over Union (IoU), Recall, and Accuracy.

7.1. U-Net Results

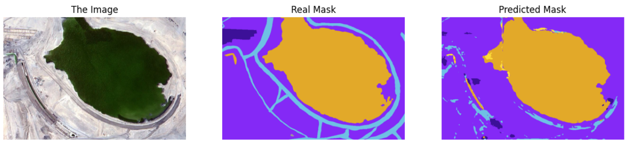

Figure 4. U-Net Predicted Mask

The U-Net model demonstrated high performance across both training and validation datasets. The low Dice Loss values indicate that the model effectively minimized segmentation errors. Specifically, U-Net exhibited a Dice loss of 0.13 (Dice coefficient of 0.87) during training and 0.19 (Dice coefficient of 0.81) during validation, reflecting its strong segmentation capabilities. Additionally, the model achieved high overall accuracy and recall, with training accuracy at 87% and validation accuracy at 81%. The recall values were also significant, with 81% for training and 79% for validation. These metrics underscore the robustness of the U-Net model in accurately identifying and classifying various classes within the aerial imagery dataset. U-Net's mask prediction performance is illustrated in Figure 4.

7.1.1 Class-Specific Performance of U-Net:

U-Net	**Label**	**Dice Loss**	**Acc.**	**Recall**
Train	**Total**	0.13	87	81
	Class 0: Building	0.39	60	80
	Class 1: Land	0.13	88	70
	Class 2: Road	0.41	57	74
	Class 3: Vegetation	0.50	40	80
	Class 4: Water	0.44	55	86
	Class 5: Unlabeled	0.25	0	94
Validation	**Total**	0.19	81	79
	Class 0: Building	0.48	39	74
	Class 1: Land	0.17	88	70
	Class 2: Road	0.43	47	72
	Class 3: Vegetation	0.60	33	79
	Class 4: Water	0.50	56	87
	Class 5: Unlabeled	0.34	0	93

Table 1. Performance of U-Net on Training and Validation Sets

Land Class: The 'Land' class, being the most frequent in the dataset, exhibited the best performance. It achieved a training accuracy of 88% and recall of 70%, with similar validation performance (88% accuracy, 70% recall). This consistency underscores the model's effectiveness in learning and generalizing for prevalent classes.

Unlabeled Class: The 'Unlabeled' class presented unique challenges because it is underrepresented in the dataset. Despite a high recall of 94% during training, indicating the model's ability to identify 'Unlabeled' pixels in the ground truth, its accuracy was zero. This discrepancy arises because accuracy also considers true negatives, and the small number of true positives for 'Unlabeled' skews the accuracy calculation. This pattern continued in the validation set, with a recall of 93% but zero accuracy.

Other Classes: Other classes like 'Building', 'Road', and 'Vegetation' showed moderate performance. For example, the 'Road' class had a validation accuracy of 47% and recall of 72%, indicating that not enough data compare the land class. Results are shown in the table 1.

The effectiveness of U-Net in segmenting the 'Land' class highlights the importance of class distribution in training. Classes with more data tend to have better-learned representations, resulting in higher performance metrics. Conversely, the significant challenges with the 'Unlabeled' class demonstrate the difficulties of dealing with rare classes.

7.2. SegNet Results

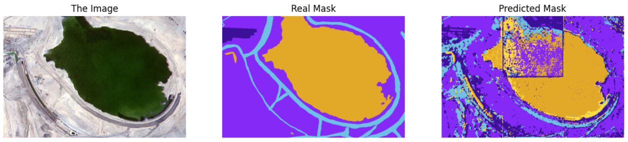

Figure 5. SegNet Predicted Mask

SegNet exhibited higher Dice Loss values and lower Recall and Accuracy compared to U-Net. The total Dice loss for SegNet was 0.31 during training and 0.33 during validation, indicating less effective segmentation performance. The model's overall accuracy was 80% for training and 78% for validation, with recall values of 74% for training and 73% for validation. The Dice coefficients for SegNet, which measure the similarity between predicted and ground truth masks, were also lower (0.69 for training and 0.67 for validation) compared to U-Net, further highlighting the model's relative inefficiency. However, due to its architecture, SegNet is faster compared to U-Net, making it more efficient in terms of computational speed. Predicted masks using the SegNet architecture are presented in Figure 5.

7.2.1 Class-Specific Performance of SegNet:

SegNet	Label	Dice Loss	Acc.	Recall
Train	**Total**	0.31	80	74
	Class 0: Building	0.54	67	78
	Class 1: Land	0.23	84	68
	Class 2: Road	0.54	49	72
	Class 3: Vegetation	0.7	38	80
	Class 4: Water	0.6	48	85
	Class 5: Unlabeled	0.98	0	94
Validation	**Total**	0.33	78	73
	Class 0: Building	0.55	51	77
	Class 1: Land	0.23	76	65
	Class 2: Road	0.55	23	67
	Class 3: Vegetation	0.69	22	77
	Class 4: Water	0.62	55	85
	Class 5: Unlabeled	0.99	0	93

Table 2. Performance of SegNet on Training and Validation Sets

Land Class: The 'Land' class, which is the most prevalent in the dataset, achieved the highest performance metrics. During training, it reached an accuracy of 84% and a recall of 68%. In validation, the model maintained results with 76% accuracy and 65% recall.

Unlabeled Class: The 'Unlabeled' class exhibited poor accuracy due to its rarity but achieved high recall (80% during training and 85% during validation). This high recall indicates the model's ability to identify 'Unlabeled' pixels in the ground truth. However, low accuracy, influenced by the high number of true negatives, highlights the model's struggle with false positives.

Other Classes: Classes such as 'Building', 'Road', 'Vegetation', and 'Water' showed variability in performance. For instance, the 'Road' class had a validation accuracy of 23% and a recall of 67%.

While SegNet can effectively segment prevalent classes, it struggles with less common ones. This underlines the necessity of balanced class distributions or specialized techniques to manage class imbalances for training effective segmentation models.

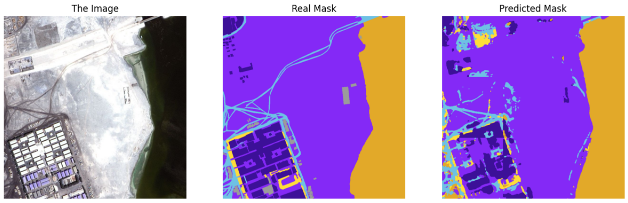

Figure 6. U-Net 2

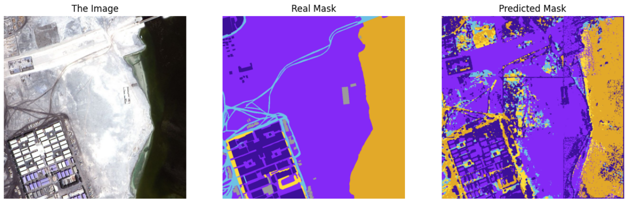

Figure 7. SegNet 2

Visual inspection of the real versus predicted masks for both models indicates that U-Net delivers more accurate and detailed segmentation, particularly in complex areas with multiple overlapping classes. U-Net's predictions consistently capture finer details and show fewer segmentation errors. In contrast, SegNet, while generally accurate, tends to miss finer details and exhibits more segmentation inaccuracies. The evaluation metrics for U-Net showed a Global Accuracy of 80.44%, a Mean IoU of 42.14%, and a Recall of 57.15%. On the other hand, SegNet achieved a Global Accuracy of 59.11%, a Mean IoU of 22.54%, and a Recall of 45.18%. The comparison between the two models' predicted masks can be seen in Figures 6 and 7.

Overall, U-Net consistently outperforms SegNet, demonstrating lower Dice loss and higher accuracy and recall across most classes. U-Net's lower Dice loss values reflect its effectiveness in minimizing segmentation errors, while its higher accuracy and recall metrics indicate robust performance in correctly identifying and classifying various classes within the imagery. Conversely, SegNet's higher Dice loss and lower performance metrics suggest greater challenges in accurately identifying and classifying the different classes in the dataset.

Both models highlight the critical importance of class distribution in training segmentation models. U-Net's superior performance underscores the advantages of effective learning and generalization from prevalent classes. In contrast, SegNet's difficulties with rare classes emphasize the need for specialized techniques to handle class imbalances, ensuring robust performance across all classes. This comparison illustrates that while U-Net excels in segmentation accuracy and recall, addressing class imbalances is essential to improve the performance of models like SegNet.

8. Conclusion

Overall, the U-Net model outperforms SegNet in both quantitative metrics and visual assessments. The experiments demonstrate that U-Net's architecture, with its encoder-decoder structure and skip connections, is better suited for the semantic segmentation of high-resolution aerial imagery. The insights gained from these experiments provide a strong foundation for further optimization and application of deep learning models in semantic segmentation tasks.

By varying key hyperparameters such as epochs, learning rates, and batch sizes, and employing robust validation techniques, we were able to achieve these performances for aerial image semantic segmentation.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015. [Online]. Available: https://arxiv.org/abs/1511.00561

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[4] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution semantic labeling with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017. [Online]. Available: https://ieeexplore.ieee.org/document/7952936

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017*, 2017. [Online]. Available: https://arxiv.org/abs/1609.02907

[7] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 61–81, 2009. [Online]. Available: https://link.springer.com/article/10.1007/s11263-005-4635-4

[8] V. Andrearczyk and P. F. Whelan, "Using filter banks in convolutional neural networks for texture classification," *Pattern Recognition Letters*, vol. 84, pp. 63–69, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865516301756

[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/pdf/2010.11929

[10] E. Guérin, K. Oechslin, C. Wolf, and B. Martinez, "Satellite Image Semantic Segmentation," *arXiv preprint arXiv:2110.05812*, 2021. [Online]. Available: https://arxiv.org/pdf/2110.05812

[11] Humans in the Loop, "Semantic Segmentation of Aerial Imagery Dataset," 2021. [Online]. Available: https://humansintheloop.org/resources/datasets/semantic-segmentation-dataset-2/