Journal of Information Systems Engineering and Management
2025, 10(31s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Transformer based UNet for Semantic Segmentation on Aerial Imagery

K. Suvarna Vani[1*], N. Sai Sruthi[2], Sk. Ershathunnisa[3], Velagapudi Sreenivas[4], G.Srilakshmi[5]

[1,2,3] Department of Computer Science and Engineering, V R Siddhartha School of Engineering, Siddhartha Academy of Higher Education, Deemed to be University, Vijayawada, Andhra Pradesh, India

[4]Department of Computer Science and Engineering, SRK Institute of Technology, Vijayawada, A.P, India

[5]Department of Information Technology, SRK Institute of Technology, Vijayawada, A.P, India

Email: [1*]suvarnavanik@gmail.com, [2]saisruthi.namala@gmail.com, [3]ershatunnisa2004@gmail.com, [4]velagapudisreenivas@gmail.com, [5]sree.gpk@gmail.com

ARTICLE INFO

Received: 30 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Semantic segmentation is an integral component of computer vision, providing detailed scene analysis by classifying each pixel in an image. It is particularly valuable in remote sensing applications, such as land cover mapping, urban change detection, and environmental protection. However, semantic segmentation often faces challenges in capturing both local and global context effectively. Traditional machine learning models encounter limitations with suboptimal feature extraction, handling noisy data, and adapting to varying data distributions. To address these challenges, deep learning models offer improved adaptability and feature learning capabilities. In particular, Transformer architectures have shown promise in modelling global information, leading to enhanced performance in various vision-related tasks, including semantic segmentation. In this work, we propose a novel approach that integrates a Transformer-based decoder into the U-Net architecture for real-time urban scene segmentation. The model combines a CNN-based encoder, utilizing ResNet-101 for feature extraction, with a Transformer-based decoder to capture both local and global contexts. This hybrid architecture allows for better complex urban element segmentation, making the model better at defining fine details and also large-scale structures. For performance evaluation, the proposed model is tested against UAVid, which results to an 89% accuracy and an 80% of MIoU; thus, confirming that the proposed model is effective in achieving a good outcome in the urban scene segmentation process.

Keywords: Semantic Segmentation, Deep Learning, Aerial Imagery, U-Net, ResNet-101, Transformer.

INTRODUCTION

Semantic segmentation is a basic computer vision concept, that is comprised of assigning to each pixel of an image a class label. This kind of pixel-level classification sets semantic segmentation apart from image classification and region-level object detection and gives finer details about the scene. Semantic segmentation is a very crucial application in many scenes, including autonomous driving that will enable safe navigation through the understanding of road scenes and, in robotics and surveillance, by action localization and classification. Moreover, in remote sensing, semantic segmentation works hand in hand with land cover mapping, urban change detection, and environmental protection. Furthermore, in medical imaging, the key ap plication of semantic segmentation is in the detection of the tumor and in the segmentation of organs.

Initially Semantic Segmentation was mostly done with Machine learning models which often struggled with effectively capturing the nuances of complex scenes. To enhance performance, it is crucial to consider both local and global contexts in the segmentation process. Local context refers to the detailed, small-scale information found in a limited region of an image. It captures fine-grained features, such as textures, edges, and small objects. Local context is crucial for distinguishing between objects that are close together or have subtle differences, like differentiating between a parked car and a moving car, or detecting small details within an object. Global con text refers to the large-scale, holistic information that spans the entire image or large portions of it. It captures relationships between objects and their surroundings, as well as the overall structure of the scene. Understanding that a road runs through an entire image, or that a building is part of a larger urban scene. It helps the model recognize spatial relationships between objects, such as knowing that cars are likely on roads and trees are more likely next to buildings. It prevents misclassification by using the scene's overall context (e.g., not classifying a tree as a car because of its larger environment). Traditional machine learning methods for semantic segmentation have relied on manually designed feature extractors and classifiers, such as conditional random fields (CRFs), support vector machines (SVMs), K-means 2 and decision trees. While these methods offer certain advantages, particularly in small-sample datasets or low-noise images, they face limitations. CRFs, although previously popular, require extensive manual feature

engineering and parameter tuning, resulting in high computational complexity and inefficiency. SVMs, though effective for binary classification, struggle with multiclass problems. Decision trees, despite being simple and interpretable, are often hindered by high-dimensional data and noise, leading to poor generalization.

The advent of deep learning has revolutionized semantic segmentation. Convolutional neural networks (CNNs), Deep Neural Networks(DNN)[14] with their automated feature extraction capabilities, have replaced traditional manual methods. Pioneering models like LeNet-5 paved the way for more sophisticated architectures such as GoogLeNet, VGG, ResNet and AlexNet. These networks have significantly advanced the field, enabling end-to-end processing and pixel-level classification through fully convolutional networks (FCNs). However, FCNs have limitations in label localization, global context handling, and multiscale processing. To address these, various architectures have been proposed. U-Net, for instance, employs a U-shaped structure to enhance context and location information, particularly in medical image seg mentation. The DeepLab family, through versions V1 to V3+, has introduced innovations like atrous spatial pyramid pooling (ASPP) and encoder-decoder architectures to maintain resolution and improve performance[11,13].

Despite these advancements, challenges remain, especially in efficiently capturing global context and hierarchical features. The recent introduction of transformer-based architectures, originally developed for natural language processing, has shown promise in addressing these issues. The Vision Trans former (ViT) like Swin Transformer[15] and its variants leverage long-range dependency modeling capabilities, offering significant performance gains in vision tasks. The encoder-decoder attention mechanism within transformers facilitates effective sequence-to-sequence transformations, enhancing predictions and providing a new direction for semantic segmentation research. In this work, we propose an innovative approach that integrates a Transformer-based decoder into the U-Net architecture for real-time urban scene segmentation. By combining a CNN-based encoder, specifically ResNet 101, with a Transformer-based decoder, we aim to harness both local and global contexts within the image while reducing computational complexity. ResNet-101's robust feature extraction capabilities enhance segmentation accuracy, providing a detailed and comprehensive understanding of urban 3 scenes. Our approach addresses the limitations of traditional and deep learning models, offering a promising solution for complex semantic segmentation tasks in diverse and dynamic environments.

RELATED WORK

Xiujuan Li et al.[1] proposed a Multi-Feature Fusion and Channel Attention Network(MFCA-Net) which is built on an encoding-decoding structure, with an improved MobileNet V2 (IMV2) and Multi-Feature Dense Fusion (MFDF) in the encoding section. The authors of this paper uses two datasets, they are Vaihingen and Gaofen Image Dataset (GID).The experimental results of this paper shows that MFCA-Net achieves 76.77 MIoU on Vaihingen and 73.94 on GID dataset. Although this model improves segmentation ac curacy and provides accurate boundary delineation of easily confused low vegetation and trees, Sometimes this model typically require large amounts of annotated training data to achieve optimal performance.

Zhongchen Wang et al.[2] proposes a Dual Encoder-Decoder Network for land cover remote sensing image segmentation. The authors of this paper uses CNN based encoder-decoder and Transformer based encoder-decoder. For CNN based encoder they used ResNet and transformer based encoder as Swin-T for parallel extraction of features. The authors used CF mod ule and NAG unit in decoder 1 by integrating self attention technique and BiseNet approach. These modules help to fully integrate the outputs of the CNN based Encoder and Transformer based Encoder at the same stage during decoding. They also use MFE module in decoder 2 along with skip connections for performing multiscale fusion of low-resolution high-channel feature maps after preliminary feature aggregation. The authors created a dataset named building and water dataset and used Gaofen Image Dataset (GID), L8SPARCS Dataset. The authors achieved 90.52 MIoU. Although the Accurate extraction of both local and global features, Effectively handling inter class ambiguity and intra class inconsistency this model results in Increase in Computational Complexity, results in overfitting for small datasets.

Venugopal et al.[3] introduce the Adaptive DeepLabv3+ model, a novel approach for semantic segmentation of UAV images that integrates DeepLabv3 with the Improved Golden Eagle Optimization Algorithm. This model lever ages ASPP with more dilation rates in encoder to efficiently capture multi-scale context information. The authors used MBRSC satellite data and aerial image segmentation dataset. The authors achieved Accuracy of 98.4% on 98.3% respectively. One of the key advantages of this approach is its ability to prevent premature convergence to suboptimal solutions. The proposed model requires adequate and diverse training data. The proposed model re quires adequate and diverse training data and computation time for dataset 1 and 2 was 136.8912 and 147.2684 seconds, respectively, which might be considered high for real-time application.

Wang et al.[4] introduce the Adaptive Feature Fusion U-Net (AFF-UNet).The model incorporates dense skip connections, an Adaptive Feature Fusion Mod ule, a Channel Attention Convolution Block(CACB), and a Spatial Attention Module by addressing the challenges like handling different sizes of objects and easily confused geo-objects by fusing context information and automatically assigning weights to different levels of feature blocks. The authors of this paper use the Potsdam and BDCI dataset and achieve 71.44,70.5 MIoU respectively.

Xing et al. [5] introduces FCUnet, which integrates. FCUnet is designed to enhance remote sensing image analysis through three main components: a deep convolution U-Net for multi scale feature abstraction, fuzzy logic units to handle uncertainties and refine segmentation, and a CRF module to incorporate spatial context and reduce spectral variability. This combination improves feature representation and segmentation accuracy by addressing the inherent uncertainties in remote sensing images. The authors used featured prediction competition (FPC) automobile data set, ISPRS, (CCF) China Computer federation. The authors achieves 0.9204 ± 0.0031 MIoU, 0.7941±0.0039 MIoU, 0.8926±0.0040 MIoU respectively. Although FCUnet provides faster inference time, It lacks of Sensitivity to shadows and misclassification of ground images.

Xin-Yi Tong et al.[6] proposed semi-automatic land cover classification scheme integrates convolutional neural network (CNN)-based image classification with interactive segmentation guided by user inputs. Initially, a CNN is employed to classify images patch-wise, providing preliminary object positions and categories. Subsequently, an interactive segmentation process is initiated, where user-guided clicks on object boundaries inform the segmentation within the patches. This interactive approach allows for finer delineation of objects and improves accuracy. The methodology is evaluated using a comprehensive dataset from Jiangsu Province, China, comprising aerial and satellite imagery, and encompasses five common land cover categories.

Zhang et al. [7] introduces TrSeg, a novel semantic segmentation network that leverages transformer architecture to efficiently capture multi-scale contextual information. Unlike traditional methods that incorporate multi scale information by integrating individual single-scale features, TrSeg integrates a transformer decoder to dynamically capture multi-scale information. The model is evaluated on two benchmark datasets: Cityscapes and RUGD. The authors of this paper uses ResNet-101 as backbone and model TrSeg achieves 79.9 MIoU on Cityscapes, ResNet- 50 as backbone and model TrSeg achieves 33.91 MIoU. Although TrSeg Outperforms other methods in capturing multi-scale information by large margins it requires additional parameters compared to other methods, which may increase the computational cost.

Wang et al.[8]. proposed a model EGDE-Net, is a specialized neural network architecture that integrates edge-guided features to enhance change detection performance. It incorporates an FDE module designed to learn and emphasize discriminative change feature maps, which are crucial for identifying differences in building structures over time. The model is evaluated on two datasets: the WHU Building Change Detection (CD) dataset and the LEVIR-CD dataset. EGDE-Net include its accurate boundary detection and its robust feature learning enabled by the FDE module, which improves the identification of changed and unchanged areas. However, the model faces challenges such as class inconsistencies and its heavy dependency on the quality and variety of datasets, which could affect its performance in unseen scenarios.

Chowdhury et al.[9]. proposed ResUNet-a model. The methodology in volves using a UNet backbone with residual connections to facilitate better gradient flow, atrous convolutions for capturing multi-scale contextual information, and Pyramid Scene Parsing Pooling to enhance scene understanding. A multi-tasking inference approach is also employed to sequentially predict object boundaries and segmentation masks. The model includes superior segmentation accuracy and improved handling of class imbalance, aided by a novel variant of the Generalized Dice loss function. However, there is an increased computational complexity, which may limit real-time applications. When evaluated on the ISPRS Potsdam dataset, ResUNet-a achieves an impressive average F1 score of 92.9%, demonstrating its effectiveness in remote sensing tasks.

*Gupta et al.[*10]. aim to develop a deep learning framework for aerial image segmentation to aid in disaster impact assessment and management, particularly in post-disaster scenarios like hurricanes and tsunamis. This research focuses to improve segmentation performance, along with open data from OpenStreetMap (OSM) to bypass the need for manual annotation, graph theory is applied to update road network data and identify changes caused by natural disasters highlighting the use of open-source data, eliminating time-consuming annotations, and a reduction in model complexity with ENetSeparable, which uses 30 percent fewer parameters than ENet while delivering comparable performance to state-of-the-art networks. However, the reliance on OSM data could pose a challenge, that if the data is outdated or incomplete, potentially affecting the accuracy of the model.

Based on the various approaches discussed, we decided to adopt ResNet 101 for feature extraction and a Transformer-based decoder for feature map ping. This combination allows us to effectively capture both local and global context, which is critical for achieving accurate segmentation results. While many existing models face challenges such as requiring large datasets, mis classification of ground features, overfitting, and increased computational complexity, our approach addresses these limitations by leveraging the strengths of CNN-based encoders for capturing fine-grained local features and Transformer-based decoders for integrating global context.

METHODOLOGY

1.1. Data Collection

The dataset we have collected is the UAVid Dataset that is publicly available in ISPRS. The dataset features images and videos captured from both oblique. These images are recorded at high resolutions ensuring clear visibility

and the ability to differentiate objects effectively, even those at a distance. To introduce diversity and prevent overfitting in learning algorithms, the dataset captured 30 different video sequences in different locations under favourable weather conditions with ample lighting. The data collection pro cess utilized modern, lightweight drones like the DJI Phantom 3 Pro and DJI Phantom 4, ensuring steady flight and clear imagery necessary for effective analysis and model training. UAVid 2020 version has 42 sequences in total. Besides the original 30 sequences (UAVid10 version), another 12 sequences have been collected to further strengthen the dataset.

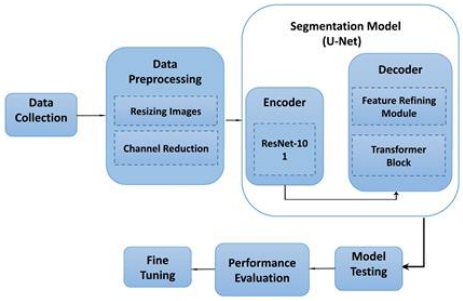


Figure 1- Proposed Model

Classes: 8(Building, Road, Static car, Tree, Low vegetation, Human, Moving car, Background, clutter).

Number of Images : 420.

Train set size: 200.

Test set size: 150.

Validation set size: 70.

Image resolution: 4096 x 2160 or 3840 x 2160.

1.2. Data Preprocessing

Data preprocessing refers to a series of essential steps undertaken to refine raw image data before it is fed into a segmentation model. This preparatory phase is pivotal in enhancing the quality of input data and subsequently improving the segmentation model's accuracy and performance.

1.2.1. Image Resizing

It is a fundamental preprocessing step in data preprocessing and computer vision tasks like image classification, object detection, and semantic segmentation. This is where one resizes the dimensions of an image concerning width and height but maintains the aspect ratio or allows the change to it, if necessary. This was a method that facilitated the model processing a smaller portion of the image but helped it better preserve spatial detail and improve segmentation accuracy. To maintain the consistency, every one high-resolution image was padded and then cropped into eight patches of size 1024 x 1024 pixels. This way, the model handles these regions of the image as segments to preserve spatial detail for further improvement of the accuracy of segmentation.

Algorithm 1 Image Resizing for Data Preprocessing

Input: Image dataset $D = [I_1, I_2, \ldots, I_n]$, Target size T
Output: Resized dataset $D_{\text{resized}} = [I'_1, I'_2, \ldots, I'_n]$

```
for i = 1 to n do
    Load image I_i
    Resize I_i to target size T using interpolation
    Store the resized image I'_i in D_resized
end for
Return resized dataset D_resized
```

Algorithm 1, titled "Image Resizing for Data Preprocessing" takes an input image dataset and resizes each image in the dataset to a specified target size. It uses interpolation during the resizing process to maintain image quality and aspect ratio. The output is a resized dataset containing images with dimensions matching the target size.

1.2.2. Channel Reduction

In semantic segmentation, labelled images often contain multiple channels representing different classes or categories (e.g., RGB channels for different semantic labels such as road, buildings, trees). However, for training a model, it is common to convert these multi-channel labelled images into single-channel images where each pixel value represents a specific class label. Algorithm 2, titled "Convert Labelled Channels to Single-Channel Images" The algorithm begins by taking a dataset of labelled images, where each image has multiple channels corresponding to different classes. It then iterates through each image in the dataset. For each image, it combines the multiple channels

Algorithm 3 U-Net with ResNet-101 Encoder and Transformer-based Decoder

Input: Input image I, Labeled image L
Output: Semantic segmentation mask M

Encoder (ResNet-101):
Load pre-trained ResNet-101 model as the encoder
Extract multi-scale features $F = \{F_1, F_2, \ldots, F_n\}$ from the input image I
Decoder (Transformer-based):
Initialize Transformer blocks (TB) and Refinement Heads (RH)
for $i = 1$ to n **do**
    Apply Transformer block (TB) to feature map F_i from the encoder and the previous decoder layer
    Apply Refinement Head (RH) to further refine features from TB
end for
Output Layer:
Generate semantic segmentation mask M from the final decoder output
Loss Function:
Calculate Dice loss $\mathcal{L}_{\text{dice}}$ and cross-entropy loss $\mathcal{L}_{\text{ce}}$ between M and L
$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{ce}}$
Optimization:
Update model parameters using backpropagation and an optimization algorithm (e.g., Adam)
Return segmentation mask M

Algorithm 2 Convert Labeled Channels to Single-Channel Images

Input: Labeled image dataset $D = [L_1, L_2, \ldots, L_n]$ where each L_i is a labeled image with multiple channels.
Output: Single-channel labeled image dataset $D_{\text{single\_channel}} = [L'_1, L'_2, \ldots, L'_n]$ where each L'_i is a single-channel labeled image.

for $i = 1$ to n **do**
    Load labeled image L_i.
    Convert L_i to a single-channel image by combining multiple channels into one channel.
    Store the single-channel labeled image L'_i in $D_{\text{single\_channel}}$.
end for
Return the single-channel labeled image dataset $D_{\text{single\_channel}}$.

into a single channel by merging them using a specific method (e.g., taking the maximum value across channels to assign the pixel with the highest probability to a class).

This merging process condenses the information from multiple channels into a single channel, making the image suitable for training with models that expect single-channel input, such as many semantic segmentation networks. After merging the channels, the algorithm stores the resulting single-channel labeled image in a new dataset. This new dataset contains single-channel labeled images that are ready to be used for training a semantic segmentation model.

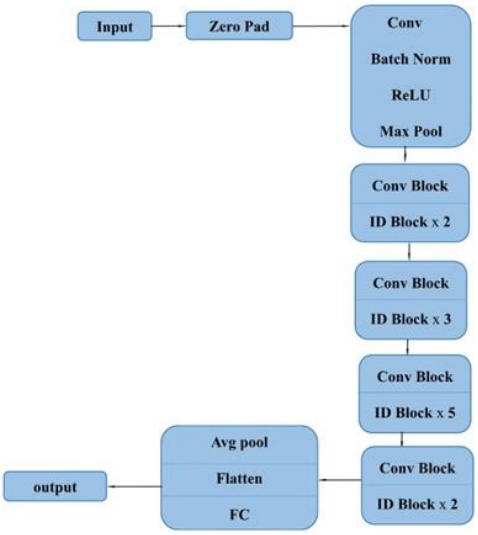


Figure 2: RseNet-101 Architecture[12]

1.3. Segmentation Model

The segmentation model architecture leverages the strengths of ResNet 101 for feature extraction, Transformer-based decoders for contextual understanding, and a multi-head loss function for effective training, resulting in accurate and robust semantic segmentation of complex urban scenes. Algorithm 3, titled "U-Net with ResNet-101 Encoder and Transformer based Decoder" contains the components as follows

1.3.1. Encoder (ResNet-101)

The encoder is based on ResNet-101, which is known for its effective ness in semantic segmentation tasks. ResNet-101 utilizes Residual Blocks (Resblocks) that enable the model to learn multi-scale features from input images. Each Resblock stage progressively down-samples the feature maps, capturing information at different levels of abstraction.

This hierarchical feature extraction is crucial for understanding the content of the images at varying levels of detail. The below Fig:2 shows the simple architecture of ResNet-101.

1.3.2. Decoder

The decoder module recovers the segmented output using features that were extracted by the encoder. It consists of Transformer Blocks specially designed to capture the global semantic contexts along with local spatial details. These blocks enhance the model's ability to interpret complex urban scenes by incorporating broad scene

semantics with detailed spatial information. Furthermore, the decoder incorporates a module called Feature Refinement Module. This will refine features further for better extraction. This stage reduces the semantic gap among various feature representations, ultimately increasing the accuracy in segmentation.

1.4. Performance Metric

The assessment of our model's performance entails the evaluation of its ac curacy through various metrics, including Overall Accuracy (OA), F1 Score, and Mean Intersection over Union (mIoU). MIoU is the average IoU computed across multiple classes or instances in a dataset. MIoU is an important metric for the evaluation of overall performance of a Semantic Segmentation model where the ability of how well the model identifies and delineates different objects or regions in the images was discovered. More the mIoU, better it is; less the mIoU, then there is scope for improvement in the model.

$$\mathrm{IoU}_i = \frac{\mathrm{Intersection}_i}{\mathrm{Union}_i}$$

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{IoU}_i$$

RESULTS

1.5. Preprocessing Results

1.5.1. Image Resizing

As the initial UAVid dataset had a much higher resolution, each individual image was each high-resolution picture—measuring either 4096 by 2160 pixels or 3840 by 2160 pixels—was padded systematically to dimensions of 1024 by 1024 pixels and then divided into eight segments, each measuring 1024 by 1024 pixels.

1.5.2. Channel Reduction

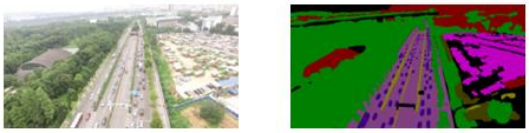
In training the model, the multiple channel images were rendered as representations consisting of a single channel (that is, a single grey-scale channel). This reduction in channels was achieved by integrating the channels into one such that each pixel in the resulting image is associated with a specific class label determined by the maximum likelihood.

1.6. Segmentation Model Results

The segmentation model was trained for the total epochs of 70 and showed progressive improvement in the performance metrics during the training time. For the evaluation of the performance of the segmentation model, accuracy and mean Intersection over Union mIoU were the primary metrics applied.

- **Accuracy**: 89%
- **mIoU**: 79.8%

(a) Image from Seq25 (b) Segmented Image for Seq25



(a) Image from Seq27 (b) Segmented Image for Seq27

The results indicate the capacity of the model to accurately classify many of the elements of an urban scene as depicted in the UAVid dataset. Training the model was done over 70 epochs; performance measures stabilized at epoch 69. Evaluation was done with mean Intersection over Union (mIoU), F1 score, and Overall Accuracy (OA). At the end of the final epoch, on the training set, the following were reached:

- **mIoU**: 0.79
- **F1-score**: 0.615
- **Overall Accuracy (OA)**: 0.896

The training mIoU stands at 0.415, which signifies the accuracy of ap proximately 79.8% of the total area across all classes by the model. This F1-score value of 0.615 reflects a balance between precision and recall during segmentation.

1.7. Class-wise Segmentation Results

- **Building:** 0.817
- **Road:** 0.642
- **Tree:** 0.741
- **Low Vegetation:** 0.781
- **Moving Car:** 0.681
- **Static Car:** 0.506
- **Clutter:** 0.495

An extensive evaluation of the effectiveness of the model on different object classes was considered. The results for each class are presented be low with their respective color code labels along with their Intersection over Union (IoU) metrics:

1.8. Overall Observations

The model had shown excellent efficiency in establishing a general range characteristics of a city, effectively covering large parts such as buildings, roads and trees, more detailed such as low vegetation, and vehicles. The elevated Intersection over Union (IoU) associated with categories like buildings (0.817), roads (0.642), and trees (0.741) reflect the model's ability to distinguish essential urban components, rendering it an invaluable instrument for intricate scene analysis comprehensiveness across various applications. Integration of a global-local framework makes the framework balance large and small elements more effectively. This hierarchical methodology improves the model's ability to Divide the major entities (such as buildings, roads) and less major elements, factors such as low vegetation and disorder. Grouping of cars—both Static and dynamic elements exemplify the model's capacity to encapsulate variability.

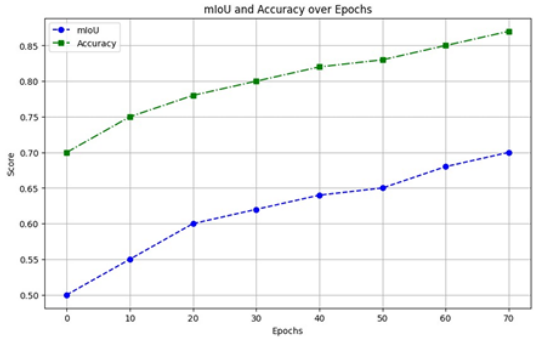


Figure 5: Graph of MIoU and Overall accuracy.

Fig 5: shows the miou and overall accuracy. These results show this global-local context framework is effective in generating precise outcomes and large semantic segmentation over a variety of object classes, for deep knowledge of complex cityscapes. More in the series, it could therefore go on to amplify its successes through further improvement of the segmentation. The process of adaptation aims to tackle increasingly complex aspects of the urban environment.

CONCLUSION AND FUTURE WORK

The current model based on U-Net architecture has proved to be very successful in capturing local details and global contextual semantic segmentation of urban environments. Nevertheless, despite robust performance, several limitations exist, particularly regarding domain shift and com plex textures. This translates into the challenges that come with the highly variable real-world conditions applied, more notably when the segmentation needs to be precise.

In future research studies, our aim will be to address these limitations by incorporating advanced learning frameworks that could feature adversarial training mechanisms. These frameworks have already shown the potential to enhance segmentation quality through boundary precision sharpening and augmenting global-local consistency.

The investigation into new techniques for handling concept drifts along with techniques for reducing reliance on large amounts of annotated training data will also be considered. Implementing these strategies will eventually lead to a better segmentation model that is adaptive and accurate, with increased generalization capabilities, making it applicable to more practical applications.

REFERENCES

[1] Li, X., Li, J. (2024). MFCA-Net: a deep learning method for Semantic Segmentation of remote sensing images. Scientific Reports, 14(1), 5745.

[2] Wang, Z., Xia, M., Weng, L., Hu, K., Lin, H. (2023). Dual encoder decoder network for land cover segmentation of remote sensing image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[3] Venugopal, P., Maddikunta, P. K. R., Gadekallu, T. R., Al-Rasheed, A., Abbas, M., Soufiene, B. O. (2023). An Adaptive DeepLabv3+ for Semantic Segmentation of Aerial Images Using Improved Golden Eagle Optimization Algorithm. IEEE Access.
[4] Wang, X., Hu, Z., Shi, S., Hou, M., Xu, L., Zhang, X. (2023). A deep learning method for optimizing Semantic Segmentation accuracy of remote sensing images based on improved UNet. Scientific reports, 13(1), 7600.
[5] Ma, X., Xu, J., Chong, Q., Ou, S., Xing, H., Ni, M. (2023). FCUnet: Refined remote sensing image segmentation method based on a fuzzy deep learning conditional random field network. IET Image Processing, 17(12), 3616-3629.
[6] Xu, L., Liu, Y., Shi, S., Zhang, H., Wang, D. (2022). Land-cover classification with high-resolution remote sensing images using Interactive segmentation. IEEE Access, 11, 6735-6747.
[7] Jin, Y., Han, D., & Ko, H. (2021). Trseg: Transformer for semantic segmentation. Pattern Recognition Letters, 148, 29-35.
[8] Chen, Z., Zhou, Y., Wang, B., Xu, X., He, N., Jin, S., Jin, S. (2022). EGDE-Net: A building change detection method for high-resolution remote sensing imagery based on edge guidance and differential enhancement. ISPRS Journal of Photogrammetry and Remote Sensing, 191, 203-222.
[9] Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet a: A deep learning framework for semantic segmentation of remotely 18 sensed data. ISPRS Journal of Photogrammetry and Remote Sensing, 162, 94-114.
[10] Gupta, A., Watson, S., &, H. (2021). Deep learning-based aerial image segmentation with open data for disaster impact assessment. Neurocomputing, 439, 22-33.
[11] Guo, Y., Nie, G., Gao, W., & Liao, M. (2023). 2d semantic segmentation: Recent developments and future directions. Future Internet, 15(6), 205.
[12] Khan, A., Khan, M. A., Javed, M. Y., Alhaisoni, M., Tariq, U., Kadry, S., ... & Nam, Y. (2022). Human Gait Recognition Using Deep Learning and Improved Ant Colony Optimization. Computers, Materials & Continua, 70(2).
[13] Hasan, K. R., Tuli, A. B., Khan, M. A. M., Kee, S. H., Samad, M. A., Nahid, A. A. (2023). Deep learning-based semantic segmentation for remote sensing: A bibliometric literature review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.
[14] Yang, N., Tang, H. (2021). Semantic segmentation of satellite images: A deep learning approach integrated with geospatial hash codes. Remote Sensing, 13(14), 2723.
[15] He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y. (2022). Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-15.