Volume 23, Issue 2: KUS: 1629: xx-xx: 13052026: 2026

KHULNA UNIVERSITY
STUDIES

journal homepage: www.kus.ku.ac.bd

Research article

Shadow-Resilient Greenery Index Estimation Using Deep Learning

Ashik Mahmud Raz, Md Anamul Islam, Azmain Inquaid Haque, Aminul Islam and G M Atiqur Rahaman

Computer Science and Engineering Discipline, Khulna University, Khulna-9208, Bangladesh

ABSTRACT

The Green Area Factor (GAF) is widely used to evaluate urban ecological quality, but conventional field-based assessment is time-consuming, labor-intensive, and difficult to apply over large areas. Although image-based methods offer a more scalable alternative, accurately identifying vegetation remains challenging because shadows cast by buildings, trees, and other urban structures often obscure green surfaces and lead to segmentation errors. To address this challenge, this study proposes a shadow-resilient deep learning framework for estimating the Greenery Index (GI) from aerial imagery as a scalable proxy for GAF. The framework combines InternImage-T with UPerNet and ResNet-34 with UNet++ using a class-wise maximum-confidence ensemble strategy. Instead of averaging predictions from multiple models, each semantic class is assigned to the model that demonstrates the highest validation performance, allowing the complementary strengths of both architectures to be fully utilized while reducing shadow-induced misclassification and improving boundary localization. The proposed framework was evaluated on two independent aerial imagery datasets acquired from satellite and UAV platforms. It achieved mIoU values of 0.88 and 0.82, F1-scores of 0.93 and 0.90, and pixel accuracies of 0.97 and 0.96. Compared with a previously published weighted-ensemble method evaluated on the same benchmark dataset, the proposed approach improved mean IoU and F1-score by six and eleven percentage points, respectively. In addition, the estimated Greenery Index showed a strong correlation with field-measured GAF values (Pearson correlation coefficient, r = 0.953) while reducing the mean absolute error from 0.140 to 0.087. These findings demonstrate that improving segmentation in shadowed regions leads to more reliable ecological indicators and provides an effective, scalable solution for urban greenery assessment, environmental monitoring, and sustainable urban planning.

ARTICLE INFO

Article timeline:
Date of Submission:
13 May, 2026
Date of Acceptance:
02 August, 2026
Article available online:
18 August, 2026

***Keywords*:**
Green Area Factor (GAF)
Semantic segmentation
Urban greenery
Satellite imagery
UAV imagery
Shadow aware learning
Deep learning
Urban sustainability

Introduction

City parks, gardens, and greenways are an important part of any modern city's sustainable design and livability. In urban areas, vegetation contributes to regulating surface temperature, air pollution control, stormwater management, biodiversity conservation, and physical and mental health. As urbanization continues to increase density, the natural vegetative cover is being replaced by impervious paving. This change reduces ecological resilience by creating more stressors on the environment. Therefore, an accurate evaluation of the amount of vegetation within urban areas is now essential for successful urban design and climate adaptability.

With the advancements made in aerial imaging and remote sensing technology, researchers have now been afforded the ability to analyze land cover at city-scale resolution. However, even with these technological advancements, detecting vegetation in urban areas continues to be a significant challenge due to the nature of urban scenes, because they are heterogeneous, have irregular spatial arrangements, different types of surface materials, and an overall complex structure. One of the key challenges in detecting vegetation within urban areas is that the shadow created from structures (such as buildings, trees, and built infrastructure) makes it difficult to identify green areas in the landscape because vegetative surfaces that are covered by shadow will have distorted spectral and textural characteristics, thus making it difficult to

*Corresponding author: <ashik110353@gmail.com>

DOI: https://doi.org/10.53808/KUS.2026.23.02.1629-se

distinguish them from non-vegetative surfaces. This is why vegetation coverage in urban areas with a high density of vertical development is often underestimated by conventional segmentation and classification methods (Dimitrovski et al., 2024; Rahaman et al., 2024; Z. Wang et al., 2023; Wu & Zhang, 2024).

Errors in vegetation mapping can propagate into urban ecological indicators, potentially affecting environmental assessment, land-use planning, and policy decisions. One of the most commonly used indicators, the Green Area Factor (GAF), evaluates the ecological value of developed urban areas by considering the various kinds of green and permeable surfaces (Rahaman et al., 2024). The GAF is a measure of ecological value in urban areas, as well as a tool for planning and designing green infrastructure. The GAF is calculated (Rahaman et al., 2024) as:

$$GAF = \frac{\sum_{i=1}^{K} A_i F_i}{\sum_{i=1}^{K} A_i} \tag{1}$$

where i represents each landscape object, K is the total number of object classes, A_i represents the area of each landscape object, and F_i reflects the contribution of each object to urban greenery. This point-based metric provides a clear and practical indicator that helps planners, architects, and ecologists make informed decisions for sustainable urban development.

The GAF method is commonly used in urban planning; it provides a consistent means of assessing the extent to which vegetation elements (trees, lawns, gardens, green roofs, permeable pavements) are integrated into new developments. GAF provides a way to objectively assess the ecological balance at the site level by assigning various weighted values to the various green components and establishing a minimum threshold for how many of these green components must be used in new developments. GAF directly contributes to the environmental sustainability of urban areas by providing a means to decrease the amount of heat generated in cities through urban heat reduction, improve the infiltration of rainwater, and promote better air quality. In addition, the wide variety of types of vegetation used in cities contributes to the biodiversity within urban areas; this is especially important in densely populated urban areas where there are few areas of natural land.

Making the structure green is not just helpful for the environment but also for social and economic well-being. Urban environments that are more livable and healthier boost visual quality while enhancing physical health. Effective use of GAF helps planners, designers, architects, and policymakers communicate better. This also helps to assess tradeoffs in a transparent manner and maintain long-term sustainability and climate-resilience goals.

The presence of high-rise buildings, transport infrastructures and tree canopy leads to shadows being a geographically persistent phenomenon in cities. The image signatures of shadows differ from the background and the spectrally and spatially homogeneous regions of interest. The spectral responses and spatial textures have a major influence on shadows. Thus, most of the green space in the shadow is usually mixed and looks like non-green or impermeable surfaces leading to systematic underestimation of the green space. Many photogrammetric studies have shown great potential in urban green space classification based on high-resolution remote sensing images and machine learning techniques and advanced feature extraction strategies (Dimitrovski et al., 2024; Rahaman et al., 2024; Z. Wang et al., 2023; Wu & Zhang, 2024).

Therefore, shadow-induced misclassification is not merely a technical limitation of image analysis; it can also affect the reliability of urban ecological assessments. When vegetation within shadowed areas is incorrectly classified as non-vegetated surfaces, the resulting vegetation estimates may be systematically underestimated. For example, shadow misclassification dilutes urban eco-indicators, resulting in ineffective checks of environmental impact and biased land-use planning, climate adaptation, and green infrastructure decisions. In other words, the planning may not reflect the real ecological situation when green spaces are consistently undervalued. As a consequence, those interventions might be less successful or even unsustainable.

Robust mapping of urban vegetation from an aerial frame requires the explicit modeling of the shadow effect. We propose a shadow-aware semantic segmentation framework to improve vegetation mapping in complex urban aerial imagery. This study proposes a shadow-resilient semantic segmentation framework designed to improve vegetation detection in both illuminated and shadowed regions of urban aerial imagery.

This research mainly aimed at creating a segmentation framework useful for detecting vegetation in shadowed urban locations. This study aims to make the computation of the Greenery Index more accurate by reducing misclassification due to shadows. More accurate vegetation mapping can support reliable greenery assessment and provide useful information for urban environmental monitoring, sustainable planning, and green infrastructure management. This paper aims at shadow detection and removal from remote sensing images for the development of green areas.

This study expects to ensure much more accurate and consistent urban vegetation maps. With the data generated from this research, we would obtain a more reliable Green Area Factor calculation. By improving the identification of vegetation under challenging illumination conditions, the proposed framework is expected to produce more accurate and consistent vegetation maps, which can subsequently support more reliable Greenery Index estimation.

Literature Review

Semantic segmentation of remote sensing imagery has become an active research area because of its importance in urban planning, environmental monitoring, precision agriculture, and land-use analysis. Compared with traditional pixel-based classification methods, deep learning techniques learn hierarchical image features directly from data and have significantly improved segmentation accuracy. Despite these advances, accurately mapping urban vegetation remains challenging because aerial scenes contain complex spatial structures, diverse land-cover types, and large variations in illumination. Among these challenges, shadows cast by buildings, trees, and other urban structures frequently obscure vegetation and reduce the reliability of segmentation results.

Early semantic segmentation methods were primarily based on convolutional neural networks (CNNs). The introduction of U-Net by (Ronneberger et al., 2015) marked a major breakthrough through its encoder-decoder architecture with skip connections, enabling the recovery of fine spatial details while preserving semantic information. U-Net has since been adapted extensively for remote sensing applications because of its simple architecture and reliable performance (Dimitrovski et al., 2024; Khan & Jung, 2024; Ma et al., 2022; Panghurian et al., 2024; Xu et al., 2024). Subsequent models further improved feature extraction and contextual understanding. DeepLab introduced atrous convolution for multi-scale feature learning (Chen et al., 2018), while DeepLabV3+ refined boundary localization using an improved decoder (Zhou & Li, 2022). PSPNet incorporated pyramid pooling to capture global contextual information (Zhao et al., 2017), and UPerNet extended this concept by combining Pyramid Pooling Modules with Feature Pyramid Networks for multi-scale scene parsing (Xiao et al., 2018). Other studies incorporated residual learning and attention mechanisms, including ResUNet-a (Diakogiannis et al., 2020) and AS-UNet++ (Nan et al., 2024), to improve feature representation and computational efficiency. HRNet (Wang et al., 2020) adopted a different strategy by maintaining high-resolution feature representations throughout the network, leading to improved localization accuracy.

Although CNN-based models achieve strong segmentation performance, they rely mainly on local receptive fields. This limits their ability to model long-range spatial relationships, particularly in complex urban environments where vegetation appears fragmented or partially occluded. Increasing network depth or adopting stronger backbones such as ResNet-34 (Nsiah et al., 2023; Panghurian et al., 2024; Rahaman et al., 2024) improves feature extraction but does not completely resolve this limitation. Ensemble learning has also been explored to combine complementary models and improve overall segmentation performance (Wang et al., 2023; Rahaman et al., 2024). However, these approaches generally optimize overall segmentation accuracy rather than explicitly addressing the uncertainty introduced by shadowed vegetation.

The introduction of Vision Transformers (ViTs) addressed many limitations of CNNs by modeling long-range spatial dependencies through self-attention mechanisms (Liu et al., 2021). Transformer-based architectures such as SegFormer (Xie et al., 2021), Swin-UPerNet (Wang et al., 2023), and several hybrid CNN-transformer frameworks (Li et al., 2022; He & Ma, 2023; Xu et al., 2024; Shi & Luo, 2024) have demonstrated superior performance in remote sensing semantic segmentation. By capturing both local details and global context, these models better distinguish visually similar land-cover classes. Nevertheless, transformer-based models generally require larger training datasets, higher computational resources, and their performance may decline when applied to imagery acquired under different environmental conditions or sensor configurations.

More recently, vision foundation models have attracted considerable attention for dense prediction tasks. InternImage (Wang et al., 2023) replaces conventional convolutions with deformable convolutions (DCNv3), allowing receptive fields to adapt dynamically according to image content. This adaptive feature extraction is particularly beneficial for modeling irregular objects such as tree canopies and fragmented vegetation. InternImage has shown competitive performance in semantic segmentation, change detection, and other computer vision tasks (Shen et al., 2024; Santosa & Kusuma, 2025; Yu et al., 2024). However, most existing studies evaluate these models under normal imaging conditions, while their robustness to heavy urban shadows remains largely unexplored.

Besides model architecture, image quality also influences segmentation accuracy. Urban aerial images frequently contain shadows, low contrast, atmospheric haze, sensor noise, and compression artifacts. These factors reduce the separability between vegetation and surrounding non-vegetation classes. Under such conditions, CNN-based methods may fail to recover weak textures, whereas transformer-based models can propagate uncertainty across large attention regions. Consequently, vegetation located within shadowed areas is often misclassified as roads, buildings, or other impervious surfaces, leading to systematic underestimation of urban greenery.

Another limitation of the current literature is the heavy reliance on overall accuracy metrics. Most studies report mean Intersection over Union (mIoU) or F1-score averaged over the entire dataset, which may conceal localized failures in shadowed regions. Furthermore, many methods are validated using data collected from a single city or imaging platform, making it difficult to assess their generalization across different urban environments. As a result, the reported performance may not fully reflect the challenges encountered in practical urban vegetation mapping.

Table 1 summarizes representative semantic segmentation approaches for remote sensing imagery. Although recent methods have substantially improved segmentation accuracy, few explicitly address shadow interference or evaluate robustness across multiple aerial datasets. This limitation highlights the need for segmentation frameworks that remain reliable under varying illumination conditions while maintaining strong generalization across different acquisition platforms.

The limitations identified above motivate the framework proposed in this study. Instead of treating shadow removal as a separate preprocessing task, the proposed method directly performs shadow-resilient semantic segmentation by combining the complementary strengths of InternImage-T with UPerNet and ResNet-34 with UNet++. A class-wise maximum-confidence ensemble strategy is introduced to preserve the strongest prediction for each land-cover class, improving vegetation detection under varying illumination while producing more reliable Greenery Index estimation across different aerial datasets.

Table 1: Comparison of Representative Semantic Segmentation Methods for Remote Sensing Imagery

Study	Model	Main Strength	Shadow-aware	Main Limitation
Ronneberger et al. (2015)	U-Net	Encoder-decoder with skip connections	No	Limited global context
Chen et al. (2018)	DeepLab	Atrous convolution and multi-scale features	No	Sensitive to illumination variation
Xiao et al. (2018)	UPerNet	Pyramid pooling with feature pyramid network	Partial	Shadow robustness not evaluated
Xie et al. (2021)	SegFormer	Efficient transformer encoder	Partial	Requires large training datasets
Wang et al. (2023)	InternImage	Adaptive deformable convolution (DCNv3)	Partial	Limited evaluation on shadow-heavy aerial imagery
Rahaman et al. (2024)	Weighted Ensemble	Improved segmentation through model fusion	Partial	Uses global weighted averaging rather than class-specific model selection
Proposed Method	InternImage-T with UPerNet and ResNet-34 with UNet++	Class-wise maximum-confidence ensemble for shadow-resilient vegetation segmentation	Yes	Performance may still decrease in extremely dense shadows or severe occlusion

Methodology

This study presents a novel and efficient deep learning framework (Figure 2) to accurately segment urban vegetation and reliably assess Greenery Index (GI) using high-resolution aerial imagery. The major goal of the research is to target the misclassification of vegetation under shaded illumination. The complete system includes preparing the data, shadow-aware data augmentation, semantic segmentation using dual-model learning, integration/refinement of ensemble and Greenery Index estimation. The framework deals with shadow interference in an adaptive and progressive manner with complementary modeling, making it more resilient to varying illumination. As per the results of our experiments, the framework consistently and stably segments vegetation without destroying the structures. The proposed framework is designed to reduce shadow-induced vegetation misclassification, thereby supporting more reliable Greenery Index estimation.

Dataset

The proposed framework was evaluated using two aerial imagery datasets acquired from different imaging platforms and geographic regions to assess its robustness under diverse urban environments and illumination conditions. The first dataset is the Örebro Municipality benchmark dataset (Rahaman et al., 2024), which contains high-resolution satellite orthophotos with pixel-level annotations for multiple urban land-cover classes. The second is a custom UAV dataset collected over the Khulna University campus, Bangladesh, using a DJI Mini 4 Pro drone. All UAV images were manually annotated at the pixel level following a consistent semantic labeling protocol. Although the datasets differ in acquisition platform and geographic characteristics, both include common urban land-cover classes such as vegetation, buildings, roads, open spaces, and shadowed regions. Evaluating the proposed framework on both datasets provides a comprehensive assessment of its robustness and generalization capability. The preprocessing and dataset partitioning procedures are described in the following subsection.

Örebro Municipality Dataset

The Örebro dataset (Rahaman et al., 2024) is a research dataset and comprises high-resolution aerial orthophotos with pixel-level annotations for several urban land cover classes including vegetation, buildings, roads and shadows. Objects that are shadowed are coded as a separate semantic class as are areas of pure shadow. This method enables explicit learning of the illumination-invariant features of models, especially for aerial imagery that is sensitive to illumination changes. This dataset can serve as a baseline in order to assess the segmentation performance under standardized conditions, allowing for comparison of the results with other research works.

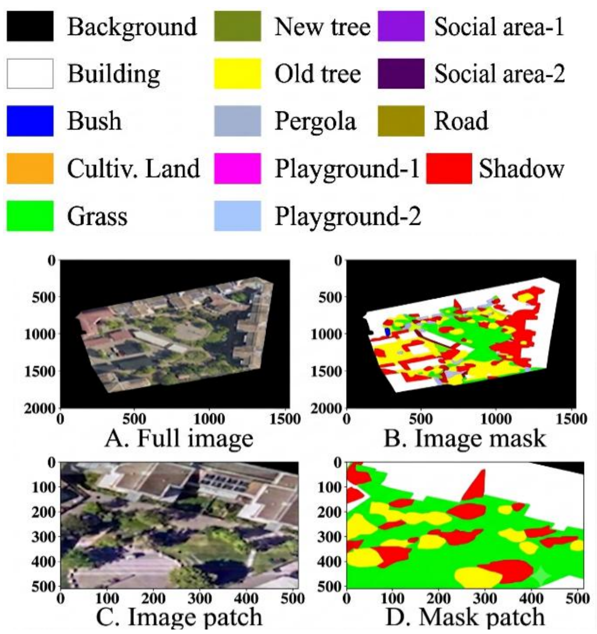


Figure 1: Representative Sample from the Örebro Municipality Dataset, Showing the Land-cover Color Palette, the Original Aerial Image, and the Corresponding Ground-truth Label Mask

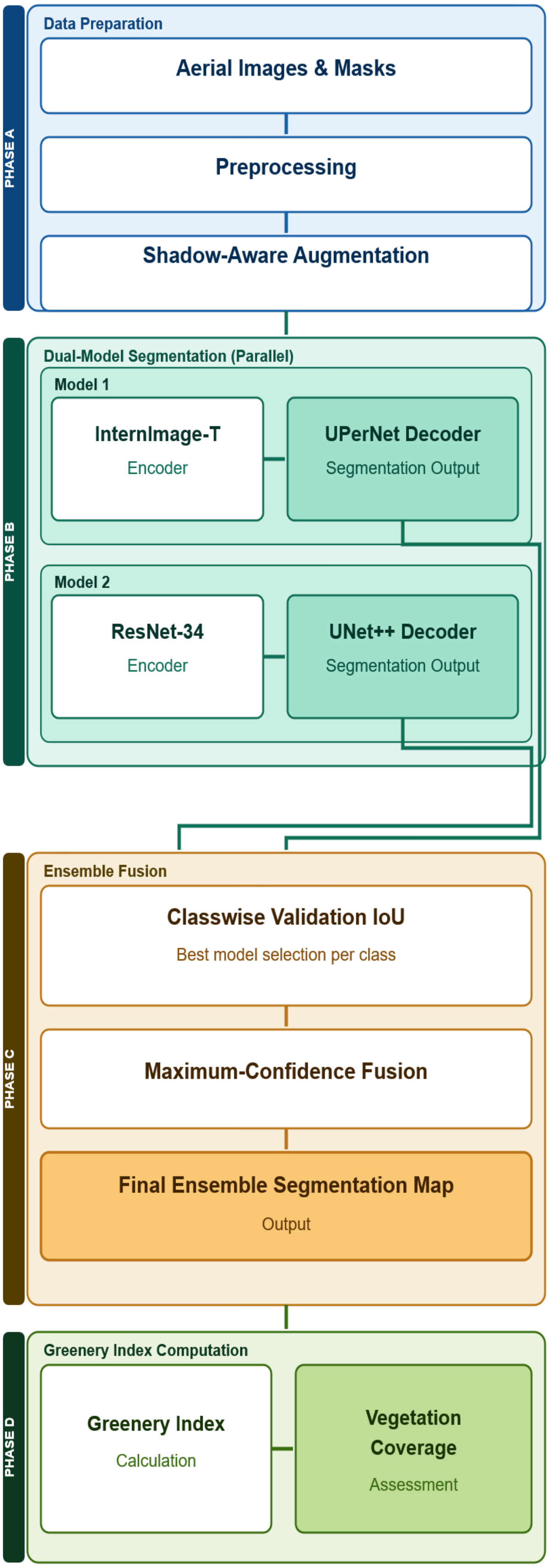


Figure 2: Proposed Shadow-resilient Ensemble Framework for Vegetation Segmentation and Greenery Index Computation

Khulna University Aerial Imagery Dataset

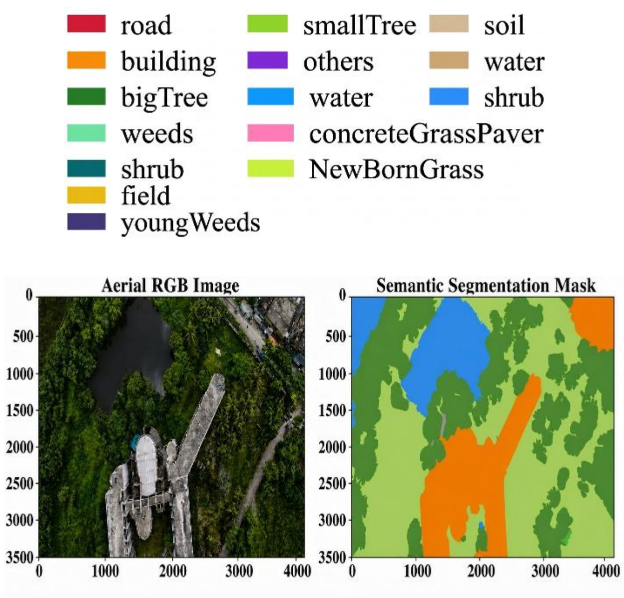
The Khulna University dataset was captured by using a DJI Mini 4 Pro drone (Figure 3). Multiple altitudes of flight were used to obtain both high-resolution images and wide area views. Low level flights focus and highlight the small structures, narrow roads and individual trees, and higher elevation flights the overall contexts and land use patterns. This multi-altitude strategy increases diversity of data and robustness of the model.

Figure 3: DJI Mini 4 Pro Drone Used for Aerial Acquisition

Annotation

Semantic consistency was achieved by manually annotating all images at the pixel-level. Shadowed regions were not included in a shadow class as was done for the Örebro dataset, but were coded to the correct land-cover class (e.g. Vegetation under shadow is recorded as vegetation). This pixel-wise semantic annotation strategy supports the estimation of area-based greenery metrics, including the Greenery Index, based on practical interpretation of urban land-cover conditions.

High-resolution images and masks were cropped into 256×256-pixel patches to normalize inputs to the models and to increase the size of this dataset. Firstly, the non-overlapping tiling was used to create the base patches, then the shifting tiling with stride 128 pixels was used to cover more space and to improve the learning of the boundaries. From this process, 720 image-mask pairs were created for training, validation, and testing; full-scale images and cropped patches of detail.



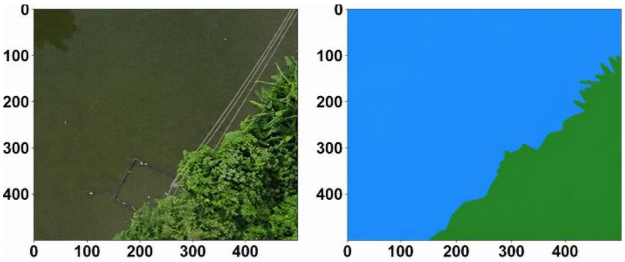


Figure 4: Representative Land-cover Color Palette, Original Aerial Imagery, and Corresponding Ground-Truth Semantic Label Masks from the Khulna University Dataset

Dataset Usage Strategy

The datasets have been utilized in separate models to test the stability of the models under different scenarios. The Örebro dataset was used in the first stage for the baseline and to verify the performance on a known test, and the Khulna University dataset was used in the second stage for evaluation of the model's performance in a local, shadow-rich urban environment. To demonstrate the testing of this model with both datasets, the evaluation strategy is shown in Figure 5. In future work, the two datasets could be combined under a common class definition to investigate the generalization capabilities of the models for mixed and diverse data distributions.

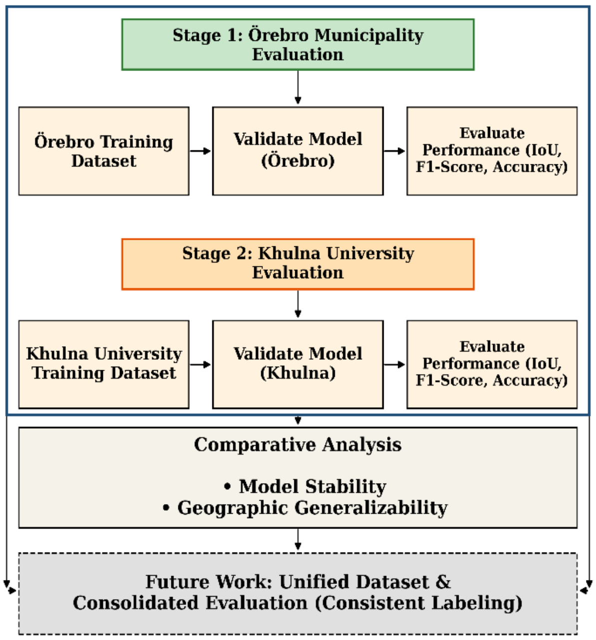


Figure 5: Diagram Showing Evaluation Strategy across Datasets

In this chapter, the two datasets of aerial images used in this study were introduced. The first set of images from Örebro are used to validate the algorithm in standard conditions, and the second set of images from Khulna University are urban scenes with shadows and localized variations. The proposed framework can be evaluated both under shadow-aware and object-centric scenarios due to different shadow annotation strategies. The datasets are combined to create a robust and comprehensive basis for evaluating the robustness and applicability of the proposed semantic segmentation model.

Data Preparation

The Örebro Municipality and Khulna University datasets were acquired using different imaging platforms, namely satellite orthophotos and UAV imagery. Consequently, the datasets differ in spatial resolution, image characteristics, and ground sampling distance. To provide consistent inputs for the proposed framework, all aerial images and their corresponding ground-truth masks were first resized to a common spatial scale before further processing. Standardizing the image scale reduces discrepancies between datasets and enables the segmentation models to learn more consistent spatial representations across different acquisition platforms.

After resizing, the images and masks were divided into the 256 × 256 pixel patches. This patch size was selected as a compromise between computational efficiency and spatial context. Larger patches provide richer contextual information but require substantially more GPU memory, whereas smaller patches may lose important neighborhood information needed to distinguish visually similar land-cover classes. A resolution of 256 × 256 preserves sufficient local context for vegetation mapping while remaining computationally efficient for training both segmentation networks.

To reduce boundary information loss during patch extraction, an overlapping tiling strategy was employed using a stride of 128 pixels, resulting in a 50% overlap between adjacent patches. Without overlap, important objects located near patch borders such as tree canopies intersecting building shadows may be split across multiple patches, making them difficult for the network to learn effectively. The overlapping strategy ensures that these boundary regions are fully represented in at least one training sample while also increasing the diversity of training data without requiring additional image acquisition.

The prepared datasets were then partitioned according to their respective experimental protocols. For the Örebro Municipality dataset, 75% of the images were used for training and 25% for validation. For the Khulna University dataset, 75% of the images were allocated for training, 20% for validation, and the remaining 5% for testing. Stratified sampling was adopted during data partitioning to preserve the distribution of land-cover classes across all subsets. This is particularly important because several classes occupy relatively small image regions compared with dominant classes such as buildings or grass. Maintaining similar class distributions in each subset helps reduce sampling bias and provides a more reliable evaluation of segmentation performance.

Before training, the RGB images were normalized using the ImageNet mean and standard deviation,

$$\mu = [0.485, 0.456, 0.406], \sigma = [0.229, 0.224, 0.225].$$

Both InternImage-T and ResNet-34 were initialized with ImageNet-pretrained weights; therefore, applying the same normalization aligns the input distribution with that used during pretraining. This improves optimization stability, accelerates convergence, and enables more effective transfer learning. In contrast, the ground-truth segmentation masks were not normalized. They were preserved as integer-valued class labels and resized using nearest-neighbor interpolation to maintain the integrity of class boundaries and prevent the creation of invalid intermediate label values.

Shadow-Aware Data Augmentation

Urban aerial imagery is affected by variations in illumination, seasonal changes, and shadows cast by buildings, trees, and other urban structures. These factors alter the appearance of vegetation and can reduce the generalization ability of semantic segmentation models. To improve robustness under diverse imaging conditions, a set of targeted data augmentation techniques was applied during training.

Geometric transformations, including horizontal flipping, vertical flipping, and random rotation, were used to increase the diversity of training samples and improve the model's invariance to viewing orientation. Photometric distortion was applied to simulate natural variations in brightness, contrast, and color caused by differences in illumination, sensor characteristics, and image acquisition conditions.

To improve feature discrimination in shadowed regions, Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied. CLAHE enhances local image contrast while limiting noise amplification, making vegetation boundaries and fine structural details more distinguishable under low-illumination conditions.

To further address the central challenge of this study, synthetic shadow augmentation was incorporated during training. This transformation generates realistic shadow patterns with varying shapes and intensities, exposing the model to a wider range of illumination conditions than those present in the original dataset. As a result, the network learns to identify vegetation from its structural and textural characteristics rather than relying primarily on pixel brightness. This improves the model's robustness to shadow-induced appearance changes and supports more reliable Greenery Index estimation across different urban environments.

Dual-Model Semantic Segmentation Framework

To improve segmentation reliability, a dual-model architecture was adopted. Instead of relying on a single network, two complementary segmentation models were trained independently:

- **Model 1:** InternImage-T encoder with a UPerNet decoder. InternImage leverages deformable convolutions (DCNv3) to capture complex spatial patterns, making it effective in heterogeneous urban scenes. UPerNet integrates Pyramid Pooling Modules (PPM) and Feature Pyramid Networks (FPN) for multi-scale feature aggregation, improving contextual understanding.

- **Model 2:** ResNet-34 encoder with a UNet++ decoder. ResNet-34 provides stable residual feature extraction, while UNet++ refines boundary delineation through nested skip connections. This configuration enhances fine-grained segmentation, particularly along vegetation edges partially obscured by shadows.

Both models were initialized with pretrained weights and optimized using a hybrid loss that combines pixel-level classification accuracy with region-level overlap.

Maximum-Confidence Ensemble Strategy

High-resolution aerial imagery presents several challenges for semantic segmentation, including large variations in object scale, class imbalance, complex urban structures, and shadow interference. Because different network architectures learn different feature representations, no single model performs equally well across all land-cover classes. During validation, InternImage-T with UPerNet consistently achieved better performance on large, homogeneous regions by capturing rich semantic context, whereas ResNet-34 with UNet++ produced more accurate predictions around fine structures and object boundaries. This complementary behavior motivated the development of the proposed ensemble strategy.

Instead of averaging the predictions from multiple models, we employ a class-wise maximum-confidence ensemble strategy that assigns each semantic class to the model demonstrating the highest validation reliability. Conventional weighted ensembles combine class probabilities using fixed or learned weights. Although effective in many applications, probability averaging can suppress highly confident predictions when the models disagree, particularly around vegetation boundaries and shadow-affected regions. This often leads to blurred object boundaries and reduced segmentation accuracy for small vegetation patches and narrow urban green structures.

To overcome this limitation, the proposed framework performs class-wise model selection rather than probability averaging. Each semantic class is assigned to the model that achieves the highest validation Intersection over Union (IoU), allowing both architectures to contribute according to their individual strengths while preserving confident predictions.

Class-Wise Model Selection

During validation, the IoU is computed independently for each semantic class c and each model m. The preferred model for class c is selected as:

$$m_c = \arg\max_{m \in \{M_1, M_2\}} \mathrm{IoU}_{\mathrm{val}}\,(m, c) \tag{2}$$

where:

- M_1 represents InternImage-T with UPerNet,

- M_2 represents ResNet-34 with UNet++,

- IoU_{val} (m, c) denotes the validation IoU for class c using model m.

This formulation ensures that each class is assigned to the architecture that empirically demonstrates superior predictive reliability.

Inference-Time Fusion

During inference, both models generate pixel-wise prediction maps.

$$\hat{y}_{M_1}(x,y) \text{ and } \hat{y}_{M_2}(x,y) \tag{3}$$

Where (x,y) denotes the pixel location in the input image. The final ensemble prediction $\hat{y}(x,y)$ is defined as:

$$\hat{y}(x,y) = \begin{cases} \hat{y}_{(M_1)(x,y)}, if\ m_{\hat{y}_{(M_1)(x,y)}} = M_1 \\ \hat{y}_{(M_2)(x,y)}, otherwise \end{cases} \tag{4}$$

The final prediction for each class is selected from the model that achieves the highest validation IoU.

Unlike conventional weighted fusion, the proposed strategy does not average class probabilities. Instead, it preserves the strengths of each individual architecture by allowing every semantic class to be predicted by the model that performs best for that class. InternImage-T with UPerNet contributes stronger global contextual understanding for large vegetation regions, while ResNet-34 with UNet++ provides finer boundary localization for smaller or structurally complex objects. By selecting predictions according to class-wise validation performance, the proposed ensemble reduces shadow-induced misclassification, preserves boundary details, and produces more reliable vegetation maps for subsequent Greenery Index estimation.

Greenery Index Computation

Following segmentation, the Greenery Index (GI) was computed to quantify vegetation coverage. The GI was calculated as the weighted ratio of vegetated area to total land area within a defined spatial unit:

$$GI = \frac{\sum(A_i \times F_i)}{\sum A_i} \tag{5}$$

where A_i represents the area of vegetation class i and F_i denotes its assigned ecological weighting factor.

Both overall GI and grid-based zonal GI were computed to enable spatially detailed ecological assessment. This allows identification of vegetation-deficient zones and supports data-driven urban planning decisions.

Loss Function

While training the model for multi-class semantic segmentation, the hybrid loss is used. The optimum prediction is achieved via a hybrid loss function in semantic segmentation. Essentially, what the hybrid loss does is optimize the similarity between the ground-truth and predicted pixel-wise segmentation masks. In aerial image semantic segmentation, the use of single loss hampers the performance of the model due to the class imbalance problem.

To overcome this issue, a hybrid loss can be created by combining Categorial Cross-Entropy and Dice loss functions. This helps the model learn pixel level & region level information simultaneously. The total loss for training is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\mathcal{CE}} + \mathcal{L}_{\mathrm{Dc}} \tag{6}$$

Cross-Entropy Loss

The Categorical Cross-Entropy loss focuses on pixel-level classification by penalizing incorrect predictions at each pixel location. It is defined as:

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{c=1}^{C} y_c \log(p_c) \tag{7}$$

where C denotes the total number of semantic classes, y_c represents the ground-truth label (one-hot encoded), and p_c is the predicted probability for class c, obtained through the softmax function.

This loss ensures that each pixel is assigned to the correct class. However, it treats all pixels equally and may not adequately handle imbalanced datasets where some classes dominate.

Dice Loss

To enhance segmentation performance at the region level, Dice Loss is incorporated. It is derived from the Dice Similarity Coefficient and directly measures the overlap between predicted and ground-truth segmentation masks. Dice Loss is defined as:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum(P \cdot T) + \epsilon}{\sum P + \sum T + \epsilon} \tag{8}$$

Where P denotes the predicted probability map, T represents the one-hot encoded ground-truth mask, and ϵ is a small constant introduced to ensure numerical stability and avoid division by zero.

Unlike Cross-Entropy, Dice Loss emphasizes region-level agreement, making it particularly effective for handling class imbalance and improving segmentation of smaller or less frequent classes. By integrating Cross-Entropy and Dice Loss, the model benefits from both detailed pixel-level supervision and global structural alignment. Cross-Entropy guides accurate classification, while Dice Loss enforces spatial coherence and improves boundary delineation. As a result, this hybrid loss function produces more accurate, smooth, and consistent segmentation outputs. It is especially effective in datasets with imbalanced class distributions and complex object boundaries, where both local precision and global consistency are critical.

Implementation Details

The proposed framework was implemented using PyTorch together with the Segmentation Models PyTorch (SMP) library. Both segmentation networks were initialized with ImageNet-pretrained backbone weights to accelerate convergence and improve feature representation. The InternImage-T encoder was paired with the UPerNet decoder, while the second model employed a ResNet-34 encoder with a UNet++ decoder.

All input images were resized to 256 × 256 pixels and normalized using the ImageNet mean and standard deviation before training. Data augmentation, including horizontal and vertical flipping, rotation, CLAHE enhancement, and synthetic shadow generation, was applied only to the training set to improve model robustness under varying illumination conditions.

Model optimization was performed using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a batch size of 6. Early stopping based on the validation mean Intersection over Union (mIoU) was employed to prevent overfitting, and the model checkpoint with the highest validation mIoU was retained for testing.

The InternImage-T with UPerNet model was trained locally on Windows Subsystem for Linux 2 (Ubuntu 22.04) using Visual Studio Code, whereas the ResNet-34 with UNet++ model was trained on Google Colab equipped with an NVIDIA Tesla T4 GPU (16 GB VRAM). Training required approximately 2 h 30 min for the Örebro dataset and 2 h 35 min for the Khulna University dataset.

Performance was evaluated using Intersection over Union (IoU), mean IoU (mIoU), Precision, Recall, F1-score, Pixel Accuracy (PA), and Mean Class Accuracy (mAcc). Using multiple evaluation metrics provides a more comprehensive assessment of segmentation quality than relying on a single performance measure.

Experimental Results and Discussion

Experimental Setup

Experiments were conducted using two complementary semantic segmentation models to exploit their respective strengths in capturing global contextual information and fine spatial details. The first model combined InternImage-T with UPerNet, while the second employed ResNet-34 with UNet++. To improve segmentation robustness, the outputs of both models were integrated using the proposed class-wise maximum-confidence ensemble strategy. Unlike conventional weighted averaging, the proposed approach assigns each semantic class to the model that demonstrated higher validation IoU for that class, thereby preserving class-specific strengths while reducing prediction uncertainty near object boundaries.

The proposed framework was evaluated on two independent aerial imagery datasets. The Örebro Municipality benchmark dataset was used to assess performance under standardized benchmark conditions, whereas the Khulna University UAV dataset was used to evaluate the framework in a real-world urban environment containing complex shadows and heterogeneous land-cover classes. For the Örebro dataset, 75% of the images were used for training and 25% for validation following the protocol adopted in the original study. For the Khulna University dataset, a stratified split of 75% training, 20% validation, and 5% testing was employed to preserve class distributions across all subsets.

All input images were normalized using the ImageNet mean and standard deviation to ensure compatibility with the pretrained backbone networks. During inference, segmentation performance was evaluated using per-class IoU, mean IoU (mIoU), Precision, Recall, F1-score, Pixel Accuracy (PA), and Mean Class Accuracy (mAcc). These complementary metrics provide a comprehensive evaluation of both overall segmentation performance and class-wise prediction quality.

Training and Validation Performance

The Örebro Municipality Dataset's training behavior for both the models is illustrated in Figure 6 and Figure 7.

Figure 6 indicates InternImage-T with UPerNet demonstrates smooth and stable convergence throughout training. There is a minimal gap between the training and validation accuracy curves denoting their closeness. This shows effective generalization and low overfitting. Stable characteristics in aerial imagery is crucial as illumination and shadow cause distribution shift often.

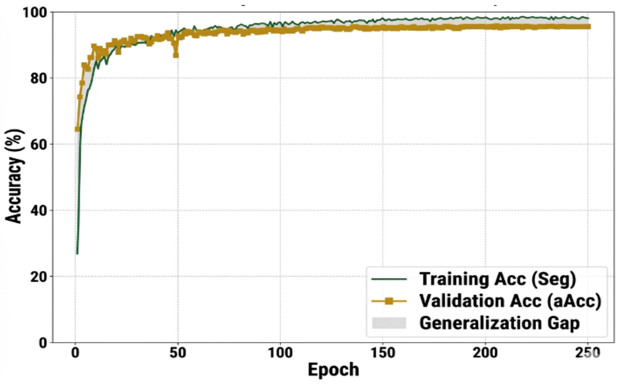


Figure 6: The Accuracy Curves for Training and Validation of InternImage-T and UPerNet

Similarly, the proposed UNet++ with ResNet-34 as a backbone detailed in Figure 7 shows that model learns meaningful spatial features in the first few epochs rapidly, after which the learning slowly stabilises.

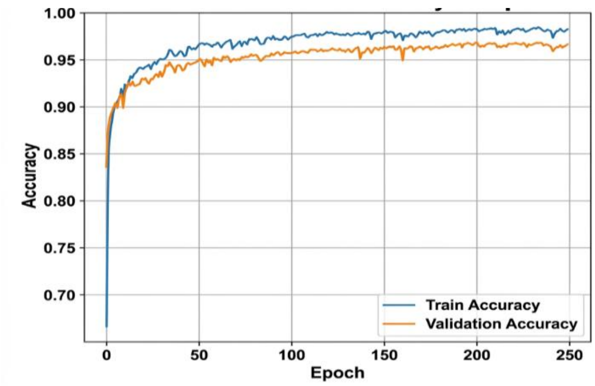


Figure 7: The Accuracy Curves for Training and Validation of UNet++ and ResNet-34

As previously depicted, the training and validation accuracy differ from each other but with a small gap, which is expected and is an indicator of balanced learning in the model.

Qualitative Results

As shown in Figure 8, the predicted masks closely follow the ground truth boundary, including in the shadow areas. Accurate segmentation of shadowed vegetation and road areas indicates that the proposed model learns robust spatial and contextual features beyond simple intensity cues.

Table 2: Confusion Matrix (0–100 %) of Proposed Maximum-Confidence Ensemble for Örebro Municipality Dataset

SPred \ True	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	100	00	00	00	00	00	00	00	00	00	00	00	00	00
1	00	95	02	00	00	01	00	00	00	00	00	00	00	02
2	00	00	99	00	00	00	00	00	00	00	00	00	00	00
3	00	02	01	90	00	03	00	01	00	00	00	00	00	03
4	00	02	01	00	94	01	00	01	00	00	00	00	00	01
5	00	01	00	01	00	96	00	01	00	00	00	00	00	01
6	00	00	00	00	00	06	92	00	00	00	00	00	00	02
7	00	01	00	01	00	01	00	96	00	00	00	00	00	01
8	03	00	00	00	00	01	00	03	86	00	00	00	00	07
9	00	01	00	00	00	00	00	00	00	95	01	00	00	02
10	10	05	01	00	00	01	00	00	00	00	80	00	00	03
11	00	02	12	01	00	02	00	00	00	00	00	82	00	01
12	00	00	02	00	00	01	00	00	00	00	00	00	96	00
13	00	01	01	00	00	01	00	00	00	00	00	00	00	97

Table 3: Mapping of Classes for the Örebro Municipality Benchmark Dataset

Class ID	Class Name	Class ID	Class Name
00	Background	07	OldTree
01	Road	08	Pergola
02	Building	09	Playground
03	Bush	10	SocialArea1
04	CultivArea	11	SocialArea2
05	Grass	12	SocialArea3
06	NewTree	13	Shadow

The model also preserves the overall shape and continuity of vegetation regions, even when their appearance is partially obscured by shadows. This demonstrates the effectiveness of the proposed shadow-aware framework in distinguishing vegetation from visually similar shadow regions. Furthermore, the consistent agreement between the predicted and ground-truth masks across different surface types highlights the model's ability to maintain reliable segmentation under challenging illumination conditions. However, class boundaries involving thin structures and mixed pixels remain challenging, while shadow-edge regions can mimic object boundaries and introduce localized segmentation errors.

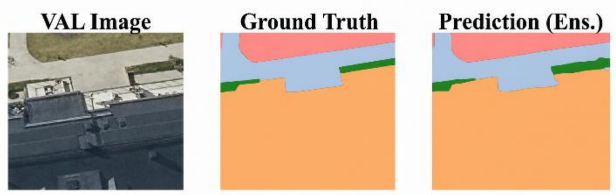


Figure 8: The Qualitative Segmentation Outcome for Ensemble

Performance Analysis

Table 2 shows the normalized confusion matrix which gives class wise segmentation performance of proposed model on dataset 1. The actual class is represented by the rows. The predicted class is represented by a column. Correct classifications are marked on the main diagonal while misclassifications get noted on the remaining parts. Table 2 reveals significant diagonal dominance in most classes, affirming the model's capability to make accurate predictions on Dataset 1. Furthermore, high classification rates are attained by several classes such as Background (1.00), Building (0.99), Grass (0.96), OldTree (0.96), SocialArea3 (0.96), Shadow (0.97), etc. These classes possess distinct visual cues enabling the model to isolate them effectively. This includes consistency in texture, shape, or distinctive color contrast.

The performance is moderate for Cultivated area (0.94), Road (0.95), Bush (0.90), NewTree (0.92). The values are still relatively higher but there is some confusion with the similar looking neighbouring classes. For example, a few pixels of Bush were classified as Grass or Cultivated Area, which is likely due to the overlying vegetation in the aerial image.

One can see relatively lower diagonal values for Pergola (0.86); SocialArea1 (0.80); SocialArea2 (0.82). These classes are more likely to be confused with the background, road, or building classes. This happens due to the similarity of the structures, shadowing effect and the lesser space that makes the boundary difficult to identify.

As seen from the concentration of values just above the diagonal in Table 2, the proposed framework can effectively differentiate between the classes. Moreover, given the dispersed nature of the off-diagonal spread and the associated low values, it becomes evident that the misclassifications are restricted to the local domain.

The proposed maximum-confidence ensemble technique was evaluated on the Örebro Municipality benchmark dataset. Moreover, compare the proposed ensemble method with that of (Rahaman et al., 2024) weighted ensemble that was already reported performances (Tables 4 and 5).

The average results of the proposed method successfully beats the baseline in all semantic classes. The Green Area Factor in addition to other interesting semantic classes achieves mean IoU of 0.88 and weighted ensemble obtains mean IoU of 0.82 (Table 4). As a result, region-oriented segmentation is better. The urban components with distinct shapes measured significant intersection over union results ranging from 0.91 to 0.97. Consequently, the model scales randomly learns the geometric patterns of various classes. In addition, it is very robust for illumination change due to the indoor and outdoor parts in Bouncing, Shadow and Playground. The ensemble technique is advantageous for vegetation classes that are highly important for Green Area Factor. Grass and Old Tree get 0.92 IoU; Bushes get 0.83 IoU; New Tree improves from 0.69 to 0.78 (Table 4). Through this approach, the model will be able to better differentiate fine-grained vegetation classes from other surfaces.

The results of F1-score (Table 5) also shows considerable improvement. According to the results, while the weighted baseline gets an average F1-score of 0.82, the proposed ensemble gets 0.93. Likewise, New Tree experiences an increase from 0.31 to 0.88. Gains of Playground Shows a 0.64 to 0.95. Likewise, Cultivated Land increased from 0.81 to 0.93. The findings indicate that the ensemble enhances not only overlap but can also guarantee reliable detection even for challenging classes.

Essentially, InternImage-T with UPerNet and ResNet-34 with UNet++ methods via confidence-based fusion gives more accurate results.

Table 4: Per-class IoU Comparison between the Weighted Ensemble and the Proposed Maximum Confidence Ensemble Model

Method	Bush	Grass	Old Tree	Shadow	Road	Cult. Land	New Tree	Pergola	Social Area	Playground	Building	Mean IoU
Weighted Ensemble	0.78	0.86	0.85	0.90	0.86	0.79	0.69	0.78	0.77	0.81	0.94	0.82
Proposed Maximum Confidence Ensemble	0.83	0.92	0.92	0.93	0.91	0.88	0.78	0.82	0.82	0.91	0.97	0.88

Table 5: Per-class F1-score Comparison between the Weighted Ensemble and the Proposed Maximum Confidence Ensemble Model

Method	Bush	Grass	Old Tree	Shadow	Road	Cult. Land	New Tree	Pergola	Social Area	Playground	Building	Avg F1
Weighted Ensemble	0.90	0.95	0.94	0.95	0.92	0.81	0.31	0.84	0.84	0.64	0.98	0.82
Proposed Maximum Confidence Ensemble	0.90	0.96	0.96	0.96	0.95	0.93	0.88	0.90	0.90	0.95	0.99	0.93

Table 6: Performance of the Proposed Model in Semantic Segmentation of Khulna University Dataset. Some Metrics Include IoU, Precision, Recall and F1-Score for each Class

Metric	Shrub	Newborn Grass	Big Tree	Small Tree	Road	Water	Weeds	Young Weeds	Soil	Playground	Building	Others	Concrete-Grass-Pav.	Avg
IoU	0.64	0.92	0.90	0.56	0.83	0.99	0.90	0.79	0.61	0.97	0.98	0.78	0.78	0.82
Precision	0.90	0.96	0.95	0.88	0.86	0.99	0.95	0.85	0.90	0.99	0.99	0.93	0.83	0.92
Recall	0.69	0.96	0.95	0.60	0.96	0.99	0.94	0.92	0.65	0.99	0.99	0.84	0.93	0.87
F1-Score	0.78	0.96	0.95	0.71	0.91	0.99	0.94	0.88	0.76	0.99	0.99	0.88	0.88	0.90

Performance Evaluation Using the Khulna University Dataset

Having achieved strong results on the benchmark dataset, we tested our proposed maximum-confidence ensemble on data collected from Khulna University in order to measure its efficacy on real-world data and prediction of GAF. After that, the data also split into 75% training, 20% validation, and 5% test with similar preparations. The ensemble created semantic segmentation maps in new areas of the campus that were not seen by the model. The average IoU of the ensemble was 82% as shown in Table 6 In particular, the model achieved an average precision of 0.92, an average recall of 0.87 and therefore, an average F1-score of 0.90. This means there is confidence in prediction for most classes. We observed that the classes, which are visually different in appearance, appeared in larger areas like Water, Buildings, Playground, NewBornGrass, BigTrees, and Weeds.

Classes such as House, Car and Fence have clear boundaries which are enhanced by contrasting colors or texture. Additionally, their shapes in our dataset are likely to be stable, which makes it easier for the model to recognize them. On the other hand, shrub, soil and small trees had much lower IoU (0.56-0.64). Classes in the area have a strong resemblance to each other and surrounding surfaces. Furthermore, they frequently possess irregular shapes of sparse coverage in our dataset. Nonetheless, their accuracy was fairly high, meaning the model was usually accurate. The primary concern is low recall, primarily stemming from overlooked areas, as opposed to false alarms.

The results obtained by using the proposed ensemble method for handling large and different objects are promising. It provides stable results for small objects and for objects that are difficult to recognize. The model results performed well for all the object categories and therefore it is a robust method for accurate GAF estimation. The model can be used for all types of campus scenes.

Training and Validation Performance

Figure 9 shows us how well the model learns. It has a jump in the first 25 epochs. The model gets a lot better fast. It goes from not being very good to being more than 90% accurate. What is really great about this is that it is very consistent. The difference between the training and the validation, which is shown in gray is very small. This stays the same for all 250 epochs. This tells us that the InternImage-T with UPerNet architecture is very good at generalizing. This means it is not just remembering the training data. The InternImage-T with UPerNet architecture is actually learning things that it can use on pictures it has not seen before. The InternImage-T with UPerNet architecture is really good at learning features.

Figure 10 shows a very effective and stable learning curve. The model exhibits a very aggressive "learning phase" in the first 25 epochs, with the accuracy rising dramatically from the baseline of ~0.5 to well above 90%. The difference between training and validation (shaded in light gray) is very small and consistent over the 250 epochs. This implies that the ResNet-34 and UNet++ architecture is very general, not only is it "memorizing" the training data but it is also learning features that would be applicable to new images that it had never seen before.

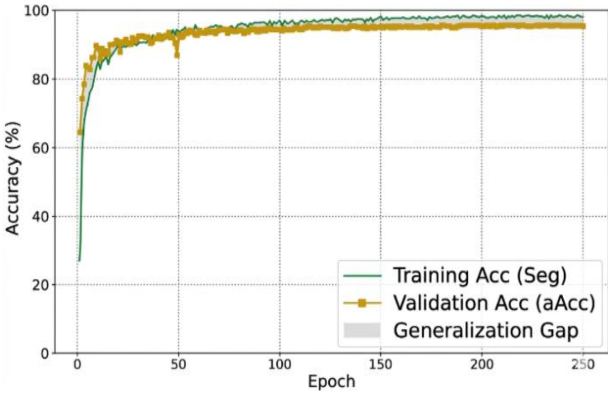


Figure 9: The Accuracy Curves for Training and Validation of InternImage-T and UPerNet

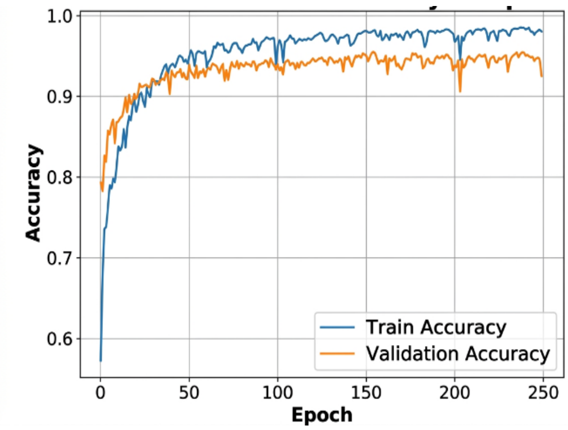


Figure 10: The Accuracy Curves for Training and Validation of UNet++ and ResNet-34

Normalized Confusion Matrix

The normalized confusion matrix is given in Table 7. It evaluates the ability of the ensemble model in predicting land-cover classes at a pixel level for 13 classes. High diagonal values are a good sign of discriminating ability especially for classes that are different in both their spectral and structural characteristics. Water (0.99), Building1 (0.99) and Field (0.99) are almost perfect with respect to accuracy, suggesting the model's ability to describe homogeneous spectral signatures and very clear geometric patterns. Similarly, the reliability of the classifications of Roads (0.96) and NewBornGrass (0.96) is also high.

The most common misclassifications occur between similar classes, both visually and in the material. SmallTree (0.61) has been misclassified with Weeds (0.20) in some instances, possibly because of similar fine scale textural features in vegetated areas and similar spectral responses. Sometimes soil (0.65) is mistaken for Road (0.25): dry soil and unpaved road surfaces have similar reflectance properties, especially under different light conditions.

Table 7: Confusion Matrix (0–100 %) of Proposed Maximum-confidence Ensemble for Khulna University Dataset

Pred \ True	0	1	2	3	4	5	6	7	8	9	10	11	12
0	96	01	01	00	00	00	00	01	00	00	00	00	00
1	01	99	02	00	00	01	00	00	00	00	00	00	00
2	00	00	95	00	00	00	00	03	00	00	00	00	01
3	05	03	09	61	00	00	00	20	00	00	02	00	00
4	06	03	02	00	84	00	00	03	00	00	00	01	02
5	25	01	00	00	00	65	00	03	00	00	02	03	00
6	00	00	00	00	00	00	99	01	00	00	00	00	00
7	01	00	03	00	00	00	00	94	00	00	01	00	01
8	03	00	04	00	00	00	00	00	93	00	00	00	00
9	08	00	02	00	00	00	00	16	00	69	00	01	05
10	00	00	00	00	00	00	00	00	00	00	99	00	00
11	00	00	01	00	01	00	00	00	00	00	00	96	01
12	01	01	05	00	01	00	00	00	00	00	00	00	92

Table 8: Mapping of Classes for the Khulna University Dataset

Class ID	Class Name	Class ID	Class Name
00	road	07	weeds
01	building1	08	concreteGrassPaver
02	bigTree	09	shrub
03	smallTree	10	field
04	others	11	NewBornGrass
05	soil	12	YoungWeeds
06	water		

Generally, errors tend to group together in classes with similar spectral and textural properties. Both these classes have common characteristics. This means that there is some level of confusion amongst them. This implies that most of the errors are due to features overlapping and not due to fluctuations in the model.

Qualitative Results

A qualitative comparison is carried out by using one validation image and one test image from the Khulna University validation and test image datasets, respectively, as shown in Figure 11. The ensemble results show better object boundaries and there is a significant decrease in misclassification when compared with the individual model results.

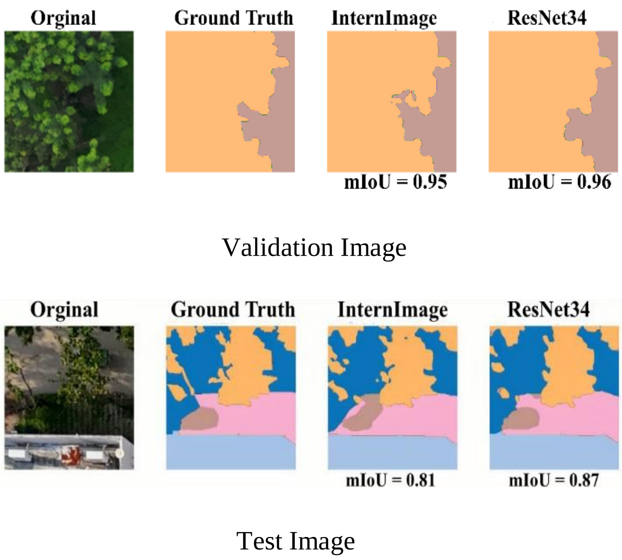


Figure 11: Prediction Comparison of InternImage-T with UPerNet and ResNet-34 with UNet++

In both samples, vegetation and built-up areas are better defined and shadowed areas are dealt with in a more uniform manner. The ensemble yields more coherent and smoother segmentation maps, especially on complex object boundaries. The visual results are consistent with the quantitative results and illustrate the strength of the proposed framework on the validation and test datasets that were not seen during the learning process.

GI Calculation

Table 9 gives the weightings factors for Greenery Index, by class.

Table 9: Class-wise Weighting Factors for Greenery Index Calculation

Class	Factor
Road	0.0
Building	0.0
Big Tree	2.4
Small Tree	1.30
Others	0.0
Soil	1.50
Water	1.0
Weeds	0.70
Concrete Grass Paver	1.50
Shrub	0.2
Field	1.2
New Born Grass	1.2
Young Weeds	0.40

The weight assignment used in this study draws on established guidelines from well-cited urban green assessment research. Weights are defined as the relative ecological contribution of each land cover type: non-green land surfaces such as roads and buildings are assigned a value of 0, while vegetation categories are assigned weights based on their canopy coverage, biomass, and environmental significance.

We then validate the performance of the ensemble prediction on the validation samples displayed in Figure 12, getting per class pixel count (Table 10) from the ensemble prediction, compute the Greenery Index (GI) with the set weights and transform the segmentation results into numeric indices.

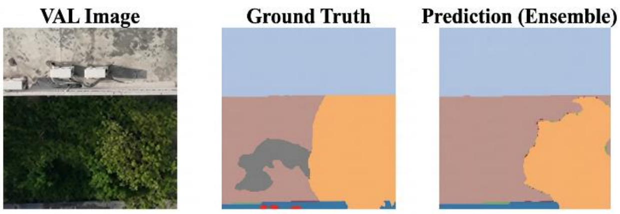


Figure 12: Representative Validation Sample Showing the Ensemble Model Output for Class-wise Greenery Index (GI) Estimation

The Greenery Index is computed as:

$$GI = \frac{\sum_c (A_c \cdot F_c)}{A_{\text{total}}} = 0.82$$

Interpretation: According to this study, the Greenery index (GI) of the ensemble prediction is 0.82. This value represents a relatively high level of vegetation coverage in the analyzed area, and the vegetation coverage distribution predicted by the model can be represented by the index. It can be applied to automated environmental assessment, monitoring of green space in urban areas, and planning activities.

Table 10: Class-wise Pixel Coverage of the Ground-truth and Ensemble Model Prediction for the Validation Sample

Class ID	Name	GT	Pred (Ens)
00	**Road**	1644	1465
01	**Building1**	24899	24921
02	**Bigtree**	18718	15623
03	**Smalltree**	0	0
04	**Others**	159	0
05	**Soil**	0	0
06	**Water**	0	0
07	**Weeds**	15622	23527
08	**Concretegrasspaver**	0	0
09	**Shrub**	4494	0
10	**Field**	0	0

Greenery Index (GI) Estimation for Some Landscape

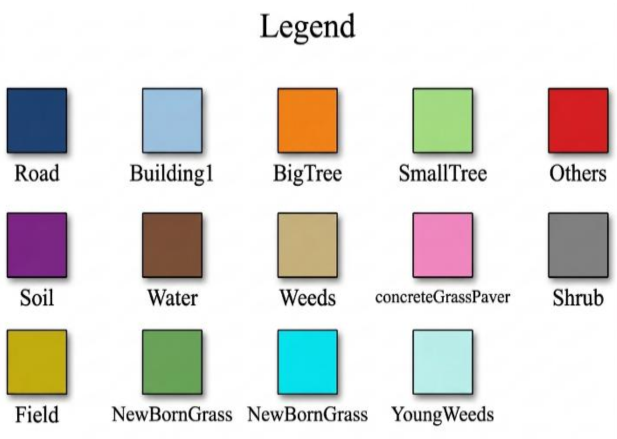


Figure 13: Masks of Segmentation Classes with Colors Indicating the Mapping of Each Class

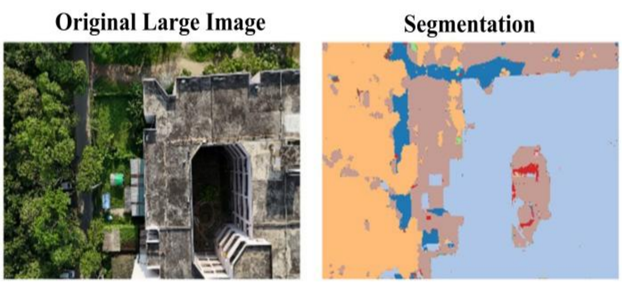


Figure 14: The Original Aerial Image, the Predicted Semantic Segmentation Mask and the Estimated Greenery Index (GI) = 0.8217 are Shown in Site_ID-6

The result of the analysis presented a high value of the Greenery Index (GI) of 0.8217 that indicates there are more natural elements in the area (Figure 14). The vegetation occupies the majority of the area with trees covering 27.8% (80,275 px) and grass 21.5% (309,651 px) of the total area. Little built-up exists with buildings representing 4.64% and roads 3.7%, with minor elements being water bodies (0.2%) and small areas of soil or burnt grass. The overall area has a green predominance with an estimated semantic segmentation mask of 0.8217, which corresponds to an estimated Green Index (GI) = 0.8217.

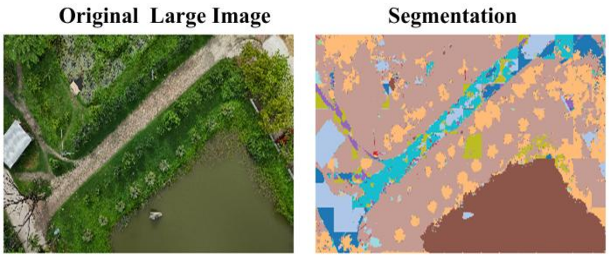


Figure 15: The Original Aerial Image, the Predicted Semantic Segmentation Mask and the Estimated Greenery Index (GI) = 0.9988 are Shown in Site_ID-7

In another area the GI is nearly maximum, 0.9988, indicating that the area is almost completely natural (Figure 15). Water bodies dominate (84%), followed by trees (15.8%) and a small amount of soil (1.1%) and a patchy amount of grass (5.2%). Building and roads make up less than 10% of human-made structures. Clearly, this area has a high GI, which indicates that it is a natural landscape.

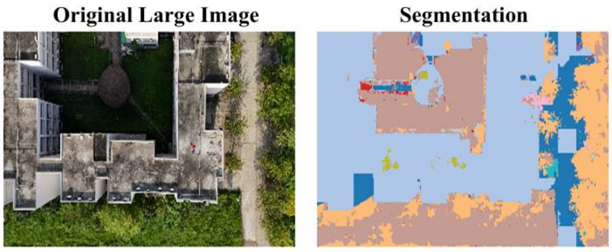


Figure 16: The Original Aerial Image, the Predicted Semantic Segmentation Mask and the Estimated Greenery Index (GI) = 0.5998 are Shown in Site_ID-8

The opposite extreme is a GI of 0.5998 which represents a more mixed landscape, where there is a mix of urban development and natural features (Figure 16). The buildings cover 45.9% of the area, while there is still significant natural cover, including trees (16.1%) and a large water body (18.4%). Small scale components include soil (0.8%), roads and concrete (0.9% each). This balance is reflected in the moderate GI, indicative of a suburban or urban environment where there is a proportion of green and blue space.

In the same way, a GI of 0.6123 indicates a well-balanced urban natural environment (Figure 17). The buildings represent 15.8% and roads for 0.9%, while natural elements dominate: trees 8. 6%, grass 7.5%, soil 1.9% and bodies 8.7%. This is slightly above 0.6 The GI indicates that it is a landscape with very slight influence from built-up areas and is therefore integrated into the green signal that it is a landscape with a very slight influence of built-up areas indicating a well-integrated green-urban landscape.

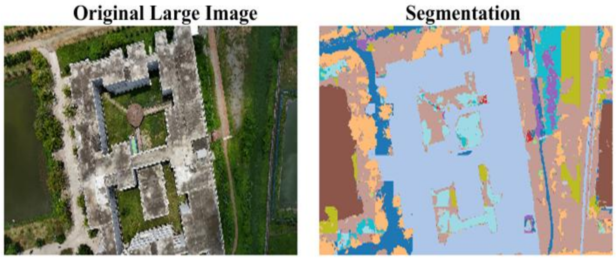


Figure 17: The Original Aerial Image, the Predicted Semantic Segmentation Mask and the Estimated Greenery Index (GI) = 0.6123 are Shown in Site_ID-9

Estimated by the trained segmentation model, the classified GI are given as (0.8217, 0.9988, 0.5998, 0.6123) which give a preliminary and automatic determination of the green cover in each of the zones. Validation of these results will be done in the future by performing site-specific measurements of greenery fraction (GYF) through field surveys. The measured values will be compared to the value predicted by the GI model, and the accuracy, precision and reliability of the model in a real-world context will be evaluated.

In Summary, the authors demonstrate the stability of the training process, high class-wise accuracy, and reliable segmentation even in shadowed regions. InternImage and UNet++ were combined into the Maximum Confidence package, which improved robustness; and the greenness index has been proposed as a tool to assess the vegetation cover of urban environments.

Statistical Validation and Comparative Analysis

The effectiveness of the proposed automated framework for estimating greenery was tested with the predicted values being compared to manually measured Green Area Factor (GAF) values from selected benchmarking sites. Comparisons were made both raw and normalized, to ensure proper statistical interpretation. The results show that the model can well simulate the real values measured with high consistency.

Table 11: Spatial Properties of the Benchmarking Sites and the Measured Green Area Factor (GAF) for Each Site

Site ID	Total Area (m^2)	Measured GAF
1	23,616.94	1.05
2	19,211.75	1.41
3	21,852.17	0.68
4	23,535.86	1.18
5	14,046.81	0.56
6	6,379.02	0.98
7	3,595.08	1.18
8	22,478.13	0.70
9	6,244.77	0.63

The spatial characteristics of the selected benchmarking sites are shown in Table 11 along with the measured GAF values. Such locations were selected to make the model useful for testing under different land-cover conditions that exist within cities. The measured values were obtained from on-site observations or reference measurements and were used as ground truth for validation. Overall, the results demonstrate that the framework can reproduce variations in GAF across heterogeneous sites. The predicted values preserve the spatial greenery trends observed in the field measurements. The close agreement between predicted and measured GAF values indicates that the framework can effectively capture differences in greenery across diverse urban environments. The results also show that the proposed method remains consistent across sites with different proportions of vegetation and built surfaces. The consistency between raw and normalized comparisons (Table 12) further supports the robustness of the estimation. This end-to-end validation demonstrates the practical value of the framework beyond conventional segmentation metrics. The findings suggest that the framework has potential for automated and scalable greenery assessment in real-world urban planning and environmental monitoring applications. Nevertheless, evaluation using larger and geographically diverse benchmarking datasets is required to further establish its generalizability.

Table 12: Normalized Measured and Estimated Values and Their Comparisons for Some Image Samples

Site ID	Estimated	Measured	Estimated (Normalized)	Measured (Normalized)
1	1.0283	1.05	0.7672	0.5765
2	1.1596	1.41	0.9799	1.0000
3	0.7120	0.68	0.2548	0.1412
4	1.1720	1.18	1.0000	0.7294
5	0.5547	0.56	0.0000	0.0000
6	0.8200	0.98	0.4298	0.4941
7	0.9900	1.18	0.7052	0.7294
8	0.5990	0.70	0.0718	0.1647
9	0.6100	0.63	0.0896	0.0824

Green Area Factor (GAF) values are also estimated using the proposed framework and compared with the measured reference GAF values for nine benchmarking sites in Table 12. The estimated values (0.5547-1.1720) are near the measured values (0.56-1.41), which show good prediction accuracy. For example, Site 5 (0.5547 vs. 0.56) and Site 4 (1.1720 vs. 1.18) show very small differences. The datasets were first min-max normalized for clarity in comparison. The estimated values are plotted in normalized form, and it is clear that they have the same overall trend as the measured values. The normalized values for Site 2 were the highest and for Site 5 were the lowest, both times, thus indicating the correct relative ranking. There are some minor deviations, but the overall consistencies between the estimated and the measured values provided here show that the proposed framework can be effective in capturing not only numerical GAF values but spatial variation of greenery as well. This facilitates its application in automatic monitoring and planning of urban vegetation.

A paired samples t-test at the 95% level of confidence was performed to further investigate statistical agreement. The test gave a t-statistic of 1.035 and a p-value of 0.331. With a p value of > 0.05, the null hypothesis is not rejected. This indicates that there is no significant difference between the estimated and measured reference value. The automated predictions are in a practical sense similar to the real observations statistically.

Further, the estimated value measured value correlation coefficient was r = 0.953, showing a very high linear relationship between the estimated and measured values. This shows that the model prediction closely matches the measured greenery value as it changes either up or down.

The framework implemented showed a lower Mean Absolute Error (MAE = 0.087) than the reported reference value (MAE = 0.140) when compared to the reference study (Rahaman et al., 2024). This decrease in error represents an improvement in the accuracy of numerical prediction and closer agreement with the ground truth observations due to the current segmentation-based estimation pipeline.

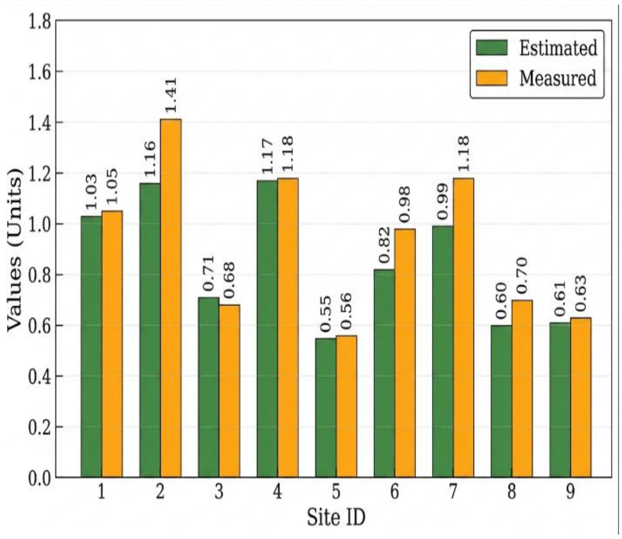


Figure 18: Estimated and measured Greenery Index (GI) values for nine sites, and the consistency of the performance of the proposed method

The Figure 18 presents a site-wise comparison between estimated and measured GI values. Overall, the estimated values closely follow the measured trends, indicating strong predictive consistency. Minor deviations are observed at certain sites (e.g., Sites 2 and 7), where the model slightly underestimates higher GI values. However, the general alignment across all sites demonstrates that the proposed framework can reliably capture spatial variations in urban greenery. This consistency supports its effectiveness for practical ecological assessment and urban planning applications.

The validation dataset includes nine benchmarking sites, but the statistical results are still encouraging. Low prediction error, high correlation and t-test difference not significant show that the proposed system is a good alternative to manual greenery assessment methods which requires a lot of time. Larger dataset can be used to enhance model robustness and generalization. The experimental results validate the proposed automated urban greenery estimation model and show high potential applicability to smart city environmental monitoring applications, especially those for which geographical data are needed, and for which manual collection is costly.

Conclusion and Future Work

This study proposed a shadow-resilient semantic segmentation framework for accurate urban vegetation mapping and Greenery Index estimation from high-resolution aerial imagery. The framework combines InternImage-T with UPerNet and ResNet-34 with UNet++ through a class-wise maximum-confidence ensemble strategy. Unlike conventional weighted ensembles that average predictions from multiple models, the proposed approach assigns each semantic class to the model demonstrating superior validation performance. This class-wise selection preserves the complementary strengths of both architectures, improves boundary localization, and reduces shadow-induced misclassification without introducing the probability smoothing commonly associated with weighted fusion.

The proposed framework was evaluated on two independent datasets representing different imaging platforms and urban environments: the Örebro Municipality benchmark dataset and a UAV dataset collected over Khulna University. Experimental results demonstrate that the proposed approach consistently improves segmentation performance under challenging illumination conditions. The framework achieved a mean Intersection over Union (mIoU) of 0.88, an F1-score of 0.93, and a pixel accuracy of 0.97 on the Örebro dataset, while achieving an mIoU of 0.82, an F1-score of 0.90, and a pixel accuracy of 0.96 on the Khulna University dataset. Compared with the weighted-ensemble baseline, the proposed class-wise maximum-confidence strategy provided more accurate vegetation segmentation, particularly in shadow-affected regions where conventional methods frequently confuse vegetation with non-vegetative surfaces. Furthermore, the estimated Greenery Index showed strong agreement with field-measured Green Area Factor values, achieving a Pearson correlation coefficient of 0.953 while reducing the mean absolute error from 0.140 to 0.087. These results demonstrate that improving segmentation reliability directly enhances the accuracy of ecological indicators used for urban planning, environmental monitoring, and sustainable green infrastructure assessment.

Although the proposed framework demonstrates promising performance, several opportunities remain for future research. First, shadow-aware feature learning can be strengthened by incorporating explicit shadow masks as an auxiliary supervision branch or additional input during training, enabling the network to learn more discriminative representations of shadowed vegetation. Second, the framework should be evaluated on larger multi-city and multi-season datasets acquired under diverse environmental and illumination conditions to investigate its cross-domain generalization capability. Future work will also explore the integration of multispectral imagery and LiDAR-derived canopy information with RGB aerial imagery to improve vegetation detection in regions where spectral ambiguity remains significant. Finally, incorporating recent vision foundation models together with domain adaptation techniques and developing lightweight model variants suitable for real-time UAV deployment would further improve the scalability and practical applicability of the proposed framework for continuous urban green infrastructure monitoring and smart city management.

Acknowledgement

This research was supported by grant (36/2024) of the Research and Innovation Center (RIC), Khulna University, Bangladesh.

Conflict of Interest

The authors declare no conflict of interest.

References

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. (2018). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848.

Chen, W., Li, J., & Zhang, Y. (2023). Hybrid CNN-Transformer network for semantic segmentation of high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2021). BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, *129*, 3051–3068.

Diakogiannis, F. I., Waldner, F., Caccetta, P., & Wu, C. (2020). ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *162*, 94–114.

Dimitrovski, I., Spasev, V., Loshkovska, S., & Kitanovski, I. (2024). U-net ensemble for enhanced semantic segmentation in remote sensing imagery. *Remote Sensing*, *16*(12), 2077.

He, J., & Ma, R. (2023). Multimodal semantic segmentation of remote sensing images with CNN-transformer fusion. *Remote Sensing*.

Khan, B. A., & Jung, J.-W. (2024). Semantic segmentation of aerial imagery using u-net with self-attention and separable convolutions. *Applied Sciences*, *14*(9), 3712.

Li, R., Zhang, M., & Wang, Z. (2024). Change detection in satellite imagery using transformer-based Siamese networks. *Remote Sensing*.

Li, X., Zhang, Y., Ding, K., & Zhang, L. (2022). Transformer-based deep learning for remote sensing image semantic segmentation. *Remote Sensing*, *14*(7), 1636.

Liu, H., & Chen, X. (2024). Vision transformer adapters for remote sensing semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Ma, L., Liu, Y., Zhang, X., & Ye, Y. (2022). Attention-based U-Net for remote sensing image segmentation. *IEEE Access*, *10*, 54212–54223.

Nan, G., Li, H., Du, H., Liu, Z., Wang, M., & Xu, S. (2024). A semantic segmentation method based on AS-Unet++ for power remote sensing of images. *Sensors*, *24*(1), 269.

Nsiah, R. A., Mantey, S., & Ziggah, Y. Y. (2023). Building segmentation from UAV orthomosaics using unet-resnet-34 optimised with grey wolf optimisation algorithm. *Smart Construction and Sustainable Cities*, *1*(1), 21.

Panboonyuen, T., & Jitkajornwanich, K. (2023). Road segmentation on aerial images using deep convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*.

Panghurian, F. P., Pranoto, H., Irwansyah, E., & Pramudya, F. S. (2024). Comparison of Resnet Models in UNet Classifier for Mapping Oil Palm Plantation Area with Semantic Segmentation Approach. *International Journal of Advanced Computer Science & Applications*, *15*(7).

Rahaman, G. A., Längkvist, M., & Loutfi, A. (2024). Deep learning based automated estimation of urban green space index from satellite image: A case study. *Urban Forestry & Urban Greening*, *97*, 128373.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 234–241.

Santosa, H., & Kusuma, G. P. (2025). Urban area road scene segmentation using InternImage-adapter model. *Commun. Math. Biol. Neurosci.*, *2025*, Article-ID.

Shen, J., Huo, C., & Xiang, S. (2024). Siamese InternImage for Change Detection. *Remote Sensing*, *16*(19), 3642.

Shi, Y., & Luo, P. (2024). Fine-grained land use classification using transformer-based semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Slyusar, V., Protsenko, M., Chernukha, A., Melkin, V., Petrova, O., Kravtsov, M., Velma, S., Kosenko, N., Sydorenko, O., & Sobol, M. (2021). Improving a neural network model for semantic segmentation of images of monitored objects in aerial photographs. *Eastern-European Journal of Enterprise Technologies*, *6*(2), 114.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., & Wang, X. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, J., Sun, L., & Huang, K. (2023). Ensemble deep learning for land cover semantic segmentation from remote sensing images. *IEEE Access*.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., & others. (2023). Internimage: Exploring large-scale vision foundation models with deformable convolutions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14408–14419.

Wang, Z., Li, J., Tan, Z., Liu, X., & Li, M. (2023). Swin-UPerNet: A semantic segmentation model for mangroves and spartina alterniflora loisel based on UPerNet. *Electronics*, *12*(5), 1111.

Wu, Y., & Zhang, M. (2024). Swin-CFNet: An attempt at fine-grained urban green space classification using Swin transformer and convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018). Unified perceptual parsing for scene understanding. *Proceedings of the European Conference on Computer Vision (ECCV)*, 418–434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*.

Xu, Q., Wang, H., & Li, M. (2024). Multi-scale semantic segmentation of satellite imagery using transformer-enhanced UNet. *Remote Sensing*.

Yu, J., Zhang, Y., Sun, F., Wang, L., & Lu, R. (2024). Solution for CVPR 2024 UG2+ Challenge Track on All Weather Semantic Segmentation. *arXiv Preprint arXiv:2406.05837*.

Yuan, Q., Wei, Y., & Meng, X. (2021). Deep learning-based semantic segmentation of remote sensing images: A review. *Remote Sensing*, *13*(23), 4772.

Zhang, L., & Du, B. (2025). A survey on semantic segmentation of remote sensing images using deep learning. *IEEE Access*.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2881–2890.

Zhou, W., & Li, Q. (2022). Improved DeepLabV3+ for semantic segmentation of satellite images. *Sensors*, *22*(14), 5312.