Semantic segmentation of green vegetation areas in remote sensing images: a multimodal 7-Channel U-Net approach

Shiping Ye [1,4][0000-0002-9771-7168], Anastasia Intyakova [2], Yafen, Xu[1,4], Sergey Ablameyko [2,3][0000-0001-9404-1206]

[1] Zhejiang Shuren University, Hangzhou, China
zjsruysp@163.com

[2] Belarusian State University, Nezavisimosti avenue 4, Minsk, Republic of Belarus
[3] United Institute of Informatics Problems of NASB, Surganov str., 6, Minsk, Republic of Belarus
ablameyko@bsu.by
[4] Zhejiang - Belarus Joint Laboratory for Intelligent Image Computing

Abstract. An improved modification of U-Net adapted to work with 7-band multimodal data (RGB + NIR + RedEdge + DSM + NDVI) for the purpose of high-precision semantic segmentation of urban green areas at remote sensing images is proposed. Specifically, the proposed model extends the encoder of the standard U-Net to accept a 7-channel input tensor, thereby enabling early fusion of spectral and elevation information. The introduction of a Digital Surface Model (DSM) provides crucial vertical structural cues, enabling the network to distinguish between spectrally similar land cover classes, such as low-lying grasslands and tall trees. Furthermore, the red-edge band and the computed NDVI index enhance sensitivity to vegetation health and mitigate misclassifications commonly found on artificial surfaces such as green roofs or colored asphalt. The algorithm's high generalization ability was proven. Testing on a completely new park landscape, not used in training, demonstrated an F1-score of 0.93. High overall segmentation quality was achieved. The final Mean IoU was 0.725, significantly exceeding the results of baseline 3-band models and classic pixel-based classification methods (such as NDVI thresholding).

Keywords: semantic segmentation, remote sensing images, green areas, digital surface model, U-net.

1 Introduction

Modern methods for monitoring urban environments are based on integrating heterogeneous Earth remote sensing (ERS) data. For semantic segmentation of vegetation, it is critical to utilize not only the optical spectrum but also specialized spectral channels and information on the spatial structure of objects [1].

The task of identifying vegetation in satellite images belongs to the class of semantic segmentation problems [2,18]. Unlike image classification, where a single label is assigned to the entire image, segmentation requires assigning a class label to each pixel in the image, forming a mask that topologically corresponds to the input data. Accounting for spatial context is crucial for land use classification, making deep learning the most promising approach [3,19].

Modern models focus on balancing high spatial resolution and computational efficiency. Three key groups of architectures are distinguished in the literature:

• U-Net-based models, which are the most common for this task [4, 5]. For example, the SA-SC-U-Net (Self-Attention with Separable Convolutions) architecture, which is one of the leading architectures as of 2026 (up to 91.78% accuracy, 81.82% IoU), uses a self-attention mechanism to capture global context [6].

• Transformer-based models: Architectures like SegFormer and CTSA-DeepLabV3+ are increasingly being adopted due to their ability to model global contextual relationship[7], which is critical for

• Hybrid models: Combine convolutional neural networks (CNNs) with transformers for super-resolution tasks (e.g., SR-UGSnet), efficiently transforming low-resolution data into detailed maps [8].

Traditional CNN architectures (such as AlexNet and VGG) have an encoder structure that successively reduces the image dimensionality by extracting abstract features. However, the final fully connected layers lose information about the spatial arrangement of objects, which is unacceptable for segmentation tasks [2,15]. A breakthrough was the emergence of fully convolutional networks (FCNs), where fully connected layers are replaced by convolutional ones and transposed convolution operations are used to restore resolution [2,14]. However, FCNs suffered from low accuracy at object boundaries due to the loss of detail during the pooling process. A further development of this idea was the U-Net architecture, proposed in 2015 [9,20]. Initially developed for biomedical applications, it demonstrated outstanding results in remote sensing due to its ability to train on small samples and high edge localization accuracy.

Clearly, the choice of architecture directly impacts the accuracy of the final results, especially in conditions with limited labeling. In this paper, we propose an improved modification of U-Net adapted to work with 7-band multimodal data (RGB + NIR + RedEdge + DSM + NDVI) for the purpose of high-precision semantic segmentation of urban green spaces. Specifically, the proposed model extends the encoder of the standard U-Net to accept a 7-channel input tensor, thereby enabling early fusion of spectral and elevation information. The introduction of a Digital Surface Model (DSM) provides crucial vertical structural cues, enabling the network to distinguish between spectrally similar land cover classes, such as low-lying grasslands and tall trees. Furthermore, the red-edge band and the computed NDVI index enhance sensitivity to vegetation health and mitigate misclassifications commonly found on artificial surfaces such as green roofs or colored asphalt.

2. Data Selection for Urban Green Zone Segmentation

The study area is the Gongchenqiao Campus of Zhejiang Shuren University, located in Hangzhou, Zhejiang Province. The campus features diverse land cover types, including trees, shrubs, lawns, buildings, roads, and sports fields, with high vegetation coverage that makes it an ideal experimental site for this study.

This study uses multispectral and oblique UAV imagery as data sources. The DJI M300 RTK drone was used as the acquisition platform. Multispectral images were captured with a MicaSense RedEdge MX camera covering five spectral bands: blue (475 nm), green (560 nm), red (668 nm), red edge (717 nm), and near-infrared (840 nm)..Oblique imagery was acquired using a PSDK 102S V3 camera, which supports independent recording of centimeter-level positional information (POS data) across five viewing angles, features mature control-free technology, and achieves high-precision 3D data collection with a ground resolution of up to 2 mm, enabling detailed representation of terrain textures and structural features.

Flight path planning was conducted using DJI Pilot 2, with automatic flight mode and equidistant image capture. For multispectral imagery, forward and side overlap rates were set to 80% and flight altitude to 110 m, yielding a ground sampling distance (GSD) of 8 cm/pixel. For oblique imagery, forward overlap was 80%, side overlap 75%, and flight altitude 100 m, yielding a GSD of 1.5 cm/pixel.

Multispectral images were processed in Pix4Dmapper: after radiometric correction and orthorectification, single-band orthophotos were output for each of the five bands at 8 cm/pixel, then stacked into a multispectral TIFF in ENVI. Oblique imagery was processed in Context Capture to generate a digital surface model (DSM) at the same spatial resolution as the DOM. As the DOM and DSM originate from different sensors, image registration was performed to ensure spatial consistency, followed by clipping to the study area boundaries in ENVI.

Urban green zone (UGZ) segmentation in remote sensing imagery requires comprehensive analysis of both spectral and elevation data [10,16]. Traditional top-down spectral views, while effective for detecting photosynthetic biomass [11], lack vertical structural insight. A distinctive feature of our approach is the integration of horizontal measurements (top-down view) with spatial-structural characteristics that simulate a perspective (ground-based) view. To recreate this structural dimension, a digital surface model (DSM) encoding absolute object heights is integrated into the analysis. The high spatial resolution (8 cm/pixel) further enables detailed discrimination of small urban elements such as isolated shrubs and narrow lawns [11].

The dataset consists of two co-registered raster layers: a five-band multispectral orthophoto (GeoTIFF, 16-bit) and a single-band DSM (32-bit float). To maximize the neural network's capacity, the input tensor was expanded to seven channels—Red, Green, Blue, NIR, RedEdge, DSM, and NDVI—enabling simultaneous analysis of object height for geometric boundary refinement and enhanced spectral features for reliable chlorophyll detection against artificial surfaces.

To generate labeled training data, a small number of representative samples were manually annotated as polygons in ArcGIS by referencing the original imagery and DSM, and used to train a baseline U-Net model. This model then predicted the remaining unlabeled imagery. The automatic classification results were manually refined by comparing them with the original imagery, the 3D terrain model derived from oblique photography, and on-site scene recognition. Vector editing was applied to correct misclassifications and remove patchy noise, after which the annotations were converted into mask images to produce the final manually corrected dataset.

3. A Modified U-Net Model for Semantic Segmentation of Multi-Channel Images

3.1. The Proposed U-Net Model

The classic U-Net architecture was originally designed for working with single-channel (medical) or three-channel (RGB) images [9]. For the current vegetation segmentation task, the U-Net input layer was modified to accept seven channels instead of the standard three.

A standard convolutional layer in deep learning libraries (PyTorch) expects a four-dimensional input tensor of dimension (B, C, H, W), where B is the batch size, C is the number of channels, and H and W are the height and width of the image.

In this work, the input tensor dimension along the channel axis was increased from C = 3 to C = 7. This required modifying the weights of the first convolutional block of the encoder. The DoubleConv network's input layer was reconfigured as follows:

$$Input: (Batch, 7{,}256{,}256) \xrightarrow{Conv2d} (Batch, 64{,}256{,}256) \qquad (1)$$

This means that the convolution kernels (filters) of the first layer have a depth of 7. This architectural solution implements the Early Fusion feature fusion strategy. Unlike processing spectrum and elevation separately in two independent networks, early integration allows the neural network to identify complex nonlinear relationships between disparate data already at the first stage of training. For example, the network can learn specific filters that are activated only when the following conditions are simultaneously met:

High value in the NIR and NDVI channels (spectral signature of living biomass).

Low value in the DSM channel (geometric signature of an object at ground level).

This combination of features uniquely identifies the class "Lawn" or "Grass", allowing it to be distinguished from spectrally similar "Green Roofs" (where NDVI is high, but DSM is also high) or from artificial stadium surfaces (where DSM is low, but NDVI is low).

3.2. ECA Module

To further enhance the model’s ability to represent multi-channel features, this study introduces an Efficient Channel Attention (ECA) module between the encoder and decoder of the U-Net. Unlike traditional Squeeze-and-Excitation (SE) modules, ECA captures cross-channel interaction information through one-dimensional fast convolutions rather than fully connected layers, thereby avoiding performance losses caused by dimensionality reduction while maintaining low computational overhead. Specifically, an ECA operation is performed on the feature maps output by the encoder before each skip connection: first, global average pooling is applied to the feature maps to obtain global description vectors for each channel; then, one-dimensional convolution with an adaptive kernel size (proportional to the number of channels) is used to compute inter-channel weights; finally, these weights are multiplied channel-by-channel with the original feature maps to achieve channel feature re-calibration. The introduction of the ECA module enables the network to automatically focus on spectral channels (such as NIR and RedEdge) and structural channels (DSM) that are more critical for vegetation segmentation, while suppressing interference from redundant channels.

3.3. Model Training

The effectiveness of a neural network model is determined not only by its architecture but also by the chosen training strategy. The specific nature of satellite data—high scene variability, noise, and pronounced object heterogeneity—requires specialized methods for hyperparameter tuning and loss function adaptation.

Semantic segmentation tasks in urban areas face a fundamental problem called class imbalance [10]. Statistical analysis of the training set masks revealed an extremely uneven distribution of object pixels:

Artificial objects (the "Non-vegetation" class) and background values occupy a dominant position (majority classes).

Target vegetation classes, especially "Moderate Vegetation" (shrubs), are sparsely represented and often comprise less than 5–10% of the total image area (minority classes).

When using the standard loss function (Cross Entropy Loss), the neural network strives to minimize global error along the path of least resistance—predicting the majority class for all pixels. This leads to model bias and an inability to detect rare objects.

In the initial stages of the study, weighted cross-entropy was used, but experiments showed that simple weighting is insufficient for confidently classifying complex examples (hard mining). To address this issue, the Focal Loss loss function was implemented [12,21]. Unlike standard entropy, Focal Loss adds a modulating factor $(1 - p_t)^\gamma$, which reduces the contribution of simple, well-classified examples (such as large areas of asphalt or background) and focuses training on difficult examples (object boundaries, sparse shrubs).

The loss function formula is:

$$FL(p_t) = -\alpha_t (1 - p_t)^\gamma \, log(p_t) \qquad (2)$$

where:

p_t is the probability of the correct class;

γ=2.0 is the focusing parameter, which controls the degree of suppression of simple examples;

α_t — is the class weighting coefficient.

To balance the contribution of classes, α_t weights calculated using the inverse frequency method with manual adjustments were applied:

Class "Background": The weight was reduced to 0.1 to minimize the influence of technical image boundaries (padding).

Class "Shrubs": The maximum weight (1.69) was assigned, forcing the model to actively adjust the weights even with single errors in this class.

Adam (Adaptive Moment Estimation) [13,17] was chosen as the optimization algorithm. Unlike stochastic gradient descent (SGD), Adam adapts the learning rate individually for each parameter, using estimates of the first and second moments of the gradients, ensuring rapid convergence on a complex loss function surface. To prevent overfitting, a set of regularization methods has been implemented:

L2 Regularization (Weight Decay): A penalty of $\lambda = 1 \cdot 10^{-4}$, proportional to the square of the weights has been added to the optimizer. This limits the growth of the parameters, making decision boundaries smoother and reducing sensitivity to noise.

Dropout (Dropout Layer): A Dropout layer (p= 0.5) has been added before the output classifier, randomly dropping 50% of the neurons at each iteration. This is equivalent to training an ensemble of thinned networks and prevents feature coadaptation.

Early Stopping: Training is terminated if the validation error does not decrease within 15 epochs. The weights of the model that showed the best result (Best Model Checkpoint) are retained, not the weights of the last epoch.

A dynamic Learning Rate Scheduling strategy has been implemented using the ReduceLROnPlateau scheduler. If the error metric (Val Loss) does not improve within 5 epochs (patience), the learning rate is automatically halved. This allows the model to take small steps in the final stages to fine-tune the weights around the global minimum.

4. Experimental Procedure, Results, and Analysis

4.1. Dataset Preprocessing

The heterogeneous data structure used in this study (different bit depths, physical meanings, and value ranges) requires a specialized normalization strategy, as using input data with a wide range of values leads to gradient instability during backpropagation. A multi-stage preprocessing procedure was implemented to normalize all channels to a single range of [0,1].

First, spectral channels are normalized. Since the original data has a 16-bit depth (N=16), simply converting to 8 bits would result in a loss of radiometric accuracy in the RedEdge channel. Therefore, linear scaling of digital counts (DN) is applied using the formula:

$$X_{norm} = \frac{X_{raw}}{2^N - 1} = \frac{X_{raw}}{65535.0} \quad (3)$$

where X_{raw}— the original pixel value, N=16 — the bit depth. This operation is performed for all five spectral channels, preserving the relative spectral characteristics of objects.

The height channel (DSM), which has a different distribution, is processed in parallel. A Min-Max normalization strategy is applied to it, with preliminary filtering of outliers (H values < -100, which arise due to sensor errors). Scaling is performed based on global scene statistics (H_min=-10.0 m, H_max=50.0 m) using the formula:

$$H_{norm} = \frac{H_{curr} - H_{min}}{H_{max} - H_{min}} \quad (4)$$

If there is no elevation difference in the image (H_max=H_min), the channel is filled with zeros to prevent division by zero.

Additionally, to improve vegetation separability, a calculated NDVI (Normalized Difference Vegetation Index) channel was added to the input tensor. It is calculated based on the normalized NIR and Red channels as

$$\text{NDVI} = \frac{NIR - Red}{\text{NIR} + \text{Red}} \quad (5)$$

The resulting values, lying in the range [-1,1], are linearly transformed to the range [0,1]. The final input tensor is formed by concatenating the spectral channels, normalized height, and NDVI index, forming an array of dimension $(\mathbf{7}, \boldsymbol{H}, \boldsymbol{W})$.

Since the original images are larger than the GPU's video memory, a sliding window method is used. The image is divided into patches of a fixed size of 256 × 256 pixels. An Overlap Strategy is implemented, with a stride of 128 pixels. This ensures 50% overlap of adjacent patches, allowing the training set to be increased by a factor of 4 and eliminating the problem of edge effects at patch boundaries.

On-the-fly data augmentation is used to prevent model overfitting. Considering the orientation invariance of aerial photographs, geometric transformations are used: random rotations of 90°, 180°, and 270°, as well as horizontal and vertical mirror reflections. Each transformation is applied with a probability of p=0.5, ensuring the model has a high variability of samples.

4.2. Post-processing and visualization

Since the original orthophotos have a resolution significantly higher than the graphics accelerator's video memory (VRAM), directly feeding the entire image to the network is impossible. To solve this problem, the predict_smooth.py software module was developed, implementing a Gaussian Blending Sliding Window (GBSW) method.

Processing algorithm:

Decomposition: The original 7-band data (RGB + NIR + RedEdge + DSM + NDVI) is read from disk. The image is split into fragments (patches) measuring 256 × 256 pixels with a step of 128 (50% overlap).

Weighting: To eliminate the "checkerboard" effect at patch boundaries, Gaussian smoothing is applied. Pixels in the center of the patch are given a higher weight than those at the edges, ensuring seamless stitching. Prediction and Assembly: The neural network generates probability maps for each class, which are then combined into a single raster array the size of the original image.

The algorithm's output is exported in GeoTIFF format with the coordinate system (CRS) preserved, allowing the resulting map to be used directly in GIS systems (QGIS, ArcGIS). The result of semantic segmentation is shown in Figure 1.

Visual analysis demonstrates the high efficiency of the proposed approach:

• Effective layer separation: The algorithm reliably distinguishes between tall trees (dark green zones) and low-growing herbaceous vegetation. This is achieved thanks to the DSM channel: objects with identical spectral responses but different heights are assigned to different classes.

• Geometric accuracy: Unlike earlier experiments, the final model correctly delineates buildings and roads (the "Asphalt" class), avoiding false positives on green roofs, as evidenced by the high IoU value for anthropogenic objects (0.86).

Figure 1. Land cover classification map obtained using the developed neural network.

4.3. Quality Assessment Metrics

Since satellite images often exhibit class imbalances (for example, the area of asphalt significantly exceeds the area of shrubs), the standard accuracy metric (the proportion of correct answers) is not representative. For an objective assessment of the quality of semantic segmentation, specialized metrics are used.

Intersection over Union (IoU, Jaccard index) is the main standard in segmentation tasks. The F1-score (Dice Coefficient) is the harmonic mean of Precision and Recall.

Using this combination of metrics allows for an objective assessment of how accurately the neural network identifies vegetation contours and minimizes false positives for anthropogenic objects. The final analysis uses metric averaging across all classes (Macro-Average), eliminating the dominance of dominant classes (background and asphalt) in the overall assessment.

To assess the qualitative improvement in accuracy, a visual comparison of maps generated in QGIS was conducted, along with a quantitative analysis based on the calculated metrics. Results were compared between the traditional threshold classification method for the Normalized Difference Vegetation Index (NDVI), pixel classification, and the developed U-Net neural network model.

4.4. Comparison with the NDVI Index Method

Analysis of map fragments revealed key shortcomings of the NDVI index method, which were successfully addressed in the current work:

1. False positives on anthropogenic objects:

- NDVI: The traditional method, based solely on spectral radiance, often misclassified building roofs (especially those painted green or with a rough texture) as lawns.

This is due to the fact that old roofing materials can have a spectral response similar to that of decaying vegetation.

- U-Net: A neural network using the height channel (DSM) accurately identified such objects as "Non-vegetation" (class 1). The model "understands" that a flat rectangular object rising 10-20 meters above the ground cannot be a lawn, even if it is green.

2. Working in shaded areas: In urban areas, a significant portion of vegetation is in deep shadow from high-rise buildings. NDVI in such areas drops sharply, leading to missed trees. A convolutional network, by analyzing the spatial context (the pixel's surroundings), successfully reconstructs vegetation class even in low-light conditions.

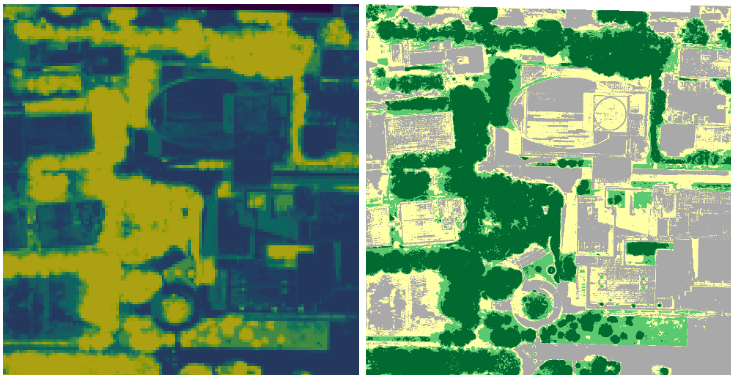
4.5. Comparison with Pixel-Based Segmentation (QGIS)

To evaluate the effectiveness of the neural network approach, a comparison was conducted with a classic pixel-based segmentation (Random Forest) implemented in a GIS environment.

• Salt-and-pepper noise: Classic methods consider each pixel in isolation. This results in high noise levels: isolated vegetation pixels appear on a continuous stretch of asphalt due to glare. The convolutional neural network (U-Net) takes context into account, resulting in consistent and smooth predicted masks. Forested areas are distinguished as dense polygons, significantly simplifying area calculations.

• Consideration of object geometry: In QGIS, class separation often relies solely on spectral luminance. The neural network was able to learn the shape of objects, distinguishing trees not only by color but also by the presence of characteristic elevations.

Comparison of segmentation results by QGIS and U-Net is shown in Fig.2.



a) b)

Figure 2. Comparison of segmentation results: a) QGIS pixel segmentation (salt and pepper effect); b) U-Net semantic segmentation (smoothed areas).

4.6. Quantitative Analysis of Metrics and Error Distribution

For objective verification of the model, metrics calculated on a holdout test set were used. Analysis of the results demonstrates significant success of the proposed architecture:

• Overall performance (Mean IoU = 0.8823): The current model configura-tion with a 7-channel input and Focal Loss demonstrates exceptional general-ization ability, achieving an overall accuracy of 0.9351. The Mean IoU reached an outstanding 0.8823, which is an excellent result for a complex multi-class task in a heterogeneous urban environment.

• Successful detection of anthropogenic objects (IoU = 0.8851): The neural network correctly identifies road surfaces and buildings, minimizing the in-fluence of spectral noise thanks to the early fusion of NDVI and DSM fea-tures. The "Asphalt" class achieved a high IoU of 0.8851, with the confusion matrix showing a correct classification rate (recall) of 0.92.

• Solution to the "Bushes" class problem: The use of Focal Loss enabled a breakthrough in detecting the most challenging minority class. The neural network demonstrates highly accurate detection for shrubs, reaching a peak IoU of 0.9513. The recall for the "Bushes" class is now exceptionally high at 0.96, with only a minor 0.04 misclassification rate into the "Trees" category.

• Forest and Grass Detection: The model correctly classifies 96% of forest pixels (Recall = 0.96), achieving an IoU of 0.8250 for the "Trees" class. For the "Grass" class, a stable IoU value of 0.7846 was achieved, with a true positive rate (recall) of 0.90.

4.7. Generalization Analysis: Spatial Cross-Validation

A key challenge in aerial imagery segmentation is the risk of overfitting on local textures (spatial overfitting). With a random split, the training and test sets may contain overlapping parts of the same objects, leading to a "spatial leakage" effect and inflated metrics.

To confirm the model's generalization ability (its ability to work in new areas), an experiment was conducted with a strict spatial split:

• Training was conducted on the northern part of the campus (80% of the area), characterized by dense development and the presence of sports facilities.

• Testing was carried out on the southern part (20% of the area), which is a park area that was not presented to the network during the training process.

Spatial Split diagram for testing the model's robustness to overfitting is shown in Figure 3.

Figure 3. Spatial Split diagram for testing the model's robustness to overfitting. The upper zone is the training set (Train), the lower zone is the independent test set (Test).

Despite significant differences in the landscape (Domain Shift), the modified model demonstrated an impressive overall accuracy of 0.9351 in the test area. This demonstrates that the developed architecture successfully learned invariant physicochemical class features (the spectral signature of chlorophyll and geometric elevation differences) rather than simply interpolating data from adjacent pixels.

The normalized error matrix (Confusion Matrix) for the proposed model on the test sample is shown in Fig. 4.

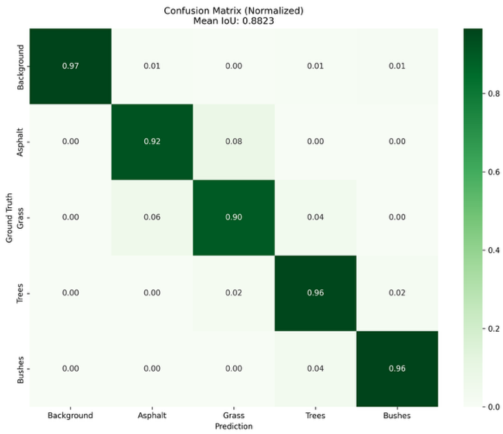


Figure 4. Confusion Matrix for the final model on the test set. The rows show the true classes, and the columns show the classes predicted by the neural network.

5. Discussion

The proposed 7-channel U-Net achieved a mean IoU of 0.8823 for urban green space segmentation, substantially outperforming traditional NDVI thresholding and pixel-based classifiers. This improvement stems from two main strategies. First, early fusion of the DSM channel provides explicit elevation information, allowing the network to disambiguate spectrally similar classes such as low-lying lawns, tall trees, and green roofs; the IoU for man-made objects reached 0.8851. Second, a class-weighted focal loss alleviates severe class imbalance by down-weighting easy background and asphalt pixels, thereby allocating more gradient resources to the rare "shrub" class and raising its recall to 0.96. Spatial cross-validation further verified generalization: when trained solely on the northern built-up area, the model attained an overall accuracy of 0.9351 on the unseen southern park landscape, indicating that it learned transferable features rather than memorizing local textures. For large-scene inference, a Gaussian Blended Sliding Window (GBSW) strategy fuses overlapping patch predictions with Gaussian weights, producing seamless classification maps free of stitching artifacts.

Several limitations of the proposed approach should be noted:

• Dependence on the quality of the DSM. Building boundary accuracy correlates with DSM resolution; at sharp elevation changes (e.g., vertical walls), hardware oversmoothing can cause geometric errors on roof edges — a constraint of the source data rather than the network.

• Semantic ambiguity of the background. The "Background" class achieves high IoU (0.9656) but occasionally contains unclassifiable small objects (temporary structures, cars, debris). Future work should introduce an explicit "Other" (Clutter) class.

6. Conclusion

This paper explored deep learning methods for automated urban vegetation inventory using high-resolution aerial imagery, employing a modified U-Net adapted for 7-band multimodal data (RGB + NIR + RedEdge + DSM + NDVI). The key results are as follows:

• The critical importance of multimodal features was confirmed. Early fusion of spectral indices and elevation data significantly improved feature separability. The DSM enabled the model to learn geometric features and reliably separate spectrally identical Grass and Forest classes, while the NDVI channel eliminated false positives on green roofs and anthropogenic surfaces (asphalt IoU: 0.8851).

• Class imbalance was successfully overcome. Replacing standard cross-entropy with a manually weighted focal loss solved the detection of minority classes, raising the recall of the complex Shrub class to 0.96.

• The algorithm's high generalization ability was proven. A strict spatial split experiment tested resistance to overfitting on local textures; on a completely new park landscape not used in training, the model achieved an overall accuracy of 0.9351, proving it learned physical and chemical reflectance and elevation profiles rather than interpolating training pixel coordinates.

• High overall segmentation quality was achieved. The final Mean IoU of 0.8823 significantly exceeded baseline 3-band models and classic pixel-based methods (e.g., NDVI thresholding). The model produces semantically coherent masks, effectively suppressing salt-and-pepper noise artifacts.

The modified 7-channel U-Net is fully justified for automated urban green space inventory, effectively suppressing noise, distinguishing vegetation layers, and achieving high performance.

References

1. Neyns, R., Canters, F.: Mapping of Urban Vegetation with High-Resolution Remote Sensing: A Review. Remote Sensing 14(4), 1031 (2022)
2. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015)
3. Yang, M.-D., Tseng, H.-H., Hsu, Y.-C., Tsai, H. P.: Semantic Segmentation Using Deep Learning with Vegetation Indices for Rice Lodging Identification in Multi-date UAV Visible Images. Remote Sensing 12(4), 633 (2020)
4. Dimitrovski, I., Spasev, V., Loshkovska, S., Kitanovski, I.: U-Net Ensemble for Enhanced Semantic Segmentation in Remote Sensing Imagery. Remote Sensing 16(12), 2077 (2024). https://doi.org/10.3390/rs16122077. Accessed 10 Feb 2026

5. Chen, K., Wang, Y., Huang, C., *et al.*: GreenNet: A dual-encoder network for urban green space classification using high-resolution remotely sensed images. International Journal of Applied Earth Observation and Geoinformation 142, 104709 (2025)
6. Khan, B. A., Jung, J.-W.: Semantic Segmentation of Aerial Imagery Using U-Net with Self-Attention and Separable Convolutions. Applied Sciences 14(9), 3712 (2024). https://doi.org/10.3390/app14093712. Accessed 12 Mar 2026
7. Li, R., Zhao, J., Fan, Y.: Research on CTSA-DeepLabV3+ Urban Green Space Classification Model Based on GF-2 Images. Sensors 25(13), 3862 (2025)
8. Chen, Y., Yin, T.: Semantic Segmentation of Urban Green Space and Water Body in High-Resolution Remote Sensing Images Based on Low-Rank Adaptation. In: 2025 37th Chinese Control and Decision Conference (CCDC) (2025)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
10. Lee, D.H., Park, H.Y., Lee, J.: A Review on Recent Deep Learning-Based Semantic Segmentation for Urban Greenness Measurement. Sensors (Basel) 24(7), 2245 (2024). https://doi.org/10.3390/s24072245. Accessed 25 Mar 2026
11. Zhang, D., Yan, Y.: Semantic segmentation of multi-source remote sensing image based on multi-path context module. Discov Appl Sci (2026). https://doi.org/10.1007/s42452-026-08338-1. Accessed 08 Apr 2026
12. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
13. Kingma, D. P., Ba, J.: Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR) (2015)
14. Karim, F., Majumdar, S., Darabi, H., *et al.*: Multivariate LSTM-FCNs for time series classification. Neural networks 116, 237-245 (2019)
15. Maninis, K. K., Pont-Tuset, J., Arbeláez, P., *et al.*: Convolutional oriented boundaries: From image segmentation to high-level tasks. IEEE transactions on pattern analysis and machine intelligence 40(4), 819-833 (2017)
16. Dekker, R. J.: Texture analysis and classification of ERS SAR images for map updating of urban areas in the Netherlands. IEEE transactions on geoscience and remote sensing 41(9), 1950-1958 (2003)
17. Arouri, Y., Sayyafzadeh, M.: An adaptive moment estimation framework for well placement optimization. Computational Geosciences 26(4), 957-973 (2022)
18. Neupane, B., Horanont, T., Aryal, J.: Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis. Remote Sensing 13(4), 808 (2021)
19. Zhang, C., Sargent, I., Pan, X., *et al.*: Joint Deep Learning for land cover and land use classification. Remote sensing of environment 221, 173-187 (2019)
20. Shamsolmoali, P., Zareapoor, M., Wang, R., *et al.*: A novel deep structure U-Net for sea-land segmentation in remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12(9), 3219-3232 (2019)
21. Ho, Y., Wookey, S.: The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. IEEE access 8, 4806-4813 (2019)