Fine-grained Vegetation Segmentation in Complex Urban Park Environments Using a Deeply Supervised Parallel SegFormer

Haixin Zhang[1], Qinying Zhang[1]

[1] Department of Landscape Architecture, Tianjin University, 300072 Tianjin, China
E-mail: haixin_zhang@tju.edu.cn, qinying_zhang@tju.edu.cn

Keywords: UAV imagery, RGB images, vegetation species, semantic segmentation, urban park, SegFormer.

Abstract

Accurate vegetation mapping in complex urban environments is essential for ecological monitoring, biodiversity assessment, and sustainable park management. However, fine-grained vegetation segmentation remains challenging because of the high diversity of plant species, overlapping canopies, and the interference of artificial objects. To address these challenges, a deeply supervised parallel architecture based on the SegFormer backbone was proposed in this paper. The model incorporated a SegFormer-ASPP-low-level (SAL) head, which fused high-level semantic representations, multi-scale contextual information, and low-level spatial details through a parallel decoding mechanism. Two auxiliary heads, a pyramid pooling module (PSP) and a fully convolutional network (FCN), were added to provide deep supervision and improve the recognition of blurred boundaries and rare categories. High-resolution UAV imagery was used to perform fine-grained semantic segmentation of 17 vegetation categories. The dataset included multiple tree species as well as non-tree classes such as *Nelumbo* sp. (lotus) and dead trees. Experimental results showed that our model achieved a mean intersection over union (mIoU) of 73.57%, outperforming architectures such as SegFormer-b1, DeepLab v3+, ConvNeXt and SCTNet. Visual analysis further demonstrated the model's robustness in complex urban park scenes, showing superior boundary delineation, improved recognition of small and spectrally similar species, and resilience to interference from artificial objects like plastic lawns and landscape lighting. The proposed approach offers valuable insights for precision forestry, ecological monitoring, and intelligent UAV-based remote sensing applications.

1. Introduction

Urban green spaces, particularly parks, play a vital role in enhancing the ecological quality, social well-being, and aesthetic value of cities (Nan et al., 2021). The internal green-space area and landscape shape index of parks are significantly correlated with the key factors that maintain low land surface temperatures and mitigate the urban heat island effect (Cai et al., 2023). The precise identification and spatial mapping of vegetation within these areas are fundamental for effective urban forest management, biodiversity monitoring, and the assessment of urban ecosystem services. Large-scale spatial information on vegetation can improve the understanding of forest biomass assessment (Zurqani, 2025) and carbon-sink assessment (Chakraborty et al., 2024), as well as assessments of wildlife habitat (Mitchell and Devisscher, 2022) and insect abundance (Li et al., 2025). With the extension of urban park construction time and the enhancement of human disturbance, the competition between designed plants and spontaneous species and the management frequency jointly determine the dynamic evolution of community structure (Chang et al., 2021), and the vegetation community often gradually deviates from the original planting design.

Traditional methods for urban tree inventory rely heavily on manual field surveys. While accurate, these ground-based approaches are often time-consuming, labor-intensive, and expensive, making them impractical for frequent, large-scale monitoring of dynamic urban vegetation. Therefore, establishing a high-frequency quantitative vegetation monitoring system is of great significance for achieving precise vegetation management (Shahtahmassebi et al., 2021).

In recent years, remote sensing technologies, especially those based on consumer-grade unmanned aerial vehicles (UAVs), have emerged as a powerful and cost-effective alternative for urban forestry applications. These platforms make frequent quantitative vegetation monitoring practically feasible as they are low-cost and easily replicated (Simon et al., 2022). UAVs offer the flexibility to acquire high-resolution imagery, capturing intricate details of tree crown structure, texture, and color from RGB optical sensors. This level of detail enables fine-grained, individual-tree classification, which remains challenging for satellite or conventional aerial imagery (Qin et al., 2022). Drone remote sensing was used to improve urban forest conservation by identifying relationships between key forest health indicators and four vegetative indices: the normalized difference vegetation index (NDVI), normalized difference red edge (NDRE), leaf chlorophyll index (LCI), and the green normalized difference vegetation index (GNDVI) (Wavrek et al., 2023). UAV images have been used for the early identification of pests such as pine wilt disease (Li and Wang, 2025) and pine shoot beetle (Wang et al., 2025a), as well as for predicting future tree deaths (O. and Vanderwel, 2022).

With the rapid advancement of artificial intelligence, deep learning–based semantic segmentation has become the most advanced approach for pixel-level image classification in remote sensing and urban ecology (Lv et al., 2023a). Unlike traditional manual feature extraction methods, these architectures automatically learn robust, generalizable representations directly from raw images (Dosovitskiy et al., 2021). Encoder–decoder frameworks such as U-Net and DeepLab improve vegetation mapping accuracy in heterogeneous landscapes (Su et al., 2022). UAV-based platforms integrated with deep learning offer an efficient and cost-effective solution for large-scale ecological monitoring (Osco et al., 2021). Intelligent cloud attention networks have further improved global urban green-space mapping performance (Chen et al., 2023). This fusion of automated deep learn-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-313-2026

ing analysis with low-cost, high-resolution UAV data presents a highly scalable and accessible solution, offering practical value for routine monitoring by municipal management departments, particularly those operating with limited budgets.

However, several challenges persist in the semantic segmentation of vegetation species in complex urban park environments. Urban forests are often characterized by high species diversity and complex canopy structures, where variations in illumination, tree age, and health lead to high intra-class variance, while similarities in crown morphology among different species result in low inter-class variance (Zheng et al., 2025). In addition, even with ultra-high resolution UAV images and deep learning networks, the segmentation accuracy is significantly degraded in dense urban forest environments due to structural and spectral complexities such as occlusion caused by canopy overlap, shadow interference, and mixed pixels at the crown boundary (Lv et al., 2023b).

To address the challenge of fine segmentation of urban vegetation species in high-resolution UAV imagery, a SegFormer-b1 baseline was enhanced with a parallel-fusion decoder that concurrently aggregated SegFormer-encoded representations, atrous spatial pyramid contexts, and low-level spatial information. Auxiliary supervision was provided by complementary FCN and PSP heads operating in parallel with the main decoder, promoting multi-scale feature refinement and boundary localization accuracy. A comprehensive experiment was conducted on UAV RGB imagery acquired over Nancuiping Park, an artificial mountainous park in Tianjin. In this paper, SegFormer-b1, DeepLab v3+, ConvNeXt and SCTNet models were selected for comparative experiments. The results showed that the proposed method achieved leading performance on various evaluation indicators. In complex urban scenes with artificial green interference, the proposed model maintains stable segmentation performance even under extreme class imbalance. The results demonstrate the effectiveness of the proposed method for fine-grained vegetation segmentation in complex urban park scenes.

2. Study Area and Dataset

2.1 Study Area

This study was conducted on Nancuiping Park (39°4’31” N, 117°8’53” E), an urban park located in Tianjin, China, which is shown in Figure 1. Nancuiping Park covers an area of 398,600 square meters, with a main peak of 52 meters high and a mountain area of 121,000 square meters. Its core is a large man-made mountain forest built on construction waste dumps, providing citizens with a unique urban mountain landscape and an important source of carbon sinks. The park is located in a warm-temperate, subhumid continental monsoon climate zone, and its vegetation is dominated by cold- and drought-tolerant temperate deciduous broad-leaved tree species. Landscape types include sparse forest grassland, gentle slope woodland, steep slope woodland, aquatic plants and other types, forming a green ecological barrier with distinct four seasons.

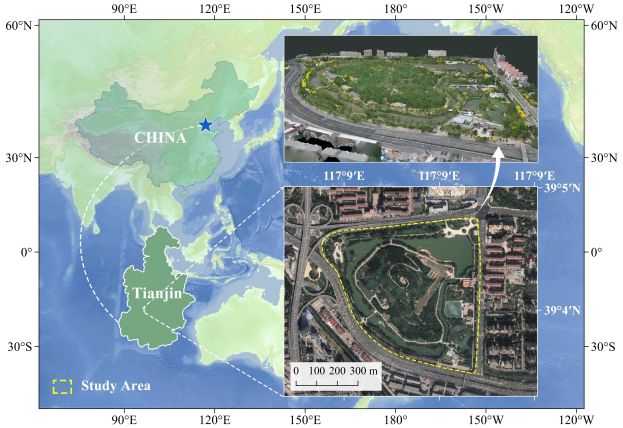

Figure 1. Study area

2.2 Data Acquisition

In this study, RGB images were acquired using a consumer-grade DJI Mini 3 Pro UAV equipped with a 1/1.3-inch f/1.7 CMOS sensor (up to 48 MP). The flight mission was conducted on 23 May 2025 at 106 m altitude. To minimize shadow effects and ensure consistent illumination, images were collected under cloudy conditions between 11:00 a.m. and 1:00 p.m.

The flight route was manually planned to ensure full coverage of the park, with images captured every 5 s at a flight speed of 8–10 m/s. In total, 315 images were collected, and an orthomosaic with a ground sampling distance of 3.826 cm/px was generated using DJI Terra software. Professionals with botanical expertise used Labelme software to manually delineate precise tree crown boundaries and assign species labels at the pixel level on 512 × 512 pixel cropped images. A total of 1220 annotated samples were used. The dataset contained 17 classes,

Species Name	Image	Pixel Ratio	Species Name	Image	Pixel Ratio
Melia azedarach		5.43 %	*Fraxinus chinensis*		4.02 %
Rhus typhina		16.43 %	*Cotinus coggygria*		0.72 %
Prunus cerasifera ’Atropurpurea’		0.89 %	*Sophora japonica* ’Golden Stem’		6.55 %
Koelreuteria paniculata		4.73 %	*Nelumbo* sp.		5.17 %
Ulmus pumila		4.74 %	*Broussonetia papyrifera*		2.38 %
Platycladus orientalis		10.22 %	*Juniperus formosana*		2.81 %
Populus tomentosa		2.65 %	Dead trees		1.21 %
Salix babylonica		13.51 %	*Ailanthus altissima*		3.50 %
Styphnolobium japonicum		14.99 %			

Table 1. Dataset

https://doi.org/10.5194/isprs-annals-XI-1-2026-313-2026

as shown in Table 1, including 13 common tree and shrub species, aquatic vegetation (*Nelumbo* sp.), a non-flowering species (*Platycladus orientalis*), and dead trees. Among them, *Rhus typhina* accounts for the largest proportion, at 16.43%, whereas *Cotinus coggygria* accounts for the smallest, at 0.72%.The artificial design of the park green space results in extreme class imbalance and similar imagery textures among some species, such as *Melia azedarach* - *Rhus typhina* - *Ailanthus altissima*, *Styphnolobium japonicum* - *Fraxinus chinensis*, which brings great challenges to the accurate segmentation and distinction.

3. Method

3.1 Data Preparation and Training Pipeline

Figure 2 illustrates the overall workflow of data acquisition, dataset construction, and model training in this study. UAV imagery is first collected using a consumer-grade platform and processed into annotated image patches, which are then used to train the proposed SAL SegFormer for patch-wise vegetation segmentation. In this study, all labeled data was randomly divided into training, validation, and test sets in an 8:1:1 ratio. In addition, some typical images of the park green area without annotation were selected as a supplementary test set to test the specific performance of the model in the urban park scene.

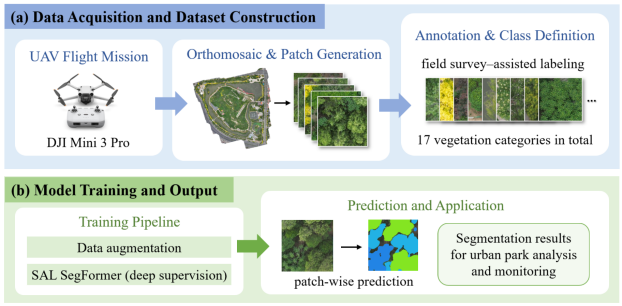

Figure 2. Workflow

During the data loading stage, online data augmentation strategies were used to enhance training diversity and reduce overfitting. The training images were first subjected to RandomResize, where the scaling factor varied dynamically between 0.5 times and 2.0 times the original size to simulate the scale differences under different flight altitudes. Subsequently, patches were randomly cropped from the resized images for training. To improve robustness and reduce overfitting, RandomFlip was also applied with a probability of 50% and PhotoMetricDistortion, which includes random adjustments to brightness, contrast, and saturation to simulate the complex lighting variations in the real world. Before being sent to the network, all image blocks underwent standardized processing through the data preprocessor. They were first converted from BGR format to RGB format, and then normalized according to the mean values and standard deviations of the ImageNet dataset.

A relatively simple preprocessing procedure was adopted for the validation and test sets. After loading, these images were only uniformly resized and then directly sent to the model for evaluation.

3.2 Four Comparison Models

3.2.1 DeepLab v3+ DeepLab v3 is a semantic segmentation model proposed by Google Research in 2017 (Chen et al., 2017). It integrates atrous convolution with an improved atrous spatial pyramid pooling (ASPP) module to capture multi-scale contextual information while preserving spatial resolution. DeepLab v3+ further introduced a learnable encoder–decoder structure for improved boundary localization and segmentation accuracy (Chen et al., 2018). Both models have shown strong performance in computer vision, especially in forestry remote sensing, where DeepLab v3+ has been used for deforestation and forest-fire monitoring (Joe et al., 2025) and automatic single-tree segmentation in forests (Shi et al., 2022).

3.2.2 SegFormer SegFormer is a transformer-based semantic segmentation architecture proposed in 2021 (Xie et al., 2021). It employs a hierarchical Mix Vision Transformer encoder to extract multi-scale features, together with a lightweight decoder that fuses features using multilayer perceptrons, achieving a favorable balance between accuracy and efficiency. In UAV remote sensing, SegFormer has demonstrated strong performance in object and land-cover delineation across diverse scenes (Spasev et al., 2024). Several variants have further extended its applicability, such as MTLSegFormer for multi-task agricultural segmentation (Goncalves et al., 2023) and ECA-SegFormer for plant disease spot segmentation under natural conditions (Yang et al., 2023).

3.2.3 ConvNeXt ConvNeXt is a pure convolutional architecture proposed in 2022 (Liu et al., 2022), which modernizes CNN design and achieves performance comparable to vision transformers. By employing large-kernel convolutions, inverted bottleneck structures, and simplified activation functions, ConvNeXt enhances receptive fields and feature representation capacity. Based on this backbone, AAUConvNeXt was used to improve crop-tilting segmentation accuracy (Zhang et al., 2024), while CMPF-UNet and ConvNeXt-U applied multi-scale fusion strategies for multi-class remote sensing segmentation and fragmented cropland extraction, respectively (Ning et al., 2024, Liu et al., 2025).

3.2.4 SCTNet Spatial Context Transformer Network (SCTNet) is a hybrid CNN–Transformer architecture designed for semantic segmentation tasks (Xu et al., 2023). It adopts a dual-branch structure to capture local spatial details and long-range contextual dependencies, and introduces a Bidirectional Fusion Module and a Scale Context Transformer to integrate multi-scale features. SCTNet has been applied to agricultural vision tasks such as eggplant segmentation (Wang et al., 2025b) and nighttime pineapple detection (Wu et al., 2024), demonstrating its effectiveness in complex visual environments.

3.3 Our Improved SegFormer Network

3.3.1 Overall Architecture This study adopted the Mix Vision Transformer (MiT-B1) as the semantic segmentation backbone, as its efficient self-attention mechanism provides strong capability for capturing multi-scale global context. However, the original SegFormer design pairs this powerful encoder with a relatively simple decoder that relies only on basic multi-scale feature upsampling, concatenation, and linear fusion. Targets with large scale variation and irregular boundaries still challenge contextual aggregation and low-level detail utilization.

To address these issues, an improved, deeply supervised parallel fusion decoder was proposed, and its overall structure is illustrated in Figure 3. The architecture consists of the customized parallel main decoder, SegFormer-ASPP-lowlevel (SAL) head, and two auxiliary decoders, PSP head and FCN head.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-313-2026 | © Author(s) 2026. CC BY 4.0 License.

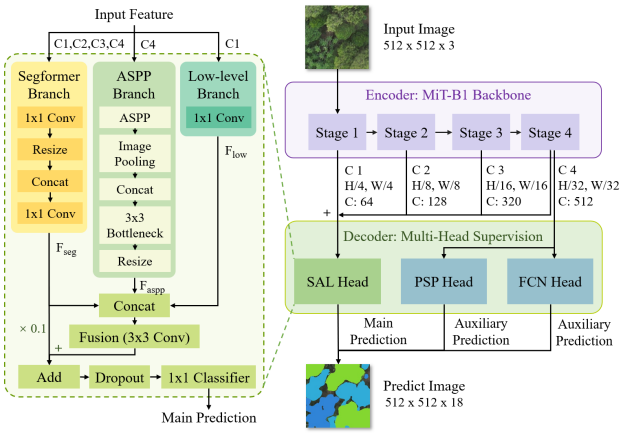

Figure 3. Architecture of the proposed SegFormer framework

The SAL Head is designed to fuse high-level semantics, multi-scale context, and fine spatial details through a parallel decoding structure. It consists of three branches: SegFormer, ASPP, and a low-level detail branch. Feature maps from the four backbone stages (C1, C2, C3, C4) are first projected with 1×1 convolutions, uniformly upsampled to 1/4 scale, concatenated, and fused by another 1×1 convolution to preserve the global multi-scale features of SegFormer. To enhance multi-scale context modeling, the ASPP branch adopts the idea of DeepLab v3+ and applies atrous spatial pyramid pooling only to the highest-level feature map (C4). Using dilation rates of 1, 6, 12, and 18 together with an image-pooling branch, it captures contextual information at multiple scales and fuses the features through a 3×3 convolution before upsampling to 1/4 scale. To compensate for the spatial detail lost during pooling and downsampling, the low-level branch directly extracts texture information from the shallowest, highest-resolution feature map (C1) through a 1×1 convolution.

Finally, outputs of the three branches are concatenated along the channel dimension and fused by a 3×3 convolution. To stabilize training and preserve the contribution of the original SegFormer branch, a residual connection is introduced, followed by Dropout and a 1×1 classifier to generate final prediction results.

3.3.2 Loss function To address the limitations of a single region-based loss, such as cross-entropy or Dice loss, in handling class imbalance and fuzzy boundaries, a hybrid loss function was employed. The total training objective consists of a main loss L_{main} and two auxiliary losses, $L_{\text{aux}}^{\text{FCN}}$ and $L_{\text{aux}}^{\text{PSP}}$, combined as follows to reinforce supervision for minority classes:

$$L_{\text{total}} = L_{\text{main}} + \lambda_1 L_{\text{aux}}^{\text{FCN}} + \lambda_2 L_{\text{aux}}^{\text{PSP}} \tag{1}$$

where λ_1 and λ_2 are the balance weights of the auxiliary heads. The main loss L_{main} is a three-part composite loss designed to optimize the model at the pixel, region, and boundary levels simultaneously:

$$L_{\text{main}} = w_1 L_{\text{ce}} + w_2 L_{\text{Dice}} + w_3 L_{\text{boundary}} \tag{2}$$

where L_{ce} is the cross-entropy loss, L_{Dice} is the Dice loss, and L_{boundary} is the boundary loss defined as:

$$L_{\text{boundary}} = \frac{1}{C} \sum_{i=1}^{C} L_{\text{Dice}}(\partial P_i, \partial G_i) \tag{3}$$

where C is the number of classes, and ∂P_i and ∂G_i denote the predicted and ground-truth boundary maps of class i, respectively.

To deal with dataset specific challenges while balancing multiple optimization objectives, the loss configuration is shown in Table 2. Among them, L_{ce}, $L_{\text{aux\_FCN}}$ and $L_{\text{aux\_PSP}}$ followed the classical configuration of their paper. To address extreme class imbalance, the weight of L_{Dice} was slightly increased to 1.2, and L_{boundary} was set to 0.15 in the overall loss function.

Component	Weight	Value	Reference
Cross Entropy	w_1	1.0	(Xie et al., 2021)
Dice Loss	w_2	1.2	(Fausto et al., 2016)
Boundary Loss	w_3	0.15	(Hoel et al., 2019)
FCN Auxiliary Head	λ_1	0.25	(Zhao et al., 2017)
PSP Auxiliary Head	λ_2	0.15	(Zhao et al., 2017)

Table 2. The loss weight configuration

4. Results and Discussion

4.1 Experimental Setup and Evaluation Metrics

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU (24 GB). The models were implemented in PyTorch 1.7.1 (CUDA 11.0) using MMSegmentation 1.2.2. DeepLab v3+ and ConvNeXt used ResNet-101-V1c and ConvNeXt-Base backbones, respectively, both initialized with ImageNet-1K pretrained weights, while SCTNet used SCTNet-Base with its domain-specific pretrained model. Training was performed with the AdamW optimizer, using an initial learning rate of 1e-4, a weight decay of 1e-2, a batch size of 8, and a maximum of 200 epochs. A cosine annealing schedule decayed the learning rate to 1e-6 over 200 cycles. To reduce overfitting and training time, an early stopping hook was applied, terminating training when the validation mIoU failed to improve by 0.01% for 50 consecutive epochs. The same early stopping configuration was used in all experiments.

Model performance was primarily evaluated by mean intersection over union (mIoU). Mean pixel accuracy (mAcc) and mean F-score (mF-score) were also reported to provide a more detailed assessment of class-level segmentation performance, as defined in equations (4)–(8). Models were evaluated on the validation set at each epoch, and the checkpoint with the best validation performance was selected for final testing.

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{6}$$

$$mAcc = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i} \tag{7}$$

$$F\text{-}score = \frac{2 \times P \times R}{P + R} \tag{8}$$

where P = Precision;
R = Recall;
TP = True positive;
FP = False positive;
FN = False negative;
$mAcc$ = Mean pixel accuracy;

N = Number of classes.

4.2 Analysis of the Vegetation Segmentation Results

In terms of model complexity, SegFormer-b1 had the least number of parameters at 55.49 M, reflecting its lightweight design concept, as shown in Table 3. The number of parameters in our model was 99.39 M, which was greater than SegFormer-b1, but still significantly lower than DeepLab v3+ (232.71M) and ConvNeXt (469.86 M). This shows that our semantic image segmentation network effectively improved the segmentation accuracy while moderately increasing the model complexity, and achieved a good trade-off between performance and efficiency.

Methods	mIoU	mF-score	mAcc	Parameters(M)
DeepLab v3+	63.11	76.11	76.46	232.71
SegFormer-b1	71.08	82.53	81.05	**55.49**
ConvNeXt	62.31	79.81	74.65	469.86
SCTNet	68.02	80.22	78.63	201.34
Ours	**73.57**	**84.03**	**84.38**	99.39

Table 3. Performance comparison on the test dataset (%)

In terms of model performance, our model achieved 73.57% in the core metric mIoU, which was 2.49% higher than the baseline model SegFormer-b1 (71.08%), and was significantly better than DeepLab v3+ (63.11%), ConvNeXt (62.31%) and SCTNet (68.02%). Similar performance advantages were also reflected in the mean F-score and mean pixel accuracy: our model achieved 84.03% and 84.38%, respectively, which were ahead of all comparison methods.

ID	DeepLab v3+	SegFormer -b1	ConvNeXt	SCTNet	Ours
Me. A.	66.67	66.62	65.66	**69.87**	67.14
Rh. T.	68.08	74.90	63.95	66.79	**75.20**
Pr. C.	33.38	**74.57**	47.21	55.26	59.27
Ko. P.	43.06	**56.03**	44.25	49.36	51.23
Ul. P.	67.02	**78.40**	61.47	68.33	77.67
Pl. O.	69.58	**78.52**	64.57	73.56	75.83
Po. T.	60.88	68.54	62.49	67.63	**76.81**
Sa. B.	81.81	**88.43**	82.28	85.70	86.84
St. J.	56.77	61.91	59.40	65.30	**68.96**
Fr. C.	38.58	37.94	33.05	41.46	**54.36**
Co. C.	79.72	85.73	89.67	85.91	**89.90**
So. J.	69.56	79.14	76.12	70.34	**79.29**
Ne. S.	93.68	**95.77**	93.44	95.41	95.67
Br. P.	69.15	83.54	59.29	86.33	**89.08**
Ju. F.	55.12	61.15	38.83	**66.55**	65.15
De. T	49.36	65.97	57.01	37.65	**66.33**
Ai. A.	66.81	71.74	56.49	69.74	**71.88**

Table 4. Class-wise IoU comparison (%)

Given the long-tail class distribution, Table 4 presents class-wise IoU comparisons to analyze model performance on rare categories. The table clearly reveals that in the dominant classes with high proportions such as *Rhus typhina*, *Salix babylonica*, and *Styphnolobium japonicum*, all models performed very close, with the difference between the highest and lowest mIoU ranges from 6.62% to 12.19%. However, the core advantage of the proposed model was reflected in its robustness to the segmentation of difficult categories with medium and low proportion. For example, for *Fraxinus chinensis* (4.02% pixels), the IoU of SegFormer-b1 was only 37.94%, while the proposed model achieved 54.36%. Similarly, for *Broussonetia papyrifera* (2.38% pixels), the proposed model (89.08%) also had an improvement of 5.5% over SegFormer-b1 (83.54%). Although SegFormer-b1 performs well on the extremely rare category *Prunus cerasifera* (0.89% pixels), our model stably and significantly improves performance on most medium and low frequency categories. The proposed deep supervision mechanism, composed of the SAL head parallel decoder and the FCN and PSP auxiliary heads, proved to be broadly effective in alleviating the impact of class imbalance and improving the segmentation accuracy.

4.3 Ablation experiments

Ablation experiments were conducted to verify each component's effectiveness, and the results were shown in Table 5. The baseline mIoU of SegFormer-b1 was 71.08%. The introduction of boundary loss $L_{boundary}$ alone improved the mIoU to 72.29%, which initially proved the effectiveness of the enhanced boundary supervision. Furthermore, FCN auxiliary head and PSP auxiliary head were added in turn for depth supervision, and the mIoU increased to 72.75% and 73.15% respectively. Finally, by replacing the baseline decoder with our proposed parallel SAL head, the complete model achieved the highest mIoU of 73.57%. This final 0.42% increase from the fully supervised baseline, combined with the earlier gains, resulted in a total improvement of 2.49% over the original SegFormer-b1. The experimental results show that all components, including the boundary loss, the dual auxiliary heads, and our SAL head architecture, bring positive gains to the model performance.

Methods	mIoU	mF-score	mAcc
Baseline	71.08	82.53	81.05
Baseline + L_{boundary}	72.29	83.17	81.70
Baseline + L_{boundary} + FCN	72.75	83.64	81.67
Baseline + L_{boundary} + FCN + PSP	73.15	83.83	82.32
SAL + L_{boundary} + FCN + PSP	**73.57**	**84.03**	**84.38**

Table 5. Results of model component ablation experiments (%)

4.4 Visual analysis

Figure 4 shows the segmentation results of each model on partial test dataset samples compared with the true annotations. Compared to other models, the segmentation boundaries generated by our model were generally smoother and more closely fit the contour of the real crown, especially when dealing with crowns with irregular shapes or edges that touch other objects, such as the full crown range of a sparse *Salix babylonica* in (a). Although the SegFormer-b1 performed better overall, it failed to successfully distinguish some similar tree species, resulting in misclassification and omissions, as shown in (c). For some tree species with small size or special morphology, such as the red-labeled trees in (d), our model also identified and delineated relatively accurately, showing good robustness.

Segmentation was further performed on an additional set of unannotated images of urban green spaces, as shown in Figure 5, to evaluate model performance in the presence of various non-vegetation objects commonly found in urban scenes. Faced with artificial debris piles, cars, and sidewalks, our model showed strong discrimination ability and effectively excluded

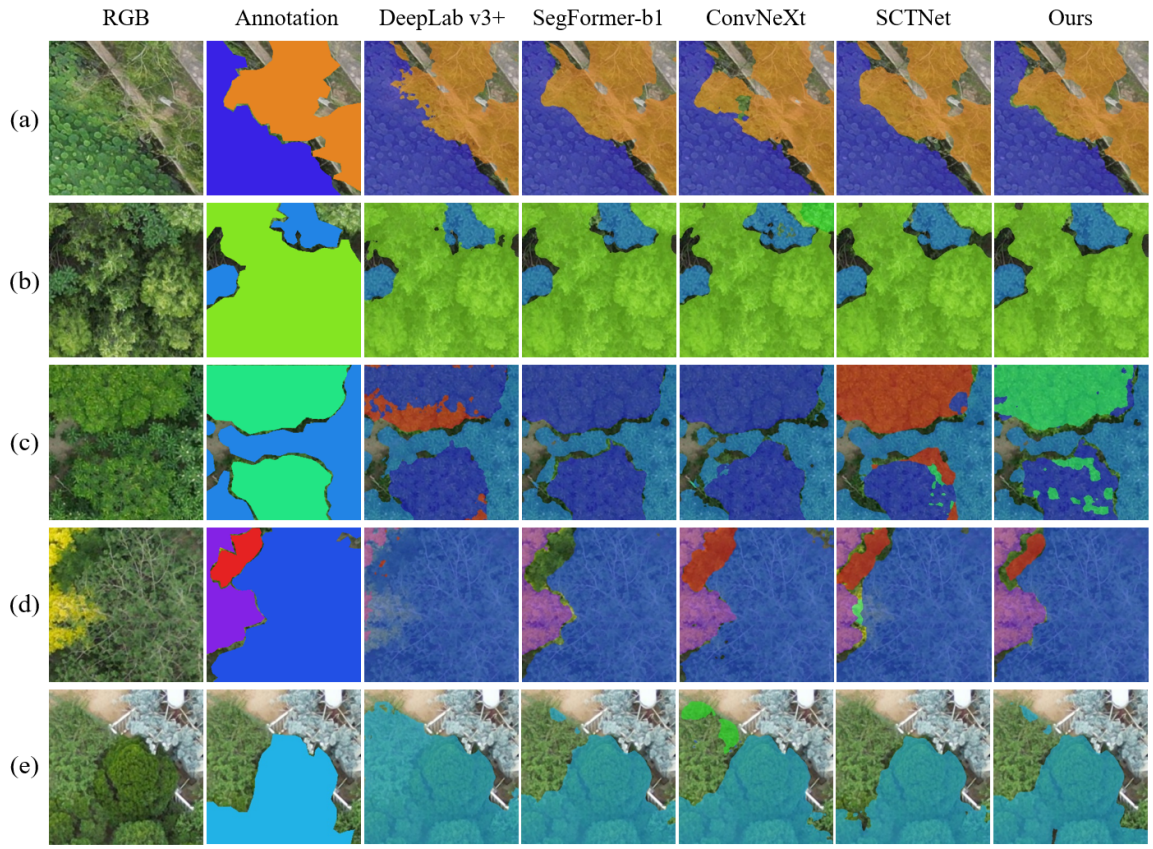

Figure 4. Comparison of segmentation results of different models

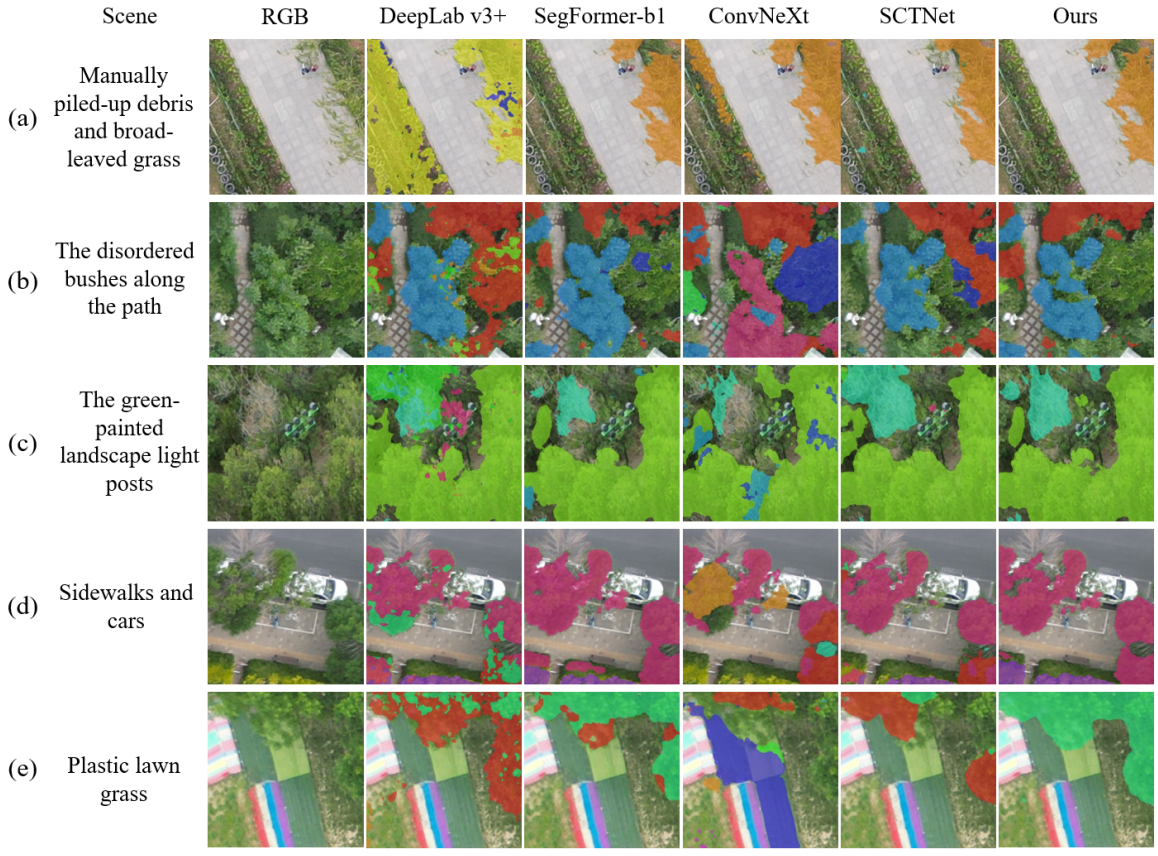

Figure 5. Segmentation results under challenging urban-park scenarios in the new dataset

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-313-2026

them from vegetation, reducing false positives. It also demonstrated a greater ability to distinguish tall trees from easily confused low-rise vegetation, such as the broad-leaved grass in (a) and the disordered bushes in (b), producing results that excluded these non-tree regions more accurately than models such as ConvNeXt. Common park facilities with vegetation-like colors, such as green landscape lamp posts in (c) and plastic fake lawns in (e), were also correctly ignored without being misclassified as trees. Importantly, the improved robustness against artificial green objects supports more reliable urban-park inventory, benefiting downstream carbon-sink estimation and biodiversity monitoring.

4.5 Limitations and Future Work

This study assumes that high-resolution RGB UAV imagery can provide sufficient spectral and structural cues for fine-grained vegetation segmentation in urban parks. However, the current dataset comes mainly from a specific park and season. Consequently, the model’s generalizability under different lighting conditions, seasonal states, or urban park environments needs further verification. Future research should therefore focus on extending cross-seasonal and cross-site experiments or exploring domain adaptation techniques to evaluate the model’s robustness in multi-temporal smart-forest monitoring tasks.

Furthermore, although the low-level detail branching helps, room for improvement remains in segmentation accuracy, particularly in challenging situations. These include severe occlusions, tiny tree crowns, shadows caused by drastic illumination changes, and extremely fine branch or crown edges. These challenges point the way for further architectural optimization.

5. Conclusion

To solve the challenges of fine-grained vegetation inventory in complex urban parks, a deeply supervised semantic segmentation model based on improved SegFormer was proposed. The Mix Vision Transformer was used as the encoder backbone to extract multi-scale features from high-resolution UAV images. A SAL head was used as a parallel fusion decoder, which integrates three key branches in parallel: (1) a SegFormer branch that processes all four-stage features to preserve global context; (2) a branch that operates specifically on the deepest features and captures multi-scale contextual semantics through ASPP modules; (3) a branch that directly exploits the shallowest features to enhance boundaries and texture details. These three feature streams are effectively fused in the decoding stage and further refined through a residual connection. In addition, a deep-supervision mechanism was introduced, in which the FCN head and PSP head were attached to the deep encoder features as auxiliary heads to provide additional supervision signals during training and promote the backbone network to learn more discriminative feature representations.

Experiments were conducted on a UAV image dataset from the Nancuiping man-made mountain urban park in Tianjin, and the proposed model was systematically compared with mainstream semantic segmentation methods, including SegFormer-b1, DeepLab v3+, ConvNeXt, and SCTNet. The proposed approach achieved the best overall accuracy and demonstrated improved capability in crown-boundary delineation, dense canopy separation, and robustness to complex urban interference. By reducing confusion caused by artificial green infrastructure, it supports routine urban ecological assessment and enables automated, low-cost, high-precision green-space management.

References

Cai, X., Yang, J., Zhang, Y., Xiao, X., Xia, J. C., 2023. Cooling island effect in urban parks from the perspective of internal park landscape. *Humanities and Social Sciences Communications*, 10, 674.

Chakraborty, D., Ciceu, A., Ballian, D., Garzón, M. B., Bolte, A., Bozic, G., Buchacher, R., Čepl, J., Cremer, E., Ducousso, A., Gaviria, J., George, J. P., Hardtke, A., Ivankovic, M., Klisz, M., Kowalczyk, J., Kremer, A., Lstibůrek, M., Longauer, R., Mihai, G., Nagy, L., Petkova, K., Popov, E., Schirmer, R., Skrøppa, T., Solvin, T. M., Steffenrem, A., Stejskal, J., Stojnic, S., Volmer, K., Schueler, S., 2024. Assisted tree migration can preserve the European forest carbon sink under climate change. *Nature Climate Change*, 14, 845–852.

Chang, C.-R., Chen, M.-C., Su, M.-H., 2021. Natural versus human drivers of plant diversity in urban parks and the anthropogenic species-area hypotheses. *Landscape and Urban Planning*, 208, 104023.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision*, 801-818.

Chen, Y., Weng, Q., Tang, L., Wang, L., Xing, H., Liu, Q., 2023. Developing an intelligent cloud attention network to support global urban green spaces mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 198, 197–209.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Fausto, M., Nassir, N., Seyed-Ahmad, A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, 565–571.

Goncalves, D. N., Junior, J. M., Zamboni, P., Pistori, H., Li, J., Nogueira, K., Gonçalves, W. N., 2023. Mtlsegformer: Multitask learning with transformers for semantic segmentation in precision agriculture. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6290–6298.

Hoel, K., Jihene, B., Christian, D., Eric, G., Jose, D., Ben, A. I., 2019. Boundary loss for highly unbalanced segmentation. *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019*, 285–296.

Joe, S., Kalukin, A., Malof, J., Xu, D., 2025. Deep Learning-Driven Multi-Temporal Detection: Leveraging DeeplabV3+/ Efficientnet-B08 Semantic Segmentation for Deforestation and Forest Fire Detection. *Remote Sensing*, 17(14), 2333.

Li, X., Wang, A., 2025. Forest pest monitoring and early warning using UAV remote sensing and computer vision techniques. *Scientific Reports*, 15, 401.

Li, Y., Schuldt, A., Bauhus, J., Belluau, M., Berthelot, S., Burghardt, K. T., Bruelheide, H., Castagneyrol, B., Chu, C., Eisenhauer, N., Ferlian, O., Fründ, J., Gebauer, T., Gravel, D., Jactel, H., Li, S., Liang, Y., Parker, J. D., Parker, W. C., Scherer-Lorenzen, M., Staab, M., Verheyen, K., Schmid, B., Ma, K., Liu, X., 2025. The tree growth–herbivory relationship depends on functional traits across forest biodiversity experiments. *Nature Ecology and Evolution*.

Liu, S., Shi, C., Xia, L., Peng, J., Ping, L., Fan, X., Teng, F., Liu, X., 2025. Lightweight Deep Learning Model, ConvNeXt-U: An Improved U-Net Network for Extracting Cropland in Complex Landscapes from Gaofen-2 Images. *Sensors*, 25(1), 261.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A ConvNet for the 2020s. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11966-11976.

Lv, J., Shen, Q., Lv, M., Li, Y., Shi, L., Zhang, P., 2023a. Deep learning–based semantic segmentation of remote sensing images: a review. *Frontiers in Ecology and Evolution*, 11, 1201125.

Lv, L., Li, X., Mao, F., Zhou, L., Xuan, J., Zhao, Y., Yu, J., Song, M., Huang, L., Du, H., 2023b. A Deep Learning Network for Individual Tree Segmentation in UAV Images with a Coupled CSPNet and Attention Mechanism. *Remote Sensing*, 15(18), 4420.

Mitchell, M. G., Devisscher, T., 2022. Strong relationships between urbanization, landscape structure, and ecosystem service multifunctionality in urban forest fragments. *Landscape and Urban Planning*, 228, 104548.

Nan, C., Malleson, N., Houlden, V., Comber, A., 2021. Using VGI and social media data to understand urban green space: a narrative literature review. *ISPRS International Journal of Geo-Information*, 10(7), 425.

Ning, L., Yu, X., Yu, M., 2024. CMPF-UNet: a ConvNeXt multi-scale pyramid fusion U-shaped network for multi-category segmentation of remote sensing images. *Geocarto International*, 39(1), 2311217.

O., B. K., Vanderwel, M. C., 2022. Predicting Tree Mortality Using Spectral Indices Derived from Multispectral UAV Imagery. *Remote Sensing*, 14(9), 2195.

Osco, L. P., Junior, J. M., Ramos, A. P. M., de Castro Jorge, L. A., 2021. A review on deep learning in UAV remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102456.

Qin, H., Zhou, W., Yao, Y., Wang, W., 2022. Individual tree segmentation and tree species classification in subtropical broadleaf forests using UAV-based LiDAR, hyperspectral, and ultrahigh-resolution RGB data. *Remote sensing of environment*, 280, 113143.

Shahtahmassebi, A. R., Li, C., Fan, Y., Wu, Y., lin, Y., Gan, M., Wang, K., Malik, A., Blackburn, G. A., 2021. Remote sensing of urban green spaces: A review. *Urban Forestry and Urban Greening*, 57, 126946.

Shi, L., Wang, G., Mo, L., Yi, X., Wu, X., Wu, P., 2022. Automatic Segmentation of Standing Trees from Forest Images Based on Deep Learning. *Sensors*, 22(17), 6663.

Simon, E., Dempewolf, J., Frey, J., Schwaller, A., Endres, E., Klemmt, H.-J., Tiede, D., Seifert, T., 2022. UAV-Based Forest Health Monitoring: A Systematic Review. *Remote Sensing*, 14(13), 3205.

Spasev, V., Dimitrovski, I., Chorbev, I., Kitanovski, I., 2024. Semantic segmentation of unmanned aerial vehicle remote sensing images using segformer. *International Conference on Intelligent Systems and Pattern Recognition*, 108–122.

Su, Y., Jian, C., Bai, H., Liu, H., He, C., 2022. Semantic Segmentation of Very-High-Resolution Remote Sensing Images via Deep Multi-Feature Learning. *Remote Sensing*, 14(3), 533.

Wang, L., Yang, G., Liu, Y., Zhong, L., Wang, S., Ma, Y., Zhan, Z., 2025a. Monitoring Pine Shoot Beetle Damage Using UAV Imagery and Deep Learning Semantic Segmentation Under Different Forest Backgrounds. *Forests*, 16(4), 668.

Wang, P., Luo, W., Liu, J., Zhou, Y., Li, X., Zhao, S., Zhang, G., Zhao, Y., 2025b. Real-time semantic SLAM-based 3D reconstruction robot for greenhouse vegetables. *Computers and Electronics in Agriculture*, 237, 110582.

Wavrek, M. T., Carr, E., Jean-Philippe, S., McKinney, M. L., 2023. Drone remote sensing in urban forest management: A case study. *Urban Forestry and Urban Greening*, 86, 127978.

Wu, F., Zhu, R., Meng, F., Qiu, J., Yang, X., Li, J., Zou, X., 2024. An Enhanced Cycle Generative Adversarial Network Approach for Nighttime Pineapple Detection of Automated Harvesting Robots. *Agronomy*, 14(12), 3002.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. 34, 12077–12090.

Xu, Z., Wu, D., Yu, C., Chu, X., Sang, N., Gao, C., 2023. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. *38th AAAI Conference on Artificial Intelligence*, 6378-6386.

Yang, R., Guo, Y., Hu, Z., Gao, R., Yang, H., 2023. Semantic segmentation of cucumber leaf disease spots based on eca-segformer. *Agriculture*, 13(8), 1513.

Zhang, P., Niu, L., Cai, M., Chen, H., Sun, X., 2024. AAUC-onvNeXt: Enhancing Crop Lodging Segmentation with Optimized Deep Learning Architectures. *Plant Phenomics*, 6, 0182.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6230–6239.

Zheng, J., Yuan, S., Li, W., Fu, H., Yu, L., Huang, J., 2025. A review of individual tree crown detection and delineation from optical remote sensing images: current progress and future. *IEEE Geoscience and Remote Sensing Magazine*, 13(1), 209-236.

Zurqani, H. A., 2025. A multi-source approach combining GEDI LiDAR, satellite data, and machine learning algorithms for estimating forest aboveground biomass on Google Earth Engine platform. *Ecological Informatics*, 86, 103052.