

Article

Disentangling Spectrally Similar Urban Vegetation via Semantic Segmentation-Guided Object Analysis and Multi-Periodic Phenological Features

Chenglong Zhu ¹, Xi Cheng ^{1,*} , Tao Liu ¹, Haoyu Wang ², Hao Lei ¹, Haiyu Wang ¹ and Zhanfeng Shen ²

¹ College of Geophysics, Chengdu University of Technology, Chengdu 610059, China; 2024020476@stu.cdut.edu.cn (C.Z.); 2024020487@stu.cdut.edu.cn (T.L.); 2024020498@stu.cdut.edu.cn (H.L.); 2023050440@stu.cdut.edu.cn (H.W.)

² Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China; wanghaoyu191@mailsucas.ac.cn (H.W.); shenzf@aircas.ac.cn (Z.S.)

* Correspondence: chengxi13@cdut.edu.cn; Tel.: +86-181-807-89799

Highlights

What are the main findings?

- A framework integrating semantically constrained object analysis and TimesNet-based multi-periodic phenological modeling was developed to disentangle spectrally similar urban vegetation types (trees, grasslands, farmlands).
- Applied to Chengdu using 0.27 m VHR imagery and Sentinel-2 time series, the framework achieved 91.6% overall accuracy with F1-scores of 92.5%, 91.9%, and 87.6% across vegetation types.

What are the implications of the main findings?

- Phenological features were important for distinguishing spectrally similar classes; removing them reduced overall accuracy by 13.1 percentage points, with F1-score decreases of 16.0 and 23.0 percentage points for grassland and farmland, respectively.
- The framework demonstrates that combining semantic-constrained object boundaries with multi-temporal phenological information effectively addresses spectral confusion and boundary fragmentation challenges in fine-grained urban vegetation mapping.

Abstract

Fine-grained classification of urban green spaces (UGSs) is important for urban ecological assessment and management but remains challenging because of spectral similarity among vegetation types and inaccurate object delineation in complex urban environments. This study proposes a pixel-to-object framework that combines semantic segmentation-guided object construction with multi-periodic phenological modeling. A semantic green-space mask derived from 0.27 m very-high-resolution imagery constrains superpixel segmentation to generate spatially coherent, boundary-aware green space object-level patches (GSOPs). Pixel-level temporal representations are then derived from Sentinel-2 normalized difference vegetation index (NDVI) time series using TimesNet, aggregated into GSOP-level phenological features, and combined with spatial attributes to classify urban trees, grasslands, and farmlands. Applied to the built-up area of Chengdu, China, the framework achieved an overall accuracy of 91.6%, with F1-scores of 92.5%, 91.9%, and 87.6% for urban trees, grasslands, and farmlands, respectively. Ablation experiments showed that removing phenological features reduced overall accuracy by 13.1 percentage points and decreased the F1-scores of grasslands and farmlands by 16.0 and 23.0 percentage points, respectively.



Academic Editor: Victor Mesev

Received: 17 May 2026

Revised: 28 July 2026

Accepted: 3 August 2026

Published: 6 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

These results demonstrate that semantically constrained object delineation and phenological information jointly reduce boundary fragmentation and improve the discrimination of spectrally similar urban vegetation types.

Keywords: urban green space; semantic segmentation; object-based classification; time-series analysis; spatiotemporal fusion

1. Introduction

Definitions of urban green space (UGS) vary across disciplines and application contexts. From an operational remote-sensing land-cover perspective, this study defines UGS as vegetation-covered land within the delineated built-up study area and classifies it into urban trees, grasslands, and remaining farmlands following previous fine-grained urban landscape mapping [1,2]. These vegetation types contribute differently to urban ecosystems through air-pollution mitigation, microclimate regulation, carbon sequestration, and urban heat-island alleviation [3–6]. The growing recognition of these benefits has driven the continuous evolution of methods for quantifying urban ecosystem services [7]. With approximately 70% of the global population projected to live in urban areas by 2050, accurate monitoring of UGS distribution is increasingly critical for sustainable urban planning [8–11]. Because urban trees, managed grasslands, and remaining farmlands provide distinct ecological functions, moving beyond binary green/non-green mapping toward fine-grained classification is essential for evidence-based urban ecological management.

Remote sensing-based UGS mapping has progressed through several methodological stages, each of which has addressed, but not fully resolved, the challenge of accurate spatial delineation. Early approaches relied on vegetation indices, such as the normalized difference vegetation index (NDVI), soil-adjusted vegetation index (SAVI), and enhanced vegetation index (EVI), to separate vegetated from non-vegetated areas using spectral thresholds [11,12]. While effective for coarse-scale mapping, these pixel-level methods are inherently limited by “salt-and-pepper” noise and blurred boundaries in complex urban scenes [13]. Machine learning classifiers such as decision trees, support vector machines (SVM), and random forests improved discrimination power but still struggle to represent the complex spatial structures present in very high-resolution (VHR) imagery [11,14].

The advent of deep learning substantially advanced UGS segmentation. Fully convolutional networks (FCNs) and their successors, notably U-Net and DeepLab, enabled end-to-end pixel-level prediction, while Transformer-based architectures such as SegFormer and Swin Transformer further enhanced performance through long-range dependency modeling [15–18]. Recent studies have further explored multi-scale feature fusion [19,20], dual-branch convolutional neural network (CNN)–Transformer architectures [21], and SegFormer-based fusion modules [22] to improve boundary precision. Despite these advances, pixel-based deep learning methods still suffer from boundary blurring and fragmented predictions, particularly for small or irregularly shaped green patches embedded in dense urban fabrics.

Object-based image analysis (OBIA) offers a complementary strategy by grouping pixels into spatially coherent segments and incorporating geometric and contextual information, which is particularly advantageous for high-resolution imagery [23–25]. However, most existing OBIA pipelines rely on generic segmentation algorithms whose boundaries are not constrained by semantic information, leading to misalignment between object boundaries and true land-cover transitions. A key unresolved challenge, therefore, is

how to combine the semantic accuracy of deep learning with the structural integrity of object-based analysis to achieve boundary-faithful green space delineation.

While spatial methods focus on where green spaces are, temporal information addresses what type they are by capturing phenological dynamics unique to each vegetation class. Multi-temporal remote sensing imagery has proven valuable for differentiating vegetation types whose spectral signatures overlap in any single acquisition [11,26]. Time-series analyses have been successfully applied to land-cover monitoring [27], crop identification through seasonal spectral modeling [28–30], and fine-scale horticultural mapping [31].

Within the UGS domain, the utilization of temporal data has evolved from simple bi-temporal comparisons to continuous phenological analysis. Early studies demonstrated the superiority of bi-temporal imagery for identifying tree species and functional vegetation types [32,33]. More recent work has leveraged dense satellite image time series to capture fine-scale phenological variations in urban trees [34], and the integration of continuous phenological features into machine learning and deep learning classifiers has improved fine-grained vegetation recognition [1,11,35].

Despite this progress, two critical limitations persist. First, most time-series studies in UGS rely on medium-resolution data (e.g., Landsat at 30 m or Sentinel-2 at 10 m), which introduces severe mixed-pixel effects in heterogeneous urban landscapes and cannot be directly aligned with VHR spatial analysis. Second, many existing studies use summary statistics (e.g., mean, maximum NDVI) or conventional sequence models that do not explicitly represent frequency-dependent temporal patterns, such as the distinct cultivation and harvesting cycles of farmland relative to the generally smoother seasonal trajectories of urban-tree vegetation, that are essential for resolving inter-class spectral confusion.

These limitations highlight the need for an object-level framework that combines semantically accurate VHR delineation with phenological information from medium-resolution time-series imagery. Accordingly, this study aims to develop and quantitatively evaluate a pixel-to-object framework for distinguishing urban trees, grasslands, and farmlands in Chengdu's built-up area. The framework integrates semantically constrained object delineation from VHR imagery with phenological information from Sentinel-2 NDVI time series. Its principal novelty lies in two aspects: a semantic-mask-constrained strategy for boundary-aware GSOP construction and an object-level cross-resolution strategy that aggregates pixel-level temporal predictions into GSOP-level phenological features and combines them with spatial and geometric attributes.

2. Study Area and Data Preparation

2.1. Study Area

Chengdu is located in the heart of the Sichuan Basin in southwestern China. The urban core is primarily situated on the flat central plains and exhibits typical subtropical monsoon climate characteristics, with warm, humid conditions and synchronized rainfall and heat patterns. The average annual temperature in 2020 was 16.6 °C, and the annual precipitation was 1229.2 mm. These climatic conditions support the growth of diverse vegetation, resulting in a high level of urban greening.

This research focuses on the core urbanized sector (i.e., the built-up zone) of Chengdu rather than the entire administrative territory (Figure 1). The study-area boundary shown in Figure 1 was delineated by integrating urban morphological characteristics, land-use dynamics, and multisource remote-sensing observations, resulting in a clearly defined spatial framework for subsequent analysis [36].

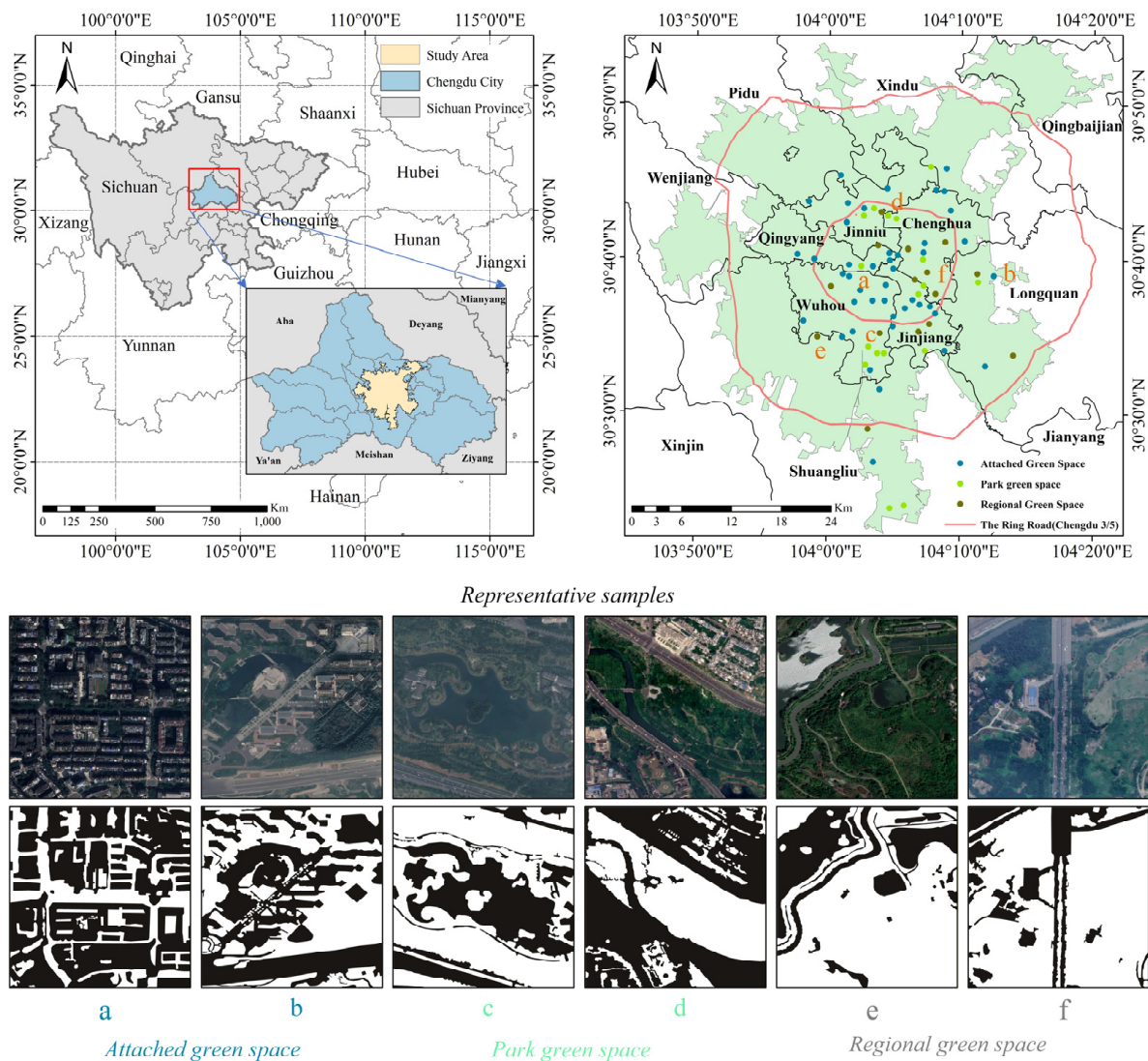


Figure 1. Location of the study area and representative samples from the semantic segmentation dataset. Samples a–f are representative 3968×3968 -pixel VHR image blocks and their corresponding binary UGS masks. Attached green space, park green space, and regional green space indicate the sample-source contexts rather than the final classification categories. In the masks, white denotes target green space, whereas black denotes the background class.

According to the Chengdu Statistical Yearbook 2021, by the end of 2020, the total green space area in the built-up zone reached 36,334 hectares, including 12,528 ha of park green space. In contrast, the green coverage area was 42,748 ha, corresponding to a green coverage rate of 43.90%. Here, green space area refers to the land area of designated green spaces, whereas green coverage area represents the vertically projected area of vegetation cover within the built-up area. The primary types of UGSs include cultivated land, garden land, and grassland, forming a diverse green landscape structure. The coexistence of dense urban vegetation, managed grassland, and remaining farmland makes Chengdu a suitable case study for fine-grained vegetation classification.

2.2. Data Preparation

2.2.1. VHR Remote Sensing Data and Training Data

The VHR remote sensing imagery used in this study was sourced from 2020 Google Earth data, with a spatial resolution of approximately 0.27 m. The three-band RGB imagery

was color-corrected, mosaicked, clipped to the study area, and reprojected to WGS 84 (EPSG:4326) for green-space extraction and spatial-feature analysis.

A total of 86 representative sample regions were strategically selected across the study area, covering diverse green space (e.g., city parks, street greenery, residential vegetation) and non-green space settings (e.g., buildings, water bodies, roads). As shown in Figure 1, these samples span the central urban districts, suburban areas, and urban-rural fringe zones, ensuring adequate representation of the spatial heterogeneity of green spaces. For each sample region, 3968×3968 pixel image blocks were extracted from the preprocessed imagery and manually labeled to delineate green space boundaries. To meet the computational requirements of deep learning models, these large blocks were systematically cropped into smaller 512×512 pixel tiles with 25% overlap, yielding a total of 7700 image tiles with corresponding binary green space masks. This dataset forms the Chengdu UGS semantic segmentation dataset and was partitioned into training, validation, and test sets at a ratio of 70%, 15%, and 15%, respectively, with tiles from the same sample region assigned to the same partition to avoid spatial leakage within the binary green-space extraction task.

2.2.2. Time-Series NDVI Data

To extract temporal phenological features of UGS, this study utilized Sentinel-2 satellite imagery (spatial resolution: 10 m) from the Google Earth Engine (GEE) platform, covering the period from January to December 2020. Due to the large spatial extent of Chengdu's built-up area and frequent cloud cover, it was challenging to obtain cloud-free imagery for the entire region at any single acquisition date.

To address this issue, a standard map-sheet-based spatial partitioning strategy was adopted. The built-up study area intersected nine standard map-sheet extents, and the portion within each map sheet was treated as a separate sub-region. This partition followed fixed cartographic extents rather than administrative boundaries and was used exclusively for low-cloud Sentinel-2 image retrieval, monthly NDVI construction, and sub-region-specific temporal modeling. It did not alter the boundary or spatial extent of the overall study area or the definition of the final GSOP classification units. For each sub-region, one relatively low-cloud Sentinel-2 image was selected for each month, yielding a 12-date annual NDVI sequence. Because the images were selected separately for each sub-region, the exact acquisition dates varied slightly across sub-regions according to local cloud conditions; however, all selected images were restricted to the 2020 calendar year, and each selected image represented the corresponding monthly phenological condition as closely as possible. This strategy maximized temporal data quality while preserving annual comparability for phenological analysis. The resulting 12-month NDVI sequence was used to characterize broad annual and seasonal vegetation dynamics.

3. Methods

3.1. Outline of the Proposed Method

This study proposes a pixel-to-object framework for fine-grained UGS classification that integrates deep semantic segmentation, superpixel-based object extraction, and phenological temporal features. As illustrated in Figure 2, the workflow consists of four stages: (1) DE-SegFormer generates binary green-space masks from VHR imagery; (2) simple Linear Iterative Clustering (SLIC) [37] superpixel segmentation refines pixel-level masks into spatially continuous GSOPs; (3) multi-dimensional features are constructed for each GSOP, including intrinsic spatial features (ISFs) from VHR imagery and phenological temporal features (PTFs) from Sentinel-2 NDVI time-series using TimesNet [38]; (4) a Random

Forest classifier categorizes GSOPs into urban trees, grasslands, and farmlands based on the integrated spatiotemporal features.

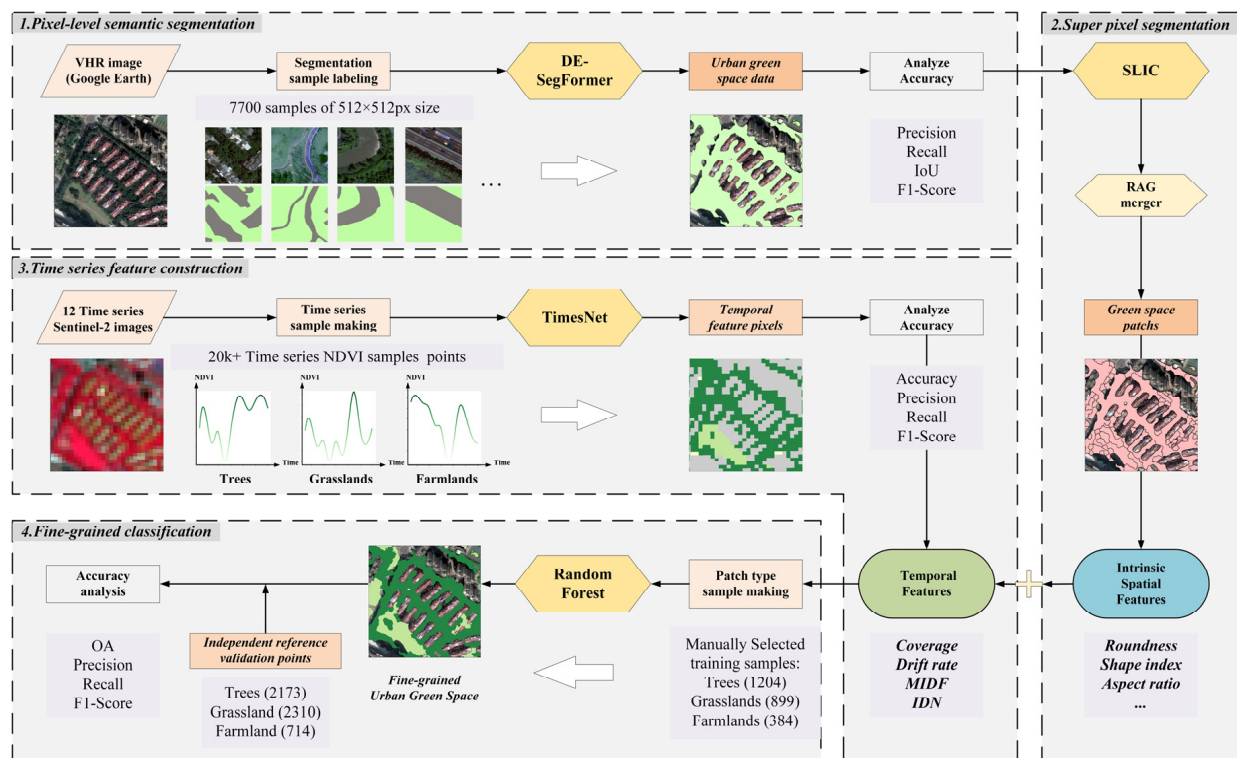


Figure 2. Workflow of the UGS fine-grained classification method.

3.2. UGS Semantic Segmentation Using DE-SegFormer

3.2.1. DE-SegFormer

UGSs in VHR imagery exhibit complex spatial patterns with blurred boundaries and diverse textures. The SegFormer model, based on the Transformer architecture, overcomes the limitations of traditional CNNs through its hierarchical encoder and lightweight MLP decoder, enabling effective multi-scale feature fusion and long-range dependency modeling.

The original SegFormer decoder projects encoder features from different stages to a common channel dimension, independently upsamples them to the same spatial resolution, and concatenates them for prediction. Although this lightweight design is computationally efficient, interactions among adjacent feature scales occur only after upsampling and concatenation. Consequently, fine spatial details from shallow layers may not be sufficiently integrated with the high-level semantic information of deeper layers. Moreover, the use of fixed bilinear interpolation limits the decoder's ability to learn task-specific boundary reconstruction, which is particularly important for small, fragmented, and irregularly shaped urban green spaces. To address these limitations, DE-SegFormer introduces two modifications while retaining the original SegFormer encoder (shown in Figure 3). First, a Merging Block (MB) progressively propagates shallow spatial details to deeper feature stages through recursive downsampling and addition before final multi-scale concatenation. This design enables each deeper feature map to incorporate both semantic context and boundary-related spatial information. Second, the final fixed bilinear upsampling operation is replaced by a learnable transposed convolution to enhance detail reconstruction. These modifications are designed specifically to improve the delineation of narrow green belts, fragmented vegetation patches, and irregular vegetation boundaries in VHR imagery.

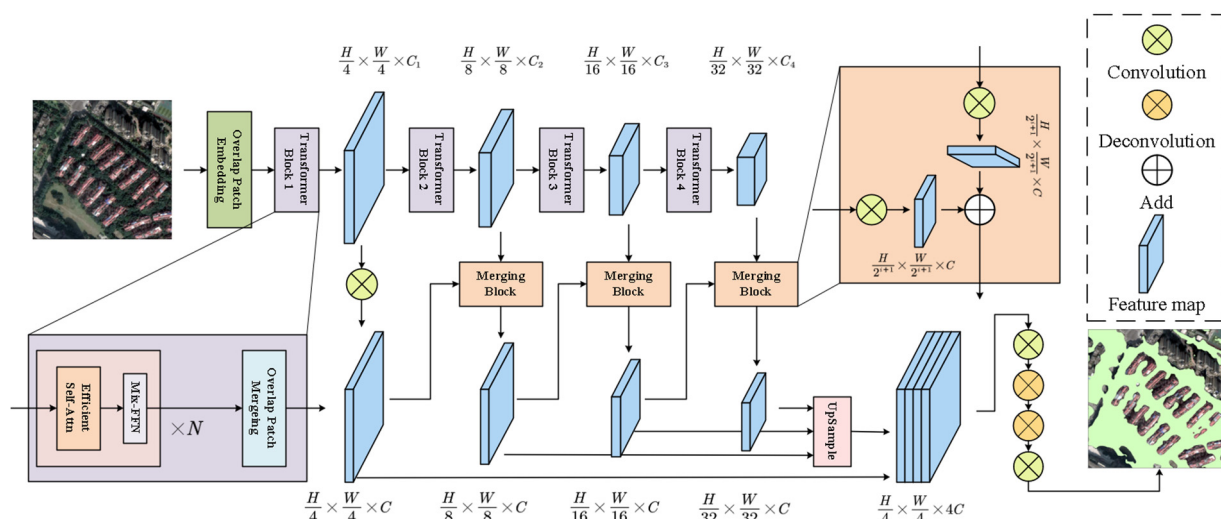


Figure 3. Network architecture of DE-SegFormer.

The MB performs hierarchical feature fusion through a recursive process. Let N_i denote the feature map from encoder stage i ($i \in \{1, 2, 3, 4\}$), N'_i denote the fused feature map after MB processing, and C denote the unified channel dimension (set to $C = 128$). The fusion process is formalized as follows:

(1) Initialization. For the shallowest feature map N_1 , only channel normalization is applied:

$$N'_1 = \text{Conv}_{1 \times 1}(N_1) \quad (1)$$

(2) Recursive fusion (for $i = 1, 2, 3$). The new feature map N'_{i+1} is obtained by downsampling the previous fused feature map N'_i and adding it to the channel-normalized current feature map N_{i+1} :

$$N'_{i+1} = \text{Down}_{3 \times 3}(N'_i) + \text{Conv}_{1 \times 1}(N_{i+1}) \quad (2)$$

where $\text{Down}_{3 \times 3}(\cdot)$ denotes a 3×3 convolution with a stride 2 for downsampling.

(3) Multi-scale concatenation. All fused feature maps N'_i are upsampled to match the resolution of N'_1 and concatenated along the channel dimension:

$$F_{\text{merged}} = \text{Concat}(N'_1, \text{Up}(N'_2), \text{Up}(N'_3), \text{Up}(N'_4)) \quad (3)$$

This operation produces a feature map of size $\frac{H}{2} \times \frac{W}{2} \times 4C$. To further enhance spatial detail representation, DE-SegFormer replaces the bilinear interpolation used in SegFormer's decoder with a learnable transposed convolution during the final upsampling stage, reducing boundary aliasing effects. This learnable upsampling operation is intended to improve spatial-detail reconstruction in the binary masks used for subsequent patch extraction.

3.2.2. Network Training Details

The experiments were conducted using PyTorch version 2.0.1 as the underlying deep learning framework, with training performed on an NVIDIA GeForce RTX 3060 GPU with 12 GB of memory. The region-level 70:15:15 training-validation-test partition described in Section 2.2.1 was retained for all segmentation experiments. To enhance the model's generalization capability and prevent overfitting, data augmentation techniques including random flipping (horizontal and vertical) and random rotation (90° , 180° , 270°) were applied during training. Due to GPU memory constraints, the batch size was set to 4. The

model was trained using the AdamW optimizer [39] with an initial learning rate of 5×10^{-5} and a weight decay of 0.01. The encoder was initialized with ImageNet pre-trained weights. To address the class imbalance between the dominant background and smaller green space objects, the Focal Loss function was employed with parameters $\alpha = 0.25$ and $\gamma = 2.0$ [40]. An adaptive learning rate scheduling strategy was adopted: the learning rate was reduced by a factor of 0.5 if the validation loss did not decrease for 3 consecutive epochs. The training process was set for a maximum of 100 epochs, with an early stopping mechanism triggered if the learning rate dropped below 1×10^{-6} or if the validation loss showed no improvement for 10 epochs. The learning rate, scheduling, and early-stopping settings were determined empirically based on the convergence of the validation loss. The model with the lowest validation loss was selected as the final model.

To quantify the additional computational cost introduced by the enhanced decoder, SegFormer and DE-SegFormer were compared under the same experimental environment and input setting, as described in Section 4.1.

3.3. Superpixel Segmentation and GSOPs Extraction

To convert the pixel-level UGS mask into object-level analysis units, SLIC was applied to the VHR imagery only within areas identified as green space by DE-SegFormer. The semantic mask constrained the analysis extent, while the color and spatial information in the VHR imagery guided the subdivision of masked green-space regions into compact superpixels that followed local image transitions. These superpixels provided the initial units for subsequent RAG-based merging and GSOP construction.

The SLIC algorithm initially generates cluster centers based on the desired number of superpixels K , and calculates the initial distance $S = \sqrt{N/K}$ between neighboring centers, where N is the number of pixels in the image [37]. After adjusting seed points to positions with the smallest gradient within the local neighborhood, the RGB pixels are converted into the CIELAB color space. A five-dimensional vector $C^i = \{l^i, a^i, b^i, x^i, y^i\}$ combining spatial coordinates and color information is formed. The weighted distance D between the pixel and the seed point is then calculated, as shown in the following equations:

$$D = \sqrt{d_{\text{lab}}^2 + \left(\frac{d_s}{S}\right)^2 m^2} \quad (4)$$

$$d_{\text{lab}} = \sqrt{(l_i - l_j)^2 + (a_i - a_j)^2 + (b_i - b_j)^2} \quad (5)$$

$$d_s = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (6)$$

where m controls the balance between color similarity and spatial proximity. Based on this distance measure D , the K-means algorithm is used for pixel clustering and iterative updating. In this study, K was adaptively calculated as $K = \text{round}(N/A_{\text{sp}})$, where A_{sp} denotes the target number of pixels per initial superpixel. For the 0.27 m VHR imagery, A_{sp} was set to 400 pixels. This setting reduced under-segmentation of small UGS patches. The compactness parameter m was set to 10 to balance color similarity and spatial proximity, while σ was empirically set to 4 based on boundary adherence and fragmentation in representative scenes.

To further reduce over-segmentation and finalize the structural representation of the green space, a hierarchical clustering method is applied. A Region Adjacency Graph (RAG) is constructed based on the initial superpixels, and neighboring superpixels are iteratively merged based on their mean color difference in CIELAB space [41,42]. The merging threshold *thresh* was set to 10. The resulting objects are spatially continuous GSOPs. Because urban trees, grasslands, and farmlands vary substantially in spatial extent and shape, the

final GSOPs do not have a fixed target size; their scales are jointly determined by the initial superpixels, local vegetation structure, and the RAG-based merging process. These GSOPs serve as the definitive analysis units for the multi-dimensional feature extraction and fine-grained classification.

3.4. Feature Space Construction of GSOPs

3.4.1. ISFs Construction of GSOPs

The inherent features of GSOPs refer to their fundamental spatial and spectral attributes, primarily including morphological features and color features. These features can be quantitatively described from multiple dimensions such as geometric shape, structural form, and spectral expression. In terms of morphology, common metrics include area, perimeter, circularity, shape index, and aspect ratio (length to width ratio), which characterize the geometric structure, compactness, and spatial distribution of GSOPs. These geometric attributes may provide complementary information because managed grassland, farmland, and fragmented tree patches can differ in shape and boundary complexity. In terms of color features, this study incorporates the average RGB values within each GSOP to provide auxiliary spectral information. Table 1 summarizes the ISFs used in this study and their computation methods.

Table 1. ISFs of GSOPs.

ISFs	Description	Formula
Area	The area of a GSOP	—
Perimeter	The perimeter of a GSOP	—
Roundness	Circularity of a GSOP	$\frac{4\pi \times \text{area}}{\text{perimeter}^2}$
Shape Index	Boundary complexity	$\frac{\text{perimeter}}{4 \times \sqrt{\text{area}}}$
Width	Width of minimum bounding rectangle	—
Length	Length of minimum bounding rectangle	—
Aspect Ratio	Length-to-width ratio of minimum bounding rectangle	$\frac{\text{length}}{\text{width}}$
MeanR	Mean red value within a GSOP	$\frac{1}{N} \sum_{i=1}^N R_i$
MeanG	Mean green value within a GSOP	$\frac{1}{N} \sum_{i=1}^N G_i$
MeanB	Mean blue value within a GSOP	$\frac{1}{N} \sum_{i=1}^N B_i$

Note: — indicates that the corresponding feature is directly derived from the GSOP geometry and therefore does not require an additional explicit formula.

3.4.2. PTFs Construction of GSOPs

To further explore the temporal variation characteristics of GSOPs, this study used 12-month Sentinel-2 NDVI time series from 2020 to construct a temporal trajectory for each 10 m pixel. The TimesNet model was applied to analyze the temporal NDVI data. Unlike traditional approaches that model one-dimensional (1D) time series directly, TimesNet transforms a 1D sequence into two-dimensional (2D) representations organized according to dominant frequencies, enabling convolution-based temporal feature extraction. As shown in Figure 4, TimesNet consists of stacked TimesBlocks that adaptively capture multi-periodic information.

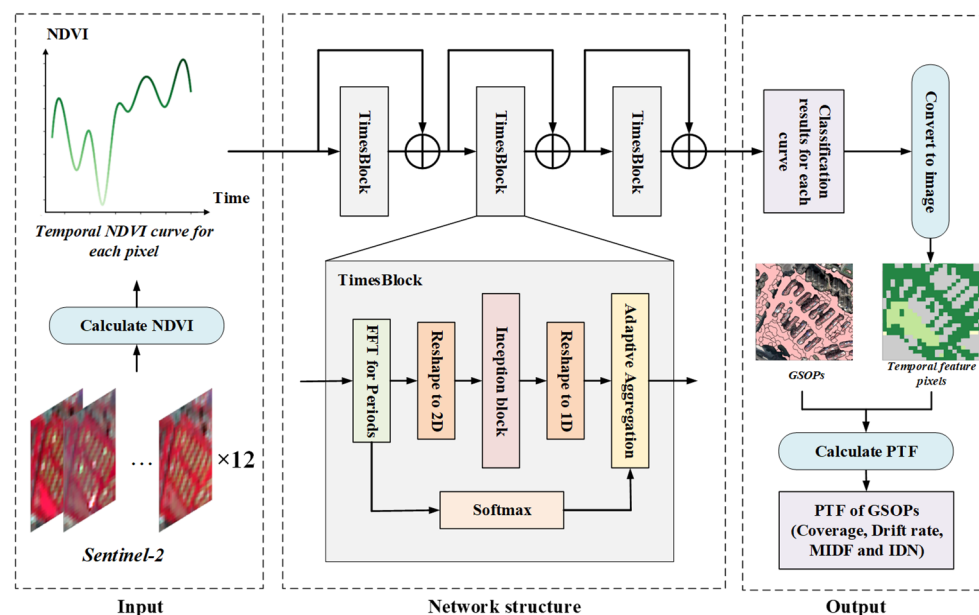


Figure 4. Architecture and input/output of the TimesNet network.

Specifically, for the l -th TimesBlock, periodic features are first extracted through Fast Fourier Transform (FFT), selecting the top k components with the largest amplitudes. A 2D time series X_{2D}^l is generated, and an Inception module extracts 2D features $\widehat{X}_{2D}^l$. These are then converted back to 1D sequence $\widehat{X}_{1D}^l$, and the final output X_{1D}^l is obtained via weighted fusion. The formula is as follows:

$$X_{2D}^{l,i} = \text{Reshape}_{p_i, f_i} \left(\text{Padding} \left(X_{1D}^{l-1} \right) \right), i \in \{1, \dots, k\} \quad (7)$$

$$\widehat{X}_{2D}^{l,i} = \text{Inception} \left(X_{2D}^{l,i} \right), i \in \{1, \dots, k\} \quad (8)$$

$$\widehat{X}_{1D}^{l,i} = \text{Trunc} \left(\text{Reshape}_{1, p_i \times f_i} \left(\widehat{X}_{2D}^{l,i} \right) \right), i \in \{1, \dots, k\} \quad (9)$$

$$X_{1D}^l = \sum_{i=1}^k \widehat{A}_{f_i}^{l-1} \times \widehat{X}_{1D}^{l,i} \quad (10)$$

where Padding indicates that zeros are added at the end of the 1D time series folding process to ensure divisibility, Reshape refers to folding the time series according to the corresponding period or frequency, and Inception denotes the 2D convolution feature extraction. $\widehat{A}_{f_i}^{l-1}$ represents the strength of the time series corresponding to the frequency after being normalized by the Softmax activation function.

Since NDVI data for green spaces exhibit clear seasonal variations, the multi-periodic modeling mechanism of TimesNet is particularly effective in capturing the underlying temporal patterns of the vegetation growth cycle. Although TimesNet is typically applied to longer time series, the twelve-month sequence covers one complete annual vegetation cycle and provides broad seasonal information for distinguishing vegetation classes. TimesNet was used as a frequency-aware temporal representation model to characterize differences in the shape and seasonal variation in these monthly trajectories. Nevertheless, the monthly sampling interval cannot resolve rapid phenological events at daily or weekly scales, and the interpretation of fine-scale or repeated sub-annual periodicities is therefore limited.

To address the challenge of cloud contamination and ensure high-quality time-series data, a spatial partitioning strategy was adopted. The study area was divided into 9 sub-regions. For each sub-region, a dedicated TimesNet model was trained independently

using the 12-month NDVI sequences specific to that area (as described in Section 2.2.2). Labeled samples were manually collected for each region, totaling approximately 1000–1700 pixel-level samples per class (urban trees, grasslands, farmlands), which were randomly split into training, validation, and test sets at a ratio of 70%, 15%, and 15%, respectively. The twelve-month NDVI trajectories of these training samples are shown in Figure 5. The trajectories exhibit class-dependent differences in seasonal amplitude and temporal variation, although substantial overlap remains among the three vegetation classes in some months. These monthly observations are therefore used to characterize broad annual and intra-annual seasonal patterns rather than short-term phenological events.

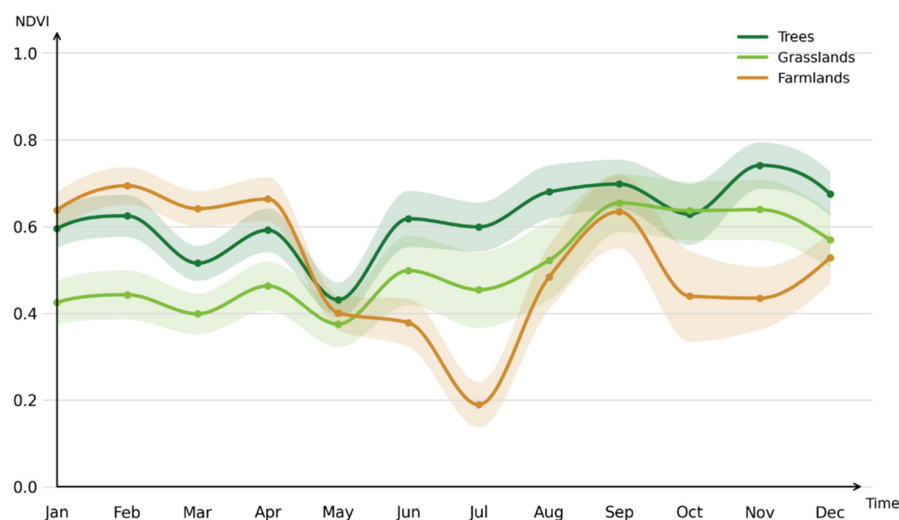


Figure 5. Monthly NDVI trajectories in 2020 for TimesNet training samples of urban trees, grasslands, and farmlands. Curves and shading indicate the mean and ± 1 standard deviation across the training samples, respectively.

The model was implemented in PyTorch and trained using the RAdam optimizer with a learning rate of 0.0001, batch size of 32, and cross-entropy loss. Training was conducted for 100 epochs with early stopping if validation loss did not improve for 10 consecutive epochs. The model was trained to classify the temporal NDVI curve of each pixel, outputting a probability map representing the likelihood of each vegetation type.

Since the temporal feature pixels (10 m resolution) differ in scale from the GSOPs (0.27 m resolution), a spatial overlay analysis was performed to derive patch-level metrics. Pixel-level temporal predictions were summarized at the GSOP level using four metrics: Coverage, Drift rate, Maximum Independent Distribution Factor (MIDF), and Independent Distribution Number (IDN).

Coverage refers to the proportion of feature pixels within a region, representing the dominance of a specific phenological type within the patch. Drift rate quantifies the spatial displacement of class-specific connected components from the GSOP centroid using area-weighted, radius-normalized centroid distances. MIDF represents the ratio of the area covered by the largest IFD, while the IDN refers to the number of IFDs. These metrics summarize the within-GSOP distribution of temporal predictions and reduce the loss of information associated with assigning only a single dominant temporal class to each object. The combination of the MIDF and IDN effectively describes the dispersion of IFDs, providing a more detailed object-level representation of temporal-prediction heterogeneity.

For each GSOP, Coverage and Drift rate are calculated for each vegetation class (urban trees, grasslands, farmlands), while the MIDF and IDN are calculated once, yielding 8 PTFs per GSOP. Let $area_f$ denote the area of feature pixels of a specific class, $area_o$ denote the GSOP area, n denote the number of IFDs, w_i denote the area weight of the i -th IFD, c_o and

c_{IFD_i} denote centroids of the GSOP and i -th IFD, $DIS(\cdot, \cdot)$ denote Euclidean distance, and r denote the GSOP radius. Table 2 summarizes the PTFs and their calculation methods.

Table 2. PTFs of GSOPs.

PTFs	Description	Formula
Coverage	the proportion of feature pixels within a GSOP	$\frac{area_f}{area_o}$
Drift rate	the deviation of the IFD within a GSOP	$\sum_{i=1}^n w_i \times \frac{DIS(c_o, c_{IFD_i})}{r}$
MIDF	the ratio of the area covered by the largest IFD	$\frac{MAX(area_{IFD})}{SUM(area_{IFD})}$
IDN	the number of IFDs	NUMBER(IFD)

3.5. Fine-Grained Classification of UGS Based on GSOPs

The random forest (RF) classifier was selected due to its robustness against noise and ability to handle high-dimensional multi-modal data. RF is an ensemble method that constructs multiple decision trees using random feature selection and bootstrap aggregation, with final classification determined by majority voting.

For each GSOP, an 18-dimensional feature vector was constructed by concatenating 10 ISFs (Table 1) and 8 PTFs: Coverage for each of the 3 classes (3 features), Drift rate for each of the 3 classes (3 features), the MIDF (1 feature), and the IDN (1 feature) from Table 2. The RF hyperparameters were selected through grid search. The final model was configured with 400 decision trees, a maximum depth of 25, and the Gini impurity index as the splitting criterion, with balanced class weights to address class imbalances.

A dataset of 2487 labeled GSOPs was constructed for training, including 1204 urban trees, 899 grasslands, and 384 farmlands. The trained RF classifier was then applied to each GSOP to generate the final fine-grained classification map.

3.6. Accuracy Evaluation Metrics and Methods

To comprehensively evaluate the accuracy of UGS extraction and classification, five metrics are employed: Overall Accuracy (OA), Precision, Recall, F1-score, and Intersection over Union (IoU) [43,44]. The validation datasets differ according to the specific task.

For the final fine-grained classification (Section 3.5), 5197 reference points were independently interpreted from temporally matched VHR and historical imagery. These points were sampled exclusively from areas that did not overlap with the GSOPs used for RF training and were not used for feature construction, model selection, or hyperparameter optimization. A stratified sampling strategy was applied across diverse urban scenes, yielding 2173 points for urban trees, 2310 for grasslands, and 714 for farmlands. The smaller farmland sample proportionally reflects its limited spatial extent within the built-up area, yet remains sufficient for robust accuracy assessment.

The evaluation metrics are defined as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (12)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (13)$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

$$IoU = \frac{Area_O}{Area_U} \quad (15)$$

where TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively. $Area_O$ is the intersection area and $Area_U$ is the union area between predicted and reference segmentation masks. For semantic segmentation, IoU is calculated for the green space class. For multi-class classification, metrics were calculated for each class independently (One-vs-All strategy) and then averaged.

4. Results

4.1. Results of Semantic Segmentation of UGS

Pixel-level semantic segmentation was performed on VHR images using DE-SegFormer. Comparison experiments were conducted with SegFormer, Swin Transformer, FCN, and DeepLabV3+ using the same dataset split, input size, augmentation strategy, and general training protocol. Performance was evaluated using Precision, Recall, IoU, and F1-score.

As presented in Table 3, DE-SegFormer achieved the highest performance across all metrics, with Precision, Recall, IoU, and F1-score values of 0.876, 0.866, 0.786, and 0.867, respectively. Compared to SegFormer (F1-score: 0.819), DE-SegFormer improved the F1-score and IoU by 4.8 and 7.2 percentage points, respectively. It also outperformed Swin Transformer (F1-score: 0.846) and CNN-based models (DeepLabV3+ and FCN).

Table 3. Accuracy results of different models in UGS semantic segmentation.

Model	Precision	Recall	IoU	F1-Score
FCN	0.828	0.811	0.702	0.816
DeepLabV3+	0.842	0.832	0.733	0.833
SegFormer	0.847	0.803	0.714	0.819
Swin Transformer	0.855	0.844	0.753	0.846
DE-SegFormer	0.876	0.866	0.786	0.867

Figure 6 compares segmentation results in challenging urban areas. As highlighted by red boxes, DE-SegFormer shows improved performance in two key aspects. First, in rows 1–2 showing dense residential areas, baseline models (FCN, DeepLabV3+) frequently miss small, fragmented green patches between buildings, while SegFormer tends to over-smooth these details. DE-SegFormer more accurately captures these fine-grained isolated patches. Second, in rows 3–4 showing roadside vegetation, DE-SegFormer produces sharper and more coherent boundaries compared to baseline models, which tend to merge vegetation with adjacent shadows or dark pavement.

As shown in Table 4, compared with SegFormer, DE-SegFormer increases the parameter count from 7.71 M to 32.60 M, the inference time from 33.4 to 91.2 ms per 512×512 tile, and the training time per epoch from 15 to 69 min. This additional computational cost, mainly introduced by recursive multi-scale fusion and learnable upsampling, yields improvements of 4.8 and 7.2 percentage points in F1-score and IoU, respectively. Thus, DE-SegFormer trades model compactness and computational speed for higher segmentation accuracy, which is acceptable for the offline mapping application considered here.

Table 4. Computational comparison between SegFormer and DE-SegFormer.

Model	Parameters (M)	Inference Time Per 512×512 Tile (ms)	Training Time Per Epoch (min)	F1-Score	IoU
SegFormer	7.71	33.4	15	0.819	0.714
DE-SegFormer	32.6	91.2	69	0.867	0.786
Relative change	+322.8%	+173.1%	+360.0%	+4.8 pp	+7.2 pp

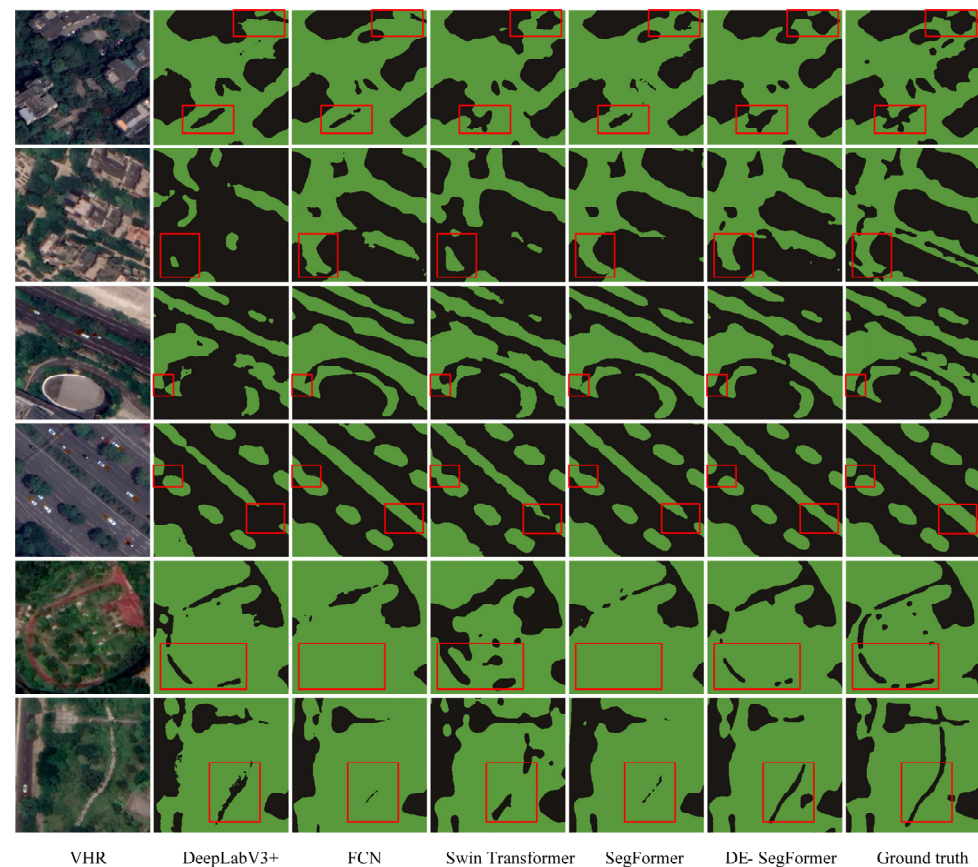


Figure 6. Results of UGS semantic segmentation with different models, where green pixels represent UGS.

4.2. Comparison of Superpixel Segmentation Methods

SLIC superpixel segmentation was compared with two commonly used methods, Felzenszwalb segmentation [45] and Watershed [46], to determine the optimal approach for refining UGS boundaries within the semantic mask. For fair comparison, the key parameters of all three methods were tuned, and the results shown in Figure 7 correspond to their selected parameter settings. For SLIC, the number of superpixels was adaptively determined according to image resolution and size, with σ set to 4 and thresh set to 10.

Figure 7 presents the segmentation results of the three methods across different UGS types, including residential areas, urban parks, and surrounding farmlands. Felzenszwalb produced relatively fragmented patches in vegetation areas, especially in tree-covered scenes with strong internal texture variation. Watershed produced irregular, jagged boundaries, with pronounced sawtooth artifacts in several regions. SLIC produced more compact and regular superpixels, with boundaries more closely following transitions between vegetation and adjacent impervious surfaces.

Overall, SLIC produced more compact and regular superpixels with boundaries closely following transitions between vegetation and adjacent impervious surfaces. Based on the GSOPs dataset constructed in Section 3.5, the average metrics of several key ISFs for different types of GSOPs are shown in Table 5. Urban trees exhibit distinct differences in Shape Index, while Aspect Ratio is particularly effective in differentiating farmlands. There is an increasing trend in Roundness values across urban trees, grasslands, and farmlands. The class means also indicate differences in RGB characteristics, although these descriptive statistics alone do not establish complete class separability.

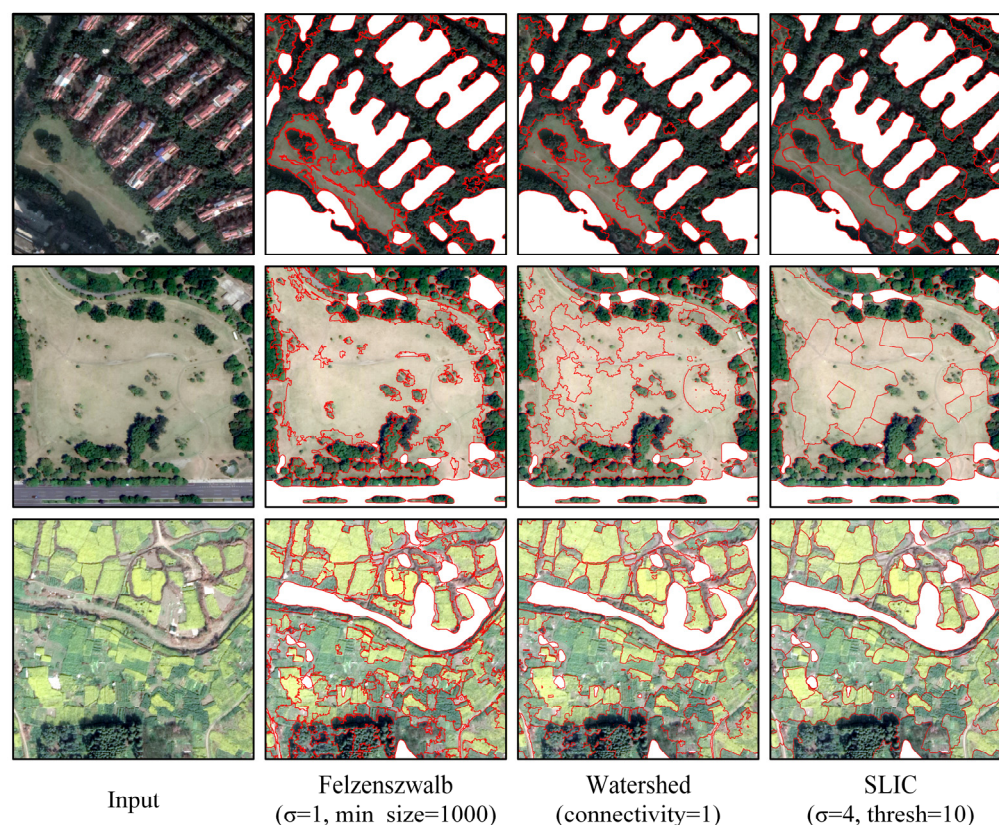


Figure 7. Segmentation results of three superpixel methods (SLIC, Felzenszwalb, and Watershed) in different UGS regions, including urban neighborhoods, urban parks, and surrounding farmlands. The red border indicates the boundary of GSOPs, and white regions indicate non-green-space areas.

Table 5. Average values of key ISFs for different types of GSOPs.

Types of GSOPs	Roundness	Shape Index	Aspect Ratio	MeanR	MeanG	MeanB
Trees	0.391	1.700	1.49	26.7	40.6	43.3
Grasslands	0.433	1.507	1.50	72.5	79.4	71.8
Farmlands	0.454	1.454	1.42	91.8	114.1	74.8

4.3. Results of Temporal Feature Construction

Three temporal modeling methods were compared for temporal feature extraction: TimesNet, PatchTST [47], and FEDformer [48]. All models used NDVI time-series data from Sentinel-2 as input and were trained using the same NDVI samples and data partitions under a consistent experimental protocol. Performance was evaluated using OA, Precision, Recall, and F1-score.

Independent TimesNet models were trained for each of the 9 partitioned sub-regions to accommodate temporal inconsistency caused by cloud cover. Table 6 presents the average performance metrics across these 9 regional models. TimesNet achieved the highest performance (OA: 0.951, F1-score: 0.941), outperforming PatchTST (OA: 0.923, F1-score: 0.911) and FEDformer (OA: 0.937, F1-score: 0.916).

Table 6. Accuracy comparison of different models in temporal feature extraction.

Method	OA	Precision	Recall	F1-Score
TimesNet	0.951	0.941	0.941	0.941
PatchTST	0.923	0.912	0.911	0.911
FEDformer	0.937	0.917	0.917	0.916

Temporal feature pixels within the region of interest for GSOPs were used to calculate metrics such as Coverage, Drift rate, MIDF, and IDN, converting pixel-level temporal features into patch-level temporal features (PTFs). As visualized in Figure 8, the ‘Temporal feature pixels’ map displays the raw classification results from TimesNet at 10 m resolution. It captures the general distribution but exhibits a distinct mosaic effect and mixed-pixel noise along boundaries. By converting these discrete predictions into patch-level metrics (Coverage, Drift rate, MIDF, and IDN), the pixel-level temporal predictions are summarized within the VHR-derived boundaries of each GSOP.

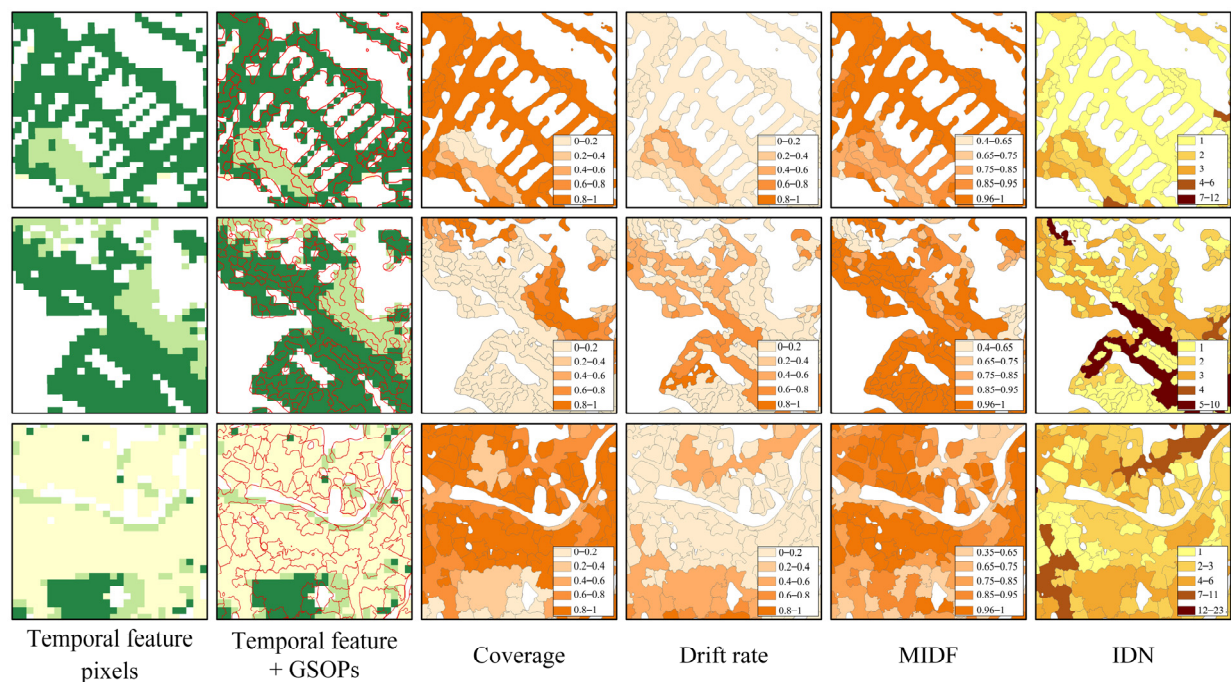


Figure 8. Temporal feature pixels (predicted by the TimesNet model) and their distribution in GSOP across different urban scenes, with PTFs (Coverage, Drift rate, MIDF, and IDN). In the temporal feature pixels, dark green represents urban trees, light green represents grasslands, and yellow represents farmlands; Coverage and Drift rate from top to bottom represent the Coverage and Drift rate for urban trees, grasslands, and farmlands, respectively.

4.4. Results of GSOPs Classification

RF, SVM, multinomial logistic regression (MLR), and k-nearest neighbors (KNNs) were compared for GSOP classification using the integrated feature space. Performance was evaluated using OA and class-level metrics (Precision, Recall, F1-score) for urban trees, grasslands, and farmlands. As shown in Table 7, RF achieved an OA of 0.916, a mean precision of 0.926, a mean recall of 0.892, and a mean F1-score of 0.907, outperforming the SVM (OA: 0.904), MLR (OA: 0.904), and KNN (OA: 0.914). As shown in Figure 9, RF demonstrated consistent performance across all three UGS types, particularly in identifying farmland (Precision: 0.956, F1-score: 0.876). A 5000-resample bootstrap analysis further yielded 95% confidence intervals of 0.909–0.923 for OA and 0.898–0.915 for Mean F1-score. For farmland, the 95% confidence intervals were 0.939–0.970 for Precision, 0.780–0.839 for Recall, and 0.858–0.895 for F1-score, quantifying the uncertainty associated with the smaller farmland validation sample.

Table 7. Accuracy comparison of different classifiers in GSOPs classification.

Classifier	OA	Mean Precision	Mean Recall	Mean F1-Score
RF	0.916	0.926	0.892	0.907
SVM	0.904	0.897	0.882	0.889
MLR	0.904	0.899	0.883	0.890
KNN	0.914	0.918	0.889	0.902

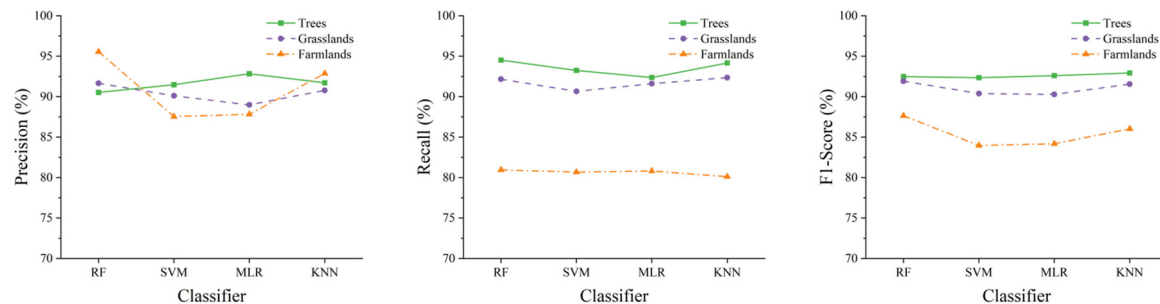
**Figure 9.** Comparison of classification performance for different UGS types (urban trees, grasslands and farmlands) across multiple classifiers. From left to right, the panels show Precision, Recall, and F1-score for urban trees, grasslands, and farmlands across the four classifiers.

Figure 10 illustrates the RF-based classification results in residential areas, urban parks, and peri-urban farmland. In the displayed examples, urban-tree patches remain spatially coherent, while grassland and farmland are mapped as separate object classes despite their similar appearance in the single-date VHR imagery. As shown in Figure 10a,b, the method accurately delineated the boundaries of urban tree patches within dense residential areas. The classification results avoided the salt-and-pepper noise typical of pixel-based methods, preserving the sharp rectangular edges of man-made green belts. Figure 10e,f show the method's capability to distinguish urban grasslands from surrounding farmlands. Despite their spectral similarity in single-date VHR imagery, the classifier correctly identified the farmland patches (yellow) and separated them from the managed grassland patches (light green).

4.5. Feature Contribution Analysis

To quantify the contribution of spatial and temporal features, ablation experiments were conducted by systematically removing each feature type from the integrated feature space. Four configurations were evaluated on the same 5197 validation points: the full framework (ISFs + PTFs), spatial features only (w/o PTFs), temporal features only (w/o ISFs), and direct GSOP-level majority voting of TimesNet predictions. For majority voting, each GSOP was assigned the dominant TimesNet-predicted class within it.

As shown in Table 8, direct GSOP-level majority voting achieved an OA of 0.678 and a Mean F1-score of 0.682, substantially lower than the PTF-only configuration (OA: 0.884; Mean F1-score: 0.878), indicating that GSOP-level PTF transformation retains more useful temporal distribution information than dominant-label assignment. Removing PTFs caused the largest performance decline, with OA and Mean F1-score decreasing by 13.1 and 15.7 percentage points, respectively, particularly for grasslands and farmlands, whose F1-scores decreased by 16.0 and 23.0 percentage points, respectively. In contrast, removing ISFs resulted in smaller losses, with OA and Mean F1-score decreasing by 3.2 and 2.9 percentage points, respectively; urban trees were most affected, with their F1-score decreasing by 3.2 percentage points. These results indicate that phenological information made the larger contribution within the evaluated feature configuration, while spatial features provided complementary information.



Figure 10. Fine-grained UGS classification results in different urban scenarios. In the six detailed comparisons labeled a–f, the red outlines indicate the boundaries of GSOPs.

Table 8. Ablation study results for feature contributions.

Configuration	OA	Mean F1-Score	Urban Trees F1-Score	Grasslands F1-Score	Farmlands F1-Score
Full feature set (ISFs + PTFs)	0.916	0.907	0.925	0.919	0.876
w/o PTFs (ISFs only)	0.785	0.750	0.846	0.759	0.646
w/o ISFs (PTFs only)	0.884	0.878	0.893	0.883	0.856
GSOP-level TimesNet majority voting	0.678	0.682	0.729	0.576	0.741

5. Discussion

5.1. Analysis of Feature Contributions and Comparison

The experimental results reveal the complementary roles of spatial and temporal features within our framework. As shown in Table 5, the intrinsic spatial features (ISFs) provide moderate inter-class differentiation: urban trees exhibit the lowest Roundness

(0.391) and highest Shape Index (1.7), reflecting irregular canopy contours along streets and between buildings. However, grasslands and farmlands show nearly identical Roundness (0.433 vs. 0.454) and Aspect Ratios (1.50 vs. 1.42), demonstrating that spatial geometry alone is insufficient to reliably separate these spectrally similar classes. Color features offer additional cues; farmlands exhibit notably higher MeanG values (114.1 vs. 79.4), but this contrast is season-dependent and may not generalize across acquisition dates.

Temporal features partly compensate for this limitation by describing seasonal dynamics that are unavailable from a single-date VHR image. TimesNet achieved an OA of 95.1% in pixel-level temporal classification (Table 6), outperforming PatchTST (92.3%) and FEDformer (93.7%). Its higher performance may be related to TimesNet's 1D-to-2D transformation mechanism, which may help characterize broad annual and intra-annual variation in the 12-month NDVI sequences.

The integrated feature space enables the RF classifier to achieve the highest OA (91.6%) and Mean F1-score (0.907) among all evaluated classifiers (Table 7). The farmland F1-score of 87.6% suggests that spatiotemporal fusion can reduce confusion that remains difficult to address using single-date imagery. Direct numerical comparisons with previous studies should be interpreted cautiously because of differences in study areas, class definitions, and validation designs. Unlike end-to-end pixel-based methods, the object-level design also reduces salt-and-pepper noise and improves the geometric coherence of mapped patches (Figure 10), which is critical for practical urban management. The multi-stage pipeline nevertheless introduces the possibility of error propagation, because inaccuracies in semantic segmentation, object construction, or temporal prediction may affect downstream classification.

5.2. Limitations and Future Directions

Several limitations merit discussion. The most significant is the resolution gap between Sentinel-2 temporal data (10 m) and VHR imagery (0.27 m). One 10 m Sentinel-2 pixel corresponds to approximately 1370 pixels at a 0.27 m ground sampling distance, meaning that small or narrow green patches may fall within a temporally contaminated pixel dominated by surrounding impervious surfaces. The mixed-pixel effect may have contributed to the lower farmland F1-score (87.6%), together with the spatial heterogeneity and smaller reference sample of this class. The PTF metrics (Coverage, Drift rate, MIDE, IDN) partially mitigate this effect but cannot fully compensate when temporal signals are severely contaminated.

Shadow interference poses a secondary challenge. Building shadows simultaneously reduce spectral contrast in VHR imagery and suppress NDVI values in time-series data, distorting phenological signals. DE-SegFormer's enhanced boundary recovery partially addresses the segmentation aspect, but temporal modeling remains vulnerable to persistent shadow effects in year-round shaded corridors, as visible in occasional misclassifications along building edges (Figure 10).

Additionally, the nine independently trained TimesNet models address cloud cover heterogeneity across sub-regions but may introduce boundary inconsistencies. Although GSOP-level PTF aggregation may reduce pixel-level discrepancies, future work should assess cross-boundary consistency and explore region-specific normalization and transfer learning toward a unified temporal model.

To address these limitations, future research should investigate higher-spatial-resolution, high-revisit time-series imagery, such as PlanetScope data. Incorporating complementary feature modalities, such as GLCM texture descriptors or LiDAR-derived height information, could further enhance discrimination in challenging scenarios. Because validation was limited to Chengdu, the reported accuracy should be interpreted as site-specific.

Cross-city validation across different climatic zones and urban morphologies is needed to assess the framework's generalizability.

6. Conclusions

This study developed a pixel-to-object framework that integrates 0.27 m VHR imagery with Sentinel-2 NDVI time series for fine-grained urban vegetation classification. The results yield four main findings. First, DE-SegFormer improved the F1-score and IoU of binary green-space segmentation by 4.8 and 7.2 percentage points, respectively, relative to SegFormer, although at substantially higher computational cost. Second, TimesNet-based temporal modeling reduces confusion between grasslands and farmlands that remains difficult to address using single-date VHR imagery, achieving a temporal classification OA of 95.1%. Third, the patch-level phenological transfer metrics (Coverage, Drift rate, MIDE, IDN) provide an object-level mechanism for integrating 10 m temporal predictions with 0.27 m VHR-derived GSOPs. Fourth, the complete framework achieved an OA of 91.6% and a Mean F1-score of 90.7%, representing improvements of 13.1 and 15.7 percentage points over the ISFs-only configuration, respectively. The final F1-scores were 92.5%, 91.9%, and 87.6% for urban trees, grasslands, and farmlands, confirming the importance of phenological information for distinguishing spectrally similar vegetation types. These results demonstrate the potential of combining semantic object delineation with temporal phenological information to reduce boundary fragmentation and improve the discrimination of spectrally similar urban vegetation types.

Author Contributions: Conceptualization, X.C. and Z.S.; methodology, C.Z. and X.C.; validation, H.W. (Haiyu Wang); formal analysis, H.W. (Haoyu Wang); investigation, C.Z.; data curation, C.Z., H.L. and T.L.; writing—original draft preparation, C.Z. and X.C.; writing—review and editing, X.C.; visualization, C.Z. and T.L.; supervision, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Chengdu University of Technology “AI Research Fund” Program (2025AI038); the Yunnan Sci-Tech Talent and Platform Plan Project (202605AF350036); and the “Engineering Research Center for Innovation Application of Low-Altitude Remote Sensing in Plateau and Mountainous Regions (Incubation)” Project of the Yunnan Province Department of Education.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Acknowledgments: During the revision of this manuscript, the authors used ChatGPT (model GPT-5.6 Sol, OpenAI) solely for language polishing, grammar checking, and minor wording refinement. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Burrewar, S.S.; Haque, M.; Haider, T.U. A Survey on Mapping of Urban Green Spaces within Remote Sensing Data Using Machine Learning & Deep Learning Techniques. In *Proceedings of the 2023 15th International Conference on Computer and Automation Engineering (ICCAE), Sydney, Australia, 3 March 2023*; IEEE: New York, NY, USA, 2023; pp. 30–34.
2. Xiao, J.; Chen, L.; Zhang, T.; Li, L.; Yu, Z.; Wu, R.; Bai, L.; Xiao, J.; Chen, L. Identification of Urban Green Space Types and Estimation of Above-Ground Biomass Using Sentinel-1 and Sentinel-2 Data. *Forests* **2022**, *13*, 1077. [[CrossRef](#)]
3. Adams, M.P.; Smith, P.L. A Systematic Approach to Model the Influence of the Type and Density of Vegetation Cover on Urban Heat Using Remote Sensing. *Landsc. Urban Plan.* **2014**, *132*, 47–54. [[CrossRef](#)]
4. Amoatey, P.; Sulaiman, H. Quantifying Carbon Storage Potential of Urban Plantations and Landscapes in Muscat, Oman. *Environ. Dev. Sustain.* **2020**, *22*, 7969–7984. [[CrossRef](#)]

5. Brack, C.L. Pollution Mitigation and Carbon Sequestration by an Urban Forest. *Environ. Pollut.* **2002**, *116*, S195–S200. [[CrossRef](#)] [[PubMed](#)]
6. Zoulia, I.; Santamouris, M.; Dimoudi, A. Monitoring the Effect of Urban Green Areas on the Heat Island in Athens. *Environ. Monit. Assess.* **2009**, *156*, 275–292. [[CrossRef](#)] [[PubMed](#)]
7. Ma, Y.; Yang, J. A Review of Methods for Quantifying Urban Ecosystem Services. *Landsc. Urban Plan.* **2025**, *253*, 105215. [[CrossRef](#)]
8. Athokpam, V.; Chamroy, T.; Ngairangbam, H. The Role of Urban Green Spaces in Mitigating Climate Change: An Integrative Review of Ecological, Social, and Health Benefits. *Environ. Rep.* **2024**, *6*, 10–14. [[CrossRef](#)]
9. Byomkesh, T.; Nakagoshi, N.; Dewan, A.M. Urbanization and Green Space Dynamics in Greater Dhaka, Bangladesh. *Landsc. Ecol. Eng.* **2012**, *8*, 45–58. [[CrossRef](#)]
10. Shahtahmassebi, A.R.; Li, C.; Fan, Y.; Wu, Y.; Lin, Y.; Gan, M.; Wang, K.; Malik, A.; Blackburn, G.A. Remote Sensing of Urban Green Spaces: A Review. *Urban For. Urban Green.* **2021**, *57*, 126946. [[CrossRef](#)]
11. Xu, Z.; Zhou, Y.; Wang, S.; Wang, L.; Li, F.; Wang, S.; Wang, Z. A Novel Intelligent Classification Method for Urban Green Space Based on High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3845. [[CrossRef](#)]
12. Gao, L.; Wang, X.; Johnson, B.A.; Tian, Q.; Wang, Y.; Verrelst, J.; Mu, X.; Gu, X. Remote Sensing Algorithms for Estimation of Fractional Vegetation Cover Using Pure Vegetation Index Values: A Review. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 364–377. [[CrossRef](#)] [[PubMed](#)]
13. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.R.; Murayama, Y.; Ranagalage, M. Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sens.* **2020**, *12*, 2291. [[CrossRef](#)]
14. Wu, Y.; Zhang, M. Swin-CFNet: An Attempt at Fine-Grained Urban Green Space Classification Using Swin Transformer and Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2503405. [[CrossRef](#)]
15. Hoeser, T.; Bachofer, F.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review—Part II: Applications. *Remote Sens.* **2020**, *12*, 3053. [[CrossRef](#)]
16. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [[CrossRef](#)]
17. Huang, Y.; Wang, L.; Zhao, P.; Zhao, Y.; Yang, Q.; Du, Y.; Ling, F. Deep Learning in Urban Green Space Extraction in Remote Sensing: A Comprehensive Systematic Review. *Int. J. Remote Sens.* **2024**, *46*, 1117–1150. [[CrossRef](#)]
18. Lee, D.H.; Park, H.Y.; Lee, J. A Review on Recent Deep Learning-Based Semantic Segmentation for Urban Greenness Measurement. *Sensors* **2024**, *24*, 2245. [[CrossRef](#)] [[PubMed](#)]
19. Cheng, Y.; Wang, W.; Ren, Z.; Zhao, Y.; Liao, Y.; Ge, Y.; Wang, J.; He, J.; Gu, Y.; Wang, Y.; et al. Multi-Scale Feature Fusion and Transformer Network for Urban Green Space Segmentation from High-Resolution Remote Sensing Images. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103514. [[CrossRef](#)]
20. Wang, W.; Cheng, Y.; Ren, Z.; He, J.; Zhao, Y.; Wang, J.; Zhang, W. A Novel Hybrid Method for Urban Green Space Segmentation from High-Resolution Remote Sensing Images. *Remote Sens.* **2023**, *15*, 5472. [[CrossRef](#)]
21. Zhang, H.; Liu, S. Double-Branch Multi-Scale Contextual Network: A Model for Multi-Scale Street Tree Segmentation in High-Resolution Remote Sensing Images. *Sensors* **2024**, *24*, 1110. [[CrossRef](#)] [[PubMed](#)]
22. Guo, Y.; Shao, Q.; Liu, F.; Lin, W.; Wang, M.; Li, Z.; Chen, S.; Liu, H.; Su, J.; Wang, X. Urban Green Space Extraction from BJ-2 Remote Sensing Image Based SegFormer Semantic Segmentation Model. *Can. J. Remote Sens.* **2025**, *51*, 2439835. [[CrossRef](#)]
23. Hofmann, P.; Strobl, J.; Nazarkulova, A. Mapping Green Spaces in Bishkek—How Reliable Can Spatial Analysis Be? *Remote Sens.* **2011**, *3*, 1088–1103. [[CrossRef](#)]
24. Wang, H.; Wang, J.; Shen, Z.; Zhang, Z.; Li, J.; Zhao, L.; Jiao, S.; Li, S.; Lei, Y.; Kou, W.; et al. Parcel-Level Mapping of Apple Orchard in Smallholder Agriculture Areas Based on Feature-Level Fusion of VHR Image and Time-Series Images. *Int. J. Remote Sens.* **2022**, *43*, 6195–6220. [[CrossRef](#)]
25. Xu, Z.; Chen, J.; Niu, H.; Fan, R.; Lu, D.; Feng, R. Boosting Urban Openspace Mapping with the Enhancement Feature Fusion of Object Geometry Prior Information from Vision Foundation Model. *Remote Sens.* **2025**, *17*, 1230. [[CrossRef](#)]
26. Neyns, R.; Canters, F. Mapping of Urban Vegetation with High-Resolution Remote Sensing: A Review. *Remote Sens.* **2022**, *14*, 1031. [[CrossRef](#)]
27. Hansen, M.C.; Egorov, A.; Potapov, P.V.; Stehman, S.V.; Tyukavina, A.; Turubanova, S.A.; Roy, D.P.; Goetz, S.J.; Loveland, T.R.; Ju, J.; et al. Monitoring Conterminous United States (CONUS) Land Cover Change with Web-Enabled Landsat Data (WELD). *Remote Sens. Environ.* **2014**, *140*, 466–484. [[CrossRef](#)]
28. Cai, Y.; Guan, K.; Peng, J.; Wang, S.; Seifert, C.; Wardlow, B.; Li, Z. A High-Performance and in-Season Classification System of Field-Level Crop Types Using Time-Series Landsat Data and a Machine Learning Approach. *Remote Sens. Environ.* **2018**, *210*, 35–47. [[CrossRef](#)]
29. Foerster, S.; Kaden, K.; Foerster, M.; Itzerott, S. Crop Type Mapping Using Spectral–Temporal Profiles and Phenological Information. *Comput. Electron. Agric.* **2012**, *89*, 30–40. [[CrossRef](#)]

30. Rustowicz, R.; Cheong, R.; Wang, L.; Ermon, S.; Burke, M.; Lobell, D. Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019.
31. Jiao, S.; Hu, D.; Shen, Z.; Wang, H.; Dong, W.; Guo, Y.; Li, S.; Lei, Y.; Kou, W.; Wang, J.; et al. Parcel-Level Mapping of Horticultural Crop Orchards in Complex Mountain Areas Using VHR and Time-Series Images. *Remote Sens.* **2022**, *14*, 2015. [[CrossRef](#)]
32. Li, D.; Ke, Y.; Gong, H.; Li, X. Object-Based Urban Tree Species Classification Using Bi-Temporal WorldView-2 and WorldView-3 Images. *Remote Sens.* **2015**, *7*, 16917–16937. [[CrossRef](#)]
33. Yan, J.; Zhou, W.; Han, L.; Qian, Y. Mapping Vegetation Functional Types in Urban Areas with WorldView-2 Imagery: Integrating Object-Based Classification with Phenology. *Urban For. Urban Green.* **2018**, *31*, 230–240. [[CrossRef](#)]
34. Bressant, C.; Herrault, P.-A.; Puissant, A. Fine-Scale Phenology of Urban Trees from Satellite Image Time Series: Toward a Comprehensive Analysis of Influencing Factors. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11685–11706. [[CrossRef](#)]
35. Wenger, R.; Bressant, C.; Roettele, L.; Forestier, G.; Puissant, A. Improving Urban Tree Species Classification with High Resolution Satellite Imagery and Machine Learning. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 4679–4682.
36. Zhang, H.; Ning, X.; Shao, Z.; Wang, H. Spatiotemporal Pattern Analysis of China’s Cities Based on High-Resolution Imagery from 2000 to 2015. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 241. [[CrossRef](#)]
37. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [[CrossRef](#)] [[PubMed](#)]
38. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
39. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
40. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [[CrossRef](#)] [[PubMed](#)]
41. Peng, B.; Zhang, L.; Zhang, D. Automatic Image Segmentation by Dynamic Region Merging. *IEEE Trans. Image Process.* **2011**, *20*, 3592–3605. [[CrossRef](#)] [[PubMed](#)]
42. Sabaneh, K.; Sabha, M. Improving SLIC Superpixel by Color Difference-Based Region Merging. *Multimed. Tools Appl.* **2024**, *83*, 47943–47961. [[CrossRef](#)]
43. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies—Part 2: Recommendations and Best Practices. *Remote Sens.* **2021**, *13*, 2591. [[CrossRef](#)]
44. Congalton, R.G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sens. Environ.* **1991**, *37*, 35–46. [[CrossRef](#)]
45. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [[CrossRef](#)]
46. Khalid, N. Review on Region-Based Segmentation Using Watershed and Region Growing Techniques and Their Applications in Different Fields. *J. La Multiapp* **2022**, *3*, 241–249. [[CrossRef](#)]
47. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In Proceedings of the 11th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
48. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; PMLR: Cambridge, MA, USA, 2022; pp. 27268–27286.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.