International Journal of Applied Earth Observation and Geoinformation 124 (2023) 103514

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Multi-scale Feature Fusion and Transformer Network for urban green space segmentation from high-resolution remote sensing images

Yong Cheng[a], Wei Wang[a], Zhoupeng Ren[b,*], Yingfen Zhao[c], Yilan Liao[b], Yong Ge[d,e,*], Jun Wang[a], Jiaxin He[a], Yakang Gu[a], Yixuan Wang[a], Wenjie Zhang[a,b], Ce Zhang[f,g]

[a] Nanjing University of Information Science & Technology, Nanjing 210044, China
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[c] China Centre for Resources Satellite Data and Application, Beijing 100094, China
[d] Key Laboratory of Poyang Lake Wetland and Watershed Research, Ministry of Education, Nanchang 330022, China
[e] School of Geography and Environment, Jiangxi Normal University, Nanchang 330022, China
[f] UK Centre for Ecology & Hydrology, Library Avenue, Lancaster LA1 4AP, UK
[g] School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK

ARTICLE INFO

Keywords:
Urban green space
Deep learning
Multi-scale feature fusion
Vegetation feature
High-resolution remote sensing images

ABSTRACT

Accurate extraction of urban green space is critical for preserving urban ecological balance and enhancing urban life quality. However, due to the complex urban green space morphology (e.g., different sizes and shapes), it is still challenging to extract green space effectively from high-resolution image. To address this issue, we proposed a novel hybrid method, Multi-scale Feature Fusion and Transformer Network (MFFTNet), as a new deep learning approach for extracting urban green space from high-resolution (GF-2) image. Our method was characterized by two aspects: (1) a multi-scale feature fusion module and transformer network that enhanced the recovery of green space edge information and (2) vegetation feature (NDVI) that highlighted vegetation information and enhanced vegetation boundaries identification. The GF-2 image was utilized to build two urban green space labeled datasets, namely Greenfield and Greenfield2. We compared the proposed MFFTNet with the existing popular deep learning models (like PSPNet, DensASPP, etc.) to evaluate the effectiveness of MFFTNet by the Mean Intersection Over Union (MIOU) benchmark on Greenfield, Greenfield2, and a public dataset (WHDLD). Experiments on Greenfield2 showed that MFFTNet can achieve a high MIOU (86.50%), which outperformed deep learning networks like PSPNet and DensASPP by 0.86% and 3.28%, respectively. Meanwhile, the MIOU of MFFTNet incorporating vegetation feature (NDVI) was further achieved to 86.76% on Greenfield2. Our experimental results demonstrate that the proposed MFFTNet with vegetation feature (NDVI) outperforms the state-of-the-art methods in urban green space segmentation.

1. Introduction

Urban green space plays a pivotal role in the ecosystem of the urban landscape and has significant connections with the urban ecological environment (Thompson et al., 2012), people's health (Astell-Burt et al., 2022), and welfare (Wang et al., 2021). It has been found that urban green space coverage has a negative correlation with heat island intensity and that increasing green space coverage has considerable impact on heat island mitigation (Wang et al., 2021). In addition, urban green space provides a suitable leisure platform for residents' lives (Tu et al., 2019), decreases stress and anxiety (Bertram and Rehdanz, 2015), and promotes healthy living (Hills et al., 2019). Unfortunately, the massive expansion of urbanisation and human activities have resulted in the encroachment upon the expanse of green space (Portillo-Quintero et al., 2012). This situation poses a severe threat to the urban ecological environment (Su et al., 2011). Consequently, there is urgent need for rapid and precise extraction and monitoring techniques for urban green space as they are pivotal to ensure the sustainable and healthy development of urban areas.

With the progress in remote sensing technology, a plethora of multi-temporal high-resolution image data is acquired using diverse sensors (Zhang et al., 2019). These high-resolution images provide intricate

* Corresponding authors.
E-mail addresses: renzp@lreis.ac.cn (Z. Ren), gey@lreis.ac.cn (Y. Ge).

https://doi.org/10.1016/j.jag.2023.103514
Received 27 July 2023; Received in revised form 21 September 2023; Accepted 29 September 2023
Available online 5 October 2023

details that are imperceptible to the naked eye, enabling the exploration of internal urban structures and they serve as invaluable resources for conducting comprehensive surveys and mapping of urban greenfield (Yin et al., 2021). By analyzing high-resolution remote sensing, a comprehensive understanding of urban green space can be achieved, facilitating informed decision-making and strategic urban planning and management. Researchers employ remote sensing images to extract information pertaining to urban green space. Currently, the methods for extracting such information from images can be categorized into four distinct types: the threshold method (Myeong et al., 2006), pixel-based classification method (Liu and Yue, 2010), object-based image analysis approach (Ardila et al., 2012), and deep learning method (Xu et al., 2020). The threshold technique is commonly used to distinguish vegetation by utilizing spectral indices like the Normalized Difference Vegetation Index (NDVI) (Tucker et al., 2005). However, the presence of complex urban backgrounds, including buildings and highways, often leads to interference and disruption of vegetation feature in remote sensing imagery (Neyns and Canters, 2022). The pixel-based technique relies on the properties of distinct wavebands to extract green patches from basic image backgrounds, while the method is largely used for low- and medium-resolution images (Räsänen and Virtanen, 2019). In an object-based approach, identification of urban green space as a whole, with noise resistance and wide applicability, and has achieved successful applications in the study of vegetation high-resolution image data (Wang et al., 2020).However, segmenting urban green space from high-resolution image data is a complicated data processing operation, and high-resolution images can provide detailed feature information while simultaneously increasing intra-target class variation (Liu et al., 2016). Above approaches, such as thresholding, require manually created features to extract greenfield information, often result in poor generalization (Spiering et al., 2020), and lack an autonomous learning process; thus, improved methods to increase the efficiency and accuracy of green space extraction are urgently needed.

Deep learning-based remote sensing imagery segmentation has gained significant prominence in recent years as computer vision techniques have advanced (He et al., 2022; Li et al., 2023). Deep learning can be described as a hierarchical feature representation network with strong capability to automatically learn complex feature representations from enormous data sets such as spectrum, texture, shape, and context (Hinton et al., 2006). Deep learning is now being selected to solve a variety of problems, including object detection (Jiang et al., 2022) and semantic segmentation (Chen et al., 2021). Numerous studies have demonstrated that Fully Convolutional Networks (FCNs) increase the accuracy of target feature extraction because of their powerful end-to-end feature representation and pixel-level segmentation capabilities (Long et al., 2015). Since these studies, several semantic segmentation models based on FCNs have been developed, including SegNet (Badrinarayanan et al., 2017), DeepLabv3+ (Chen et al., 2018), DenseASPP (Yang et al., 2018), and PSPNet (Zhao et al., 2017). Although these models have shown capabilities in semantic segmentation tasks, direct application upon urban greenfield extraction tasks is difficult due to the complex urban landmark structure, often composed of different elements such as buildings and roads, each with its own features such as spectral characteristics, texture, and spatial context. Consequently, researchers proposed different models and networks according to the characteristics of urban greenfield on the basis of the above networks, such as Xu et al. (2020) and Kattenborn et al. (2021). However, only retrieving shallow information is no longer sufficient for the task of green space segmentation. Feature extraction generates abundant information across different levels in the feature maps during feature extraction, and accurate segmentation needs to take into account the contextual relationship between green spaces and their surroundings. In addition, a single convolution used in Xu et al. (2020) can only capture local image features and cannot effectively fuse urban green space information extracted from high-resolution images, thus leading to poor accuracy of segmentation results. What's more, the intricate and diverse spatial scales of features in different remote sensing images also bring huge challenges to model feature extraction, and although some scholars (Kuai et al., 2022) have developed a multiscale feature extraction module using the concept of cavity convolution to enhance segmentation accuracy, its reception field is limited, and the multi-scale feature information of the object green space has not been fully explored.

In this study, we present an innovative approach, namely the Multi-scale Feature Fusion and Transformer Network (MFFTNet), to address the aforementioned challenges of poor feature information fusion and limited perceptual field in urban green space segmentation from high-resolution image data. To capture sharp green space objects by gradually recovering spatial information, MFFTNet used an encoding–decoding structure and built a fusion module in the encoder's feature map to improve the integrity of the generated green space images. To collect multi-scale context information, the encoder's feature maps were sent through a transformer, and ultimately, vegetation feature (NDVI) was included to enhance the training. The key contributions of this research included:

1. A deep learning network, MFFTNet, was proposed based on the encoding–decoding framework for automatical extraction of urban greenfield from high-resolution image data, and the ablation experiments showed that the optimized backbone, transformer, and fusion module in the MFFTNet significantly enhanced the performance.
2. Using the GF-2 images collected from Changping District, Beijing, two deep learning urban green space labeled datasets (Greenfield and Greenfield2) were built, which contributed to the research and application of deep learning in urban green space segmentation. Comparative experiments on the WHDLD, Greenfield, and Greenfield2 datasets confirmed that MFFTNet outperformed other deep learning networks and that the model was robust and efficient.
3. Vegetation feature (NDVI) was incorporated into the MFFTNet for urban green space segmentation study. The results demonstrated that incorporating NDVI could increase the richness of MFFTNet learning and optimize segmentation results, demonstrating the efficiency of vegetation feature in improving urban green space segmentation.

2. Study area and dataset construction

Existing public semantic segmentation datasets, such as ISPRS-Vaihingen and Potsdam (https://www.isprs.org/commissions/comm3/wg4/semantic-labeling.html), are used for characterizing land cover surfaces, such as impervious surfaces, buildings, and other structures, which do not adequately represent the complex urban green space. For this reason, Men et al. (2021) employed high-resolution images to create urban green space labeled datasets; nevertheless, these high-resolution image urban green space datasets are not open source to be used freely. As a result, developing a high-quality, high-resolution urban green space labeled dataset remains a pressing issue to be addressed.

2.1. Study area

In Beijing, China's capital, the creation of "green concepts" is highly valued. As a mega-city with rapid urbanization in China's northern plains, it is an ideal location to investigate the response of urban green space to urbanization, having seen significant urban sprawl in recent decades (Zhang et al., 2022). Changping District in Beijing was selected as the study area in this research (Fig. 1). The Changping District is located in the mid-latitude zone, and the vegetation consists primarily of deciduous broad-leaved woods with a low number of evergreen trees.

2.2. Dataset construction

This experiment used high-resolution images (GF-2, June 4, 2017) as the data source to effectively extract urban green space. The data was

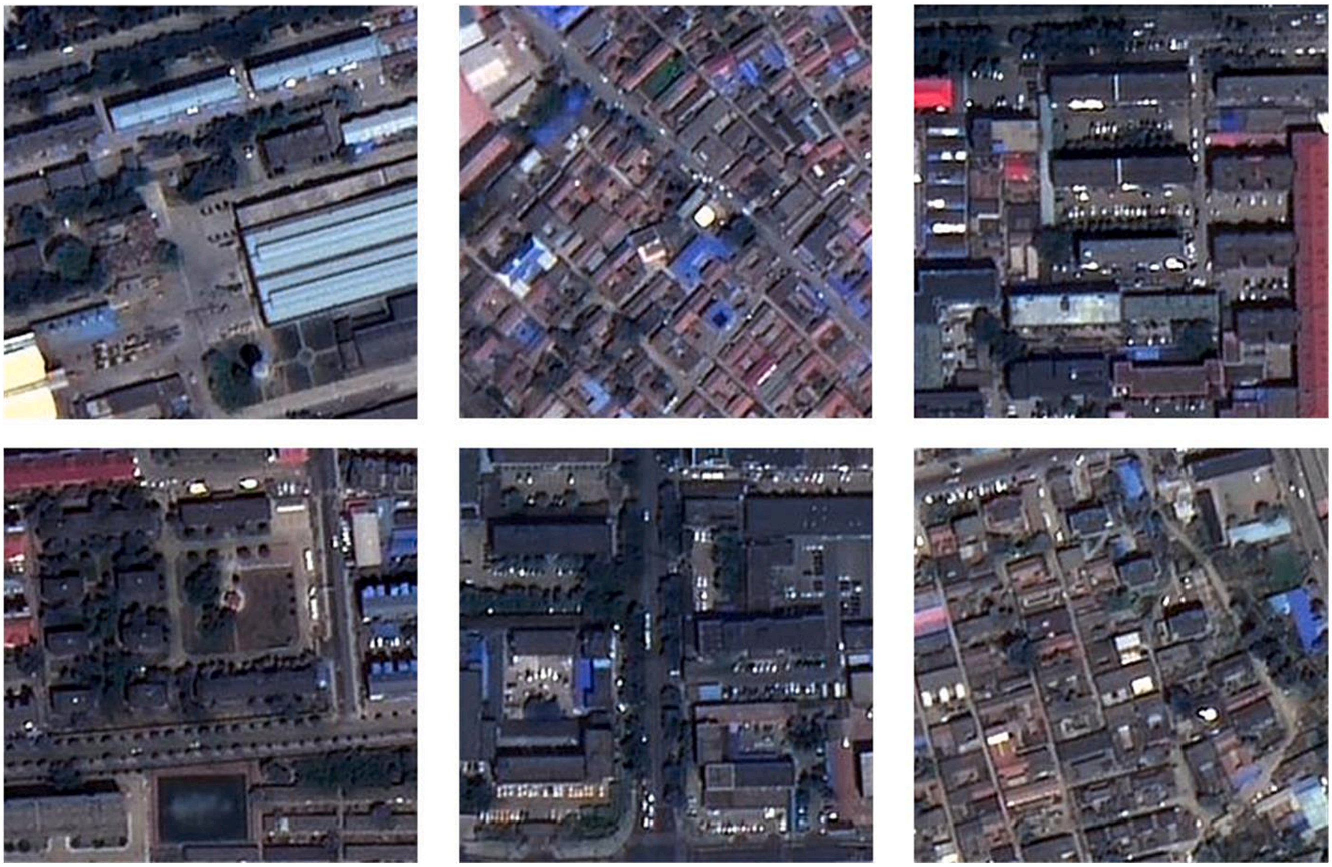

Fig. 1. Example map of some greenfield in the study area.

obtained from the China Resources Satellite Application Center. The GF-2 PMS (panchromatic multispectral sensor) has a 1-m spatial resolution panchromatic band and four 4-m spatial resolution multispectral bands, including blue, green, red, and near-infrared. The spatial resolution of the sub-satellite point reaches 0.8 m with a revisit time of 5 days. The data pre-processing includes orthographic correction, image fusion, and clipping.

All processed images were labeled at the pixel level, and classified into two categories: green space and background, with backdrop pixels having RGB of (0, 0, 0) and greenfield pixels having RGB of (0, 255, 0), as shown in Fig. 2. Image cropping was utilized to partition the original image and the label image into 256 × 256 pixels to accommodate the limited computational resources, and the final dataset of urban greenfield images, Greenfield, was obtained by utilizing the training, validation and test data split as 3:1:1. The number of image pairs of training set, validation set and test set were 1162, 387, and 387, respectively.

Typical false color image mixed with 4-3-2 bands (NIR, R, G) were used to demonstrate the urban green space features. All captured images were labeled at the pixel level, and classified into two categories: green space and background (Fig. 3). Image cropping was utilized to segment the original image and the label image into 256 × 256 pixels to accommodate the limited computational resources, as the second dataset (Greenfield2). The number of image pairs in the training set, validation set, and test set was 1109, 370, and 370, respectively.

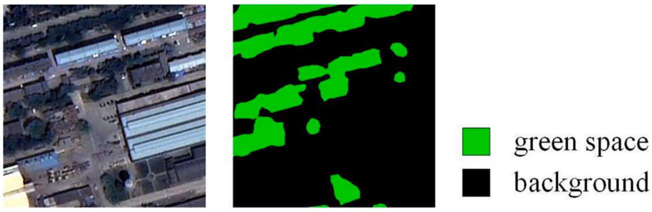

Fig. 2. Example of Greenfield dataset.

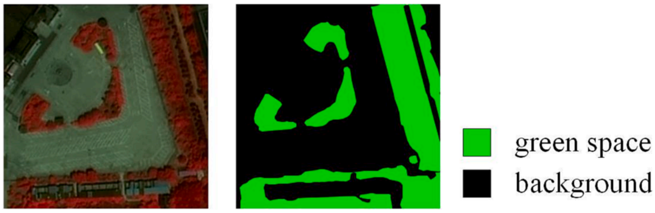

Fig. 3. Example of Greenfield2 dataset.

2.3. Vegetation feature

The spectral information in high-resolution image data is minimal, and the training data used for green space extraction is often RGB three-band information with insufficient feature richness, making high segmentation accuracy difficult to achieve. The green region shows low reflectivity in the visible band and high reflectance in the near-infrared band in GF-2. This research combines GF-2 multi-temporal remote sensing images and introduces vegetation feature for urban green areas. The NDVI (Tucker et al., 2005), a widely used index for plant growth assessment, is chosen as the vegetation feature. The formula is shown below:

$$NDVI = \frac{NIR - R}{NIR + R} \tag{1}$$

where NIR and R denote the near-infrared and red bands, respectively.

The NDVI of the study area was calculated by formula (1), and the vegetation feature (NDVI) was super-imposed onto the false color image, which was mixed with NDVI as the fourth band via image fusion to get the enhanced urban green space vegetation feature image data. The results of a conventional false color image and vegetation feature superimposition are shown in Fig. 4.

3. Methods

We proposed MFFTNet based on the codec framework to implement urban green space semantic segmentation. The MFFTNet encoder component extracted green space features from shallow to deep using Res2Net as the backbone and the transformer to synthesize green space context information. The Fusion module was built to extract multi-scale information from high-level features, which were then integrated with contextual information acquired from levels of network within the backbone. Finally, convolutional layers were used to generate the greenfield extraction. Fig. 5 depicts the overall workflow.

3.1. Multi-scale backbone

It is critical to extract high-precision feature information from images when segmenting urban green space semantically, however, as the number of encoder convolution layers grows, the visual information is lost significantly. This work used Res2Net as the backbone to improve image information extraction performance and minimize gradient disappearance.

Res2Net performs deep learning tasks by building residual connections with hierarchy within a single residual block rather than one single 3 × 3 convolution (Gao et al., 2019). Unlike ResNet (He et al., 2016), Res2Net presents a new multiscale combination strategy and cross-layer connection structure to further increase the network's representational capability and learning efficiency (Fig. 6). Res2Net, in particular, adds a new multiscale combination module to each residual block for extracting features at several target scales. By expanding the number and size of sub-blocks, the multiscale combination module can adaptively extend the network's perceptual field and feature characterization capability. To prevent feature information bottlenecks, Res2Net adds cross-layer connections between different levels of feature maps, allowing lower-level features to be transferred more fully to higher-level features. A multi-scale feature encoder with a Res2Net backbone was built to improve the extraction of deep features and multi-scale context information from urban green space. Res2Net's single 7 × 7 convolutional kernel extracts green space features, resulting in the model's inability to find reliable features and objects in images. To improve the model's representation performance, two 5 × 5 convolutional kernels were utilized instead. Furthermore, this paper used the Relu6 activation function for the Rectified Linear Unit (Relu) (Sandler et al., 2018). The Relu limit to a maximum output of 6 is Relu6. The numerical resolution of the model is improved by restricting the magnitude of the Relu6 function, while negative values are filtered out to enhance the overall generalization of capabilities.

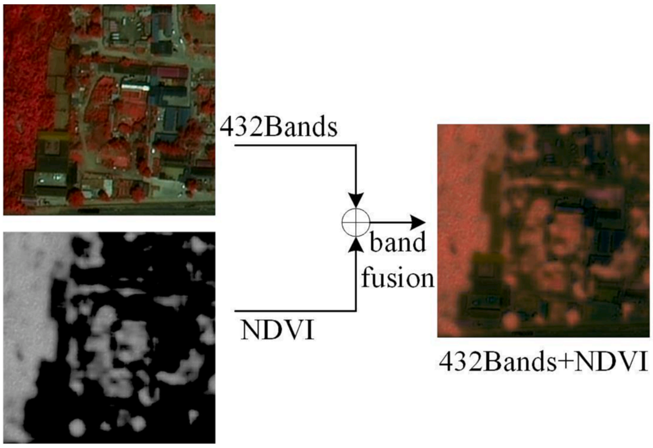

Fig. 4. False color image and NDVI feature overlay results.

3.2. Transformer module

The transformer has recently offered a significant boost to numerous computer vision methods. Given the computational complexity and parameter volume of the model, as well as the synthesis of information on the length and distance of the street tree green space, the EdgViT-related module was utilized as a component of the Transformer attention branch of this study (Pan et al., 2022). The EdgViT module introduces a local–global-local information exchange bottleneck, as illustrated in Fig. 7, through three key operations. Firstly, local aggregation utilizes deep convolution to integrate local information from neighboring tokens. Secondly, sparse attention globally provides a small collection of representative markers to facilitate long-distance information sharing via self-attention. Lastly, local propagation employs transposed convolution to propagate learned global context information from representative tokens to nearby tokens.

The model's sensing field limits the model's capacity to perceive green space targets in remote sensing image frames within a complicated city. To broaden the perceptual field, this paper introduced the EdgViT module, which can extract small-scale green space like street trees while also effectively identifying large-scale green space sections. Self-focusing enables effective learning of global information and long-range dependence, which aids in avoiding the interference of shadow occlusion of buildings, diversity of imaging conditions, and similarity of green space spectra with other features and enables the model to segment green space fractions effectively.

3.3. Fusion module

Green space features more intricate edge shapes than other objects, such as urban buildings and roadways. Fig. 8 depicts the structure of a multi-scale fusion module (Fusion) used in this paper to fuse information from multiple layers and refine edge details. Within the codec structure, the incorporation of rich high-level feature category information can greatly aid in the classification of low-level features. Similarly, leveraging low-level feature location information can significantly enhance the spatial positioning accuracy of high-level features.

The multi-scale feature fusion module (Fusion), depicted in Fig. 8, is made up of five parallel branches. To improve feature extraction capability, deep-level characteristics are first upscaled to align with the low-level features from the alternate branch. Subsequently, strip convolution is utilized to refine the information present within both deep and low-level feature sets. DOConv of sizes 1 × 3 and 3 × 1 (Cao et al., 2022), an activation function Gelu (Liu et al., 2022) and a Group Normalization (GN) layer (Wu and He, 2018) make up the majority of the bar convolution structure. Simultaneously, the input data is subjected to feature fusion by acting on one 1 × 1 convolutional kernel and three 3 × 3 convolutional kernels. Following the convolutional layer operation, the information taken from various branches is summed up and transferred jointly to the next layer of the network to produce feature information after a full connection. This paper uses DOConv instead of traditional convolution in the fusion module, which is a deep hyperparametric convolutional layer with additional learnable parameters, to solve the problems of slow convergence and insufficient generalization ability during deep convolutional neural network training. Dynamically creating convolution kernels based on diverse aspects of the input data improves model generalization and accuracy. DOConv integrates an adaptive channel attention mechanism to dynamically adjust channel weights, thereby enhancing the model's accuracy. Meanwhile, the Fusion module incorporates an asymmetric convolution kernel to boost

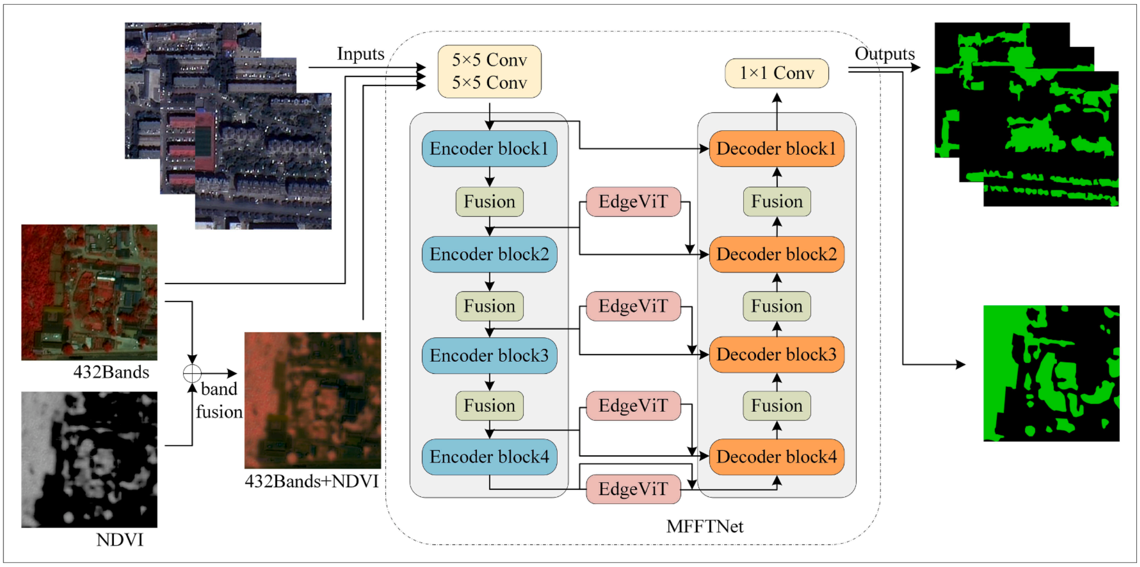

Fig. 5. Overall experimental workflow of the proposed MFFTNet.

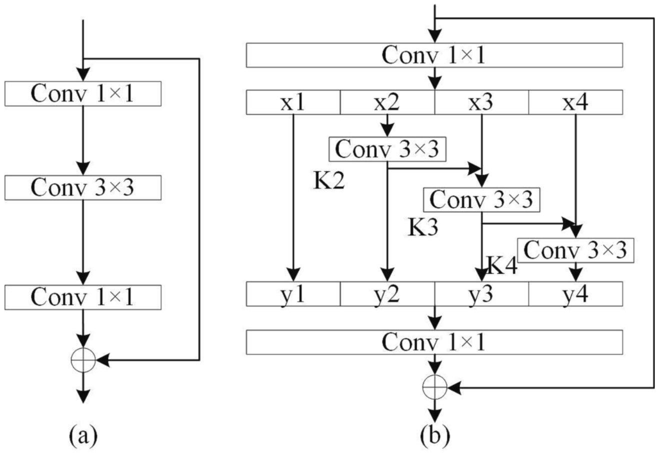

Fig. 6. (a) for the ResNet backbone; (b) for the Res2Net backbone. The x1, x2, x3, and x4 denoted the input feature maps for different scale branches in the Res2Net module; K2, K3, and K4 denoted the convolutional kernel sizes for feature fusion; and y1, y2, y3, and y4 denoted the output feature maps for different scale branches.

the attractiveness of greenfield characteristics by incorporating nonlinear changes. BN primarily addresses issues such as gradient explosions during the training phase. In urban green space semantic segmentation tasks, the batch size is frequently small, making BN training useless. As a result, we introduced GN, which improved green space segmentation performance by grouping channels.

4. Experiment

4.1. Experimental environment and evaluation metrics

The experiment was conducted on a Windows 10 platform equipped with an NVIDIA GeForce RTX 3060 GPU and 12 GB of graphics memory. The deep learning framework comprised PyTorch 1.7.1 and CUDA 11.6. The SGD optimizer, a cosine annealing strategy, and the Cross Entropy Loss function were used for network optimization, where the weight decay was set as 1e-4. We set the baseline learning rate as 0.001, the adjustment multiple to be 0.98, and the adjustment interval as 3. Moreover, the batch size was set as 2 when training, and the number of trainings was parameterized as 300.

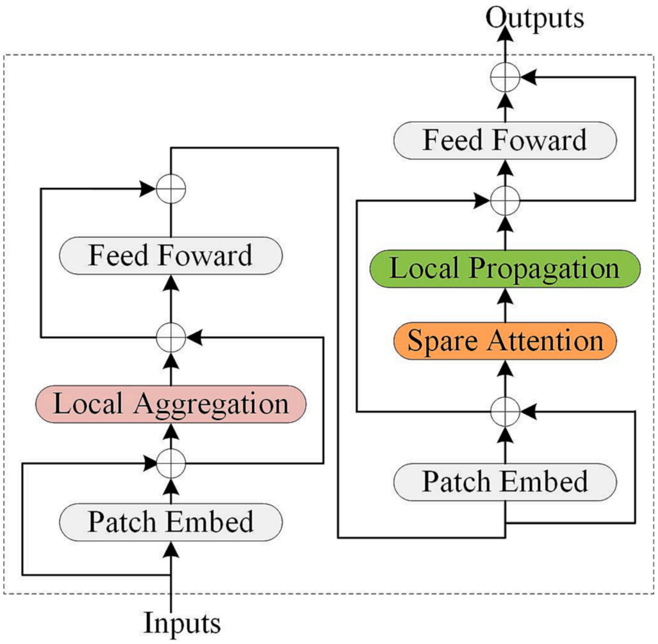

Fig. 7. EdgeViT module of the Transformer branch.

The model's resilience and efficacy were gauged using five quantitative measures: Pixel Accuracy (PA), Mean Pixel Accuracy (MPA), F1 score, Frequency Weighted Intersection Over Union (FWIOU), and Mean Intersection Over Union (MIOU). These metrics were utilized to compare the predicted values with the ground truth and evaluate the performance in practical applications. The formulas for the above evaluation metrics are shown below:

$$PA = \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{2}$$

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{3}$$

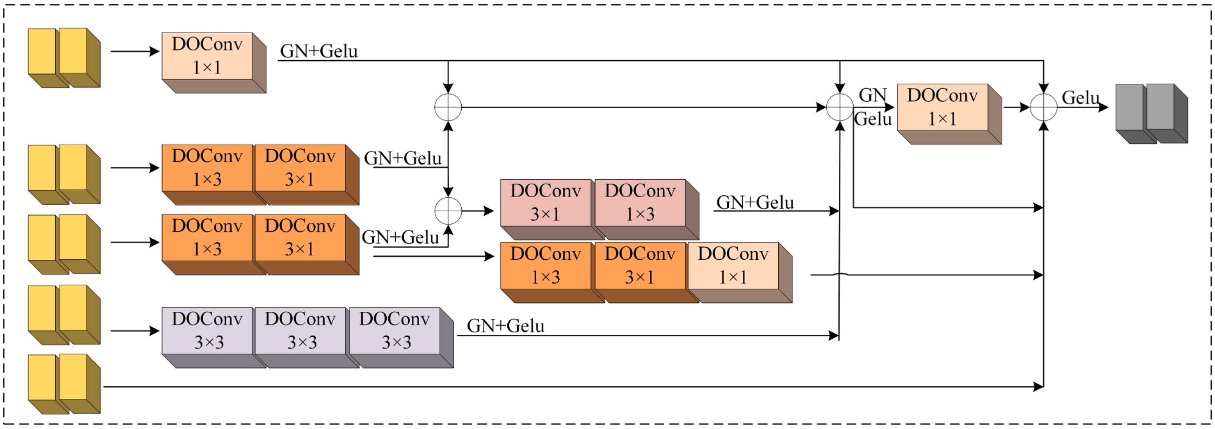

Fig. 8. Fusion module. DOConv denoted an over-parameterized convolutional layer, GN denoted Group Normalization, and Gelu denoted an activation function.

$$P = \frac{p_{ii}}{p_{ii} + p_{ij}} \tag{4}$$

$$R = \frac{p_{ii}}{p_{ii} + p_{ji}} \tag{5}$$

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{6}$$

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{7}$$

$$FWIOU = \frac{1}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} p_{ij} p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{8}$$

where P_{ii}, P_{ji}, P_{ij}, and P_{jj} represent the corresponding true and false positives, false and true negatives, respectively. Ambiguity arises when precision rate (P) and recall rate (R) are compared separately, so a reconciled average F1-score is introduced in this experiment to measure accuracy and recall together. In addition, the MIOU evaluates the resemblance between the true greenfield pixels and the predicted greenfield pixels, and the higher the MIOU value, the greater the similarity.

4.2. Ablation experiments

Table 1 summarizes the ablation tests performed on the Greenfield dataset to test the usefulness of each module. In the first row, baseline was used for feature extraction, and the extraction results were generated via straight upsampling. In this paper, U, DUC- Fusion, SK, Fusion, and EdgeViT modules were added in turn from the second to the last row, where U denoted two 5 × 5 convolutions and the Relu6 activation function, DUC- Fusion denoted a module consisting of Dense Upsampling Convolution (DUC) (Wang et al., 2018) and the Fusion module, SK denoted the selective kernel attention module (Li et al., 2019c), and EdgeViT denoted the Transformer module.

MIOU increases 0.34 % and PA improves 0.24 % when compared to the base network while using the optimized backbone, indicating that network performance is improved further (Table 1). Two 5 × 5 convolutional operations can improve the model's feature extraction performance, and the activation function Relu6 can boost the model's numerical resolution and overall generalization capacity. By constructing the DUC- Fusion module, the MIOU is improved by 0.04 %; the MIOU is also improved by constructing the SK module on top of it. MIOU is improved by 0.50 % as compared to the base network while using Fusion. The Fusion module fuses information from several layers, includes comprehensive category information for high-level features to guide the classification of low-level features, and supplements low-level feature location information with high-level features. Furthermore, when comparing SK and EdgeViT modules, EdgeViT outperforms SK by 0.11 % in MIOU. Overall, the modified module improves segmentation accuracy, and using the aforesaid improvement technique results in a 0.95 percentage point gain, demonstrating the efficacy of the improvement strategy.

Table 1
Ablation experiments on the Greenfield.

Method	PA (%)	MIOU (%)	Flops (G)	Params (M)
baseline	77.85	63.63	7.95	12.05
baseline + U	78.09	63.97	38.20	12.15
baseline + U + DUC-Fusion	78.17	64.01	44.13	26.13
baseline + U + DUC-Fusion + SK	78.18	64.03	134.45	55.62
baseline + U + Fusion	78.48	64.47	38.20	12.15
baseline + U + Fusion + EdgeViT	78.62	64.58	58.90	33.32

4.3. Comparison experiments

4.3.1. Comparison experiment of the WHDLD dataset

The WHDLD is a densely labeled dataset that originated from extensive remote sensing imagery of Wuhan. It has been specifically curated for semantic segmentation task (Shao et al., 2020). The dataset covers a diverse range of landforms, including building, road, pavement, vegetation, bare soil, and water. In this study, the WHDLD was optimized by designating five categories of non-urban green space as the background. Additionally, irrelevant data was eliminated, resulting in a refined dataset consisting of two categories: green space and background. In addition, FCN8s, SegNet, DeepLabv3+, DenseASPP, DensASPP (mobilenet), PSPNet, DFN (Yu et al., 2018), ShuffleNetV2 (Ma et al., 2018), DFANet (Li et al., 2019a), DABNet (Li et al., 2019b),

Table 2
Experimental results in the WHDLD.

Method	PA(%)	MPA(%)	F1(%)	MIOU(%)	FWIOU(%)
DFANet	88.23	86.64	82.04	77.29	79.24
DensASPP(mobilenet)	88.56	87.26	82.53	77.95	79.70
DensASPP	89.25	88.16	83.52	79.18	80.76
ESPNetv2	89.23	88.45	83.49	79.26	80.67
DFN	90.02	88.78	84.65	80.45	82.05
FCN8s	90.20	89.17	84.90	80.83	82.30
ShuffleNetV2	90.12	89.69	84.77	80.88	82.06
DeepLabv3+	90.88	90.08	85.89	82.08	83.39
DABNet	90.87	90.12	85.88	82.09	83.38
PSPNet	91.44	90.46	86.75	83.02	84.37
SegNet	91.70	91.35	87.08	83.67	84.72
MFFTNet	91.94	91.52	87.43	84.07	85.13

ESPNetv2 (Mehta et al., 2019), and MFFTNet were used for comparison experiments (Table 2).

FCN8s achieves pixel-level semantic segmentation by employing a full convolutional structure. SegNet continues to pool indexes in the codec structure in order to refine edge segmentation information. DeepLabv3+ adds arbitrary control over the resolution of encoder-extracted features in the codec structure, balancing accuracy and time via null convolution. DenseASPP segments targets using a densely connected structure. DensASPP (mobilenet) achieves quick segmentation tasks by utilizing a lightweight network, mobilenet. PSPNet presents a more extended, global contextual information integration network based on multiple image regions based on spatial pyramid pooling. DFN builds a top-down framework to optimize features at each level in order to get characteristics with inter-class distinctions and refine bounds. ShuffleNetV2 performs a quick segmentation task. With numerous connection architectures, DFANet includes a semantic segmentation coding module. DABNet uses asymmetric convolution and dilated convolution to efficiently generate bottleneck layers for improved segmentation performance. ESPNetv2 has developed a spatial pyramid of depth-expanding, separable convolutions that can be applied to edge devices.

Notably, the proposed model in this paper, MFFTNet, surpasses other networks in terms of urban green space segmentation accuracy across all metrics. The corresponding scores for the five metrics are as follows: PA, 91.94 %; MPA, 91.52 %; F1, 87.43 %; MIOU, 84.07 %; and FWIOU, 85.13 % (Table 2). Deeplabv3+, DABNet, and PSPNet fail to effectively identify the surrounding green space part and have poor classification results due to the influence of buildings, as shown in the first row of Fig. 9; in contrast, MFFTNet effectively identifies the green space around buildings and is close to the real surface condition. In the third rows of Fig. 9, each network can effectively identify green space for single presence; however, when the urban surface is complex, building shadows, the diversity of imaging conditions, and the similarity of green space spectra with other features inhibit the accurate estimation of green space extraction from remote sensing images, and Deeplabv3+, DABNet, and PSPNet do not effectively extract the green space part. The MFFTNet provides green space segmentation in difficult urban situations with high robustness by upgrading the network and performing multi-scale feature fusion.

4.3.2. *Comparison experiment of the Greenfield dataset*

A set of comparative experiments on the Greenfield dataset was chosen from current approaches to further test the effectiveness and rationale of MFFTNet in urban green space segmentation tasks. Table 3 displays the experimental outcomes.

Table 3 presents the performance metrics of various networks evaluated on the Greenfield dataset. The evaluation metrics utilized in this study include PA, MPA, F1, MIOU, and FWIOU. Notably, the MFFTNet exhibits the highest segmentation accuracy and outperforms other networks across all metrics for the urban greenfield segmentation task. Specifically, the scores for the five metrics are as follows: PA, 78.62 %; MPA, 79.02 %; F1, 70.80 %; MIOU, 64.58 %; and FWIOU, 64.66 % (Table 3).

Several example images from the Greenfield dataset were selected for experimentation; the objective was to demonstrate the model's capability in detecting objects of varying sizes, shapes, and distributions. The visualization of detection results can be observed in Fig. 10. It is clear that the MFFTNet can detect the majority of the objects. Deeplabv3+ classification results disregard a huge number of Greenfield components, and the classification results are unsatisfactory. The PSPNet classification results demonstrate that the targets' boundaries are reasonably smooth, but they fail to recognize minor elements of the areas, such as roadside trees (Fig. 10, second row f). Although SegNet and DenseASPP classification scores have improved, these two

Table 3
Experimental results in the Greenfield.

Method	PA(%)	MPA(%)	F1(%)	MIOU(%)	FWIOU(%)
DFANet	75.37	76.15	67.05	60.38	60.34
DensASPP(mobilenet)	77.10	77.62	69.01	62.58	62.61
PSPNet	77.49	78.42	69.64	63.19	63.12
DensASPP	77.58	78.13	69.60	63.23	63.24
ESPNetv2	77.75	78.21	69.77	63.43	63.47
SegNet	77.79	78.45	69.90	63.53	63.51
ShuffleNetV2	77.92	78.27	69.94	63.63	63.72
FCN8s	78.08	78.42	70.13	63.84	63.94
DeepLabv3+	78.07	78.55	70.16	63.86	63.91
DFN	78.14	78.52	70.21	63.93	64.01
DABNet	78.18	78.56	70.26	63.98	64.07
MFFTNet	78.62	79.02	70.80	64.58	64.66

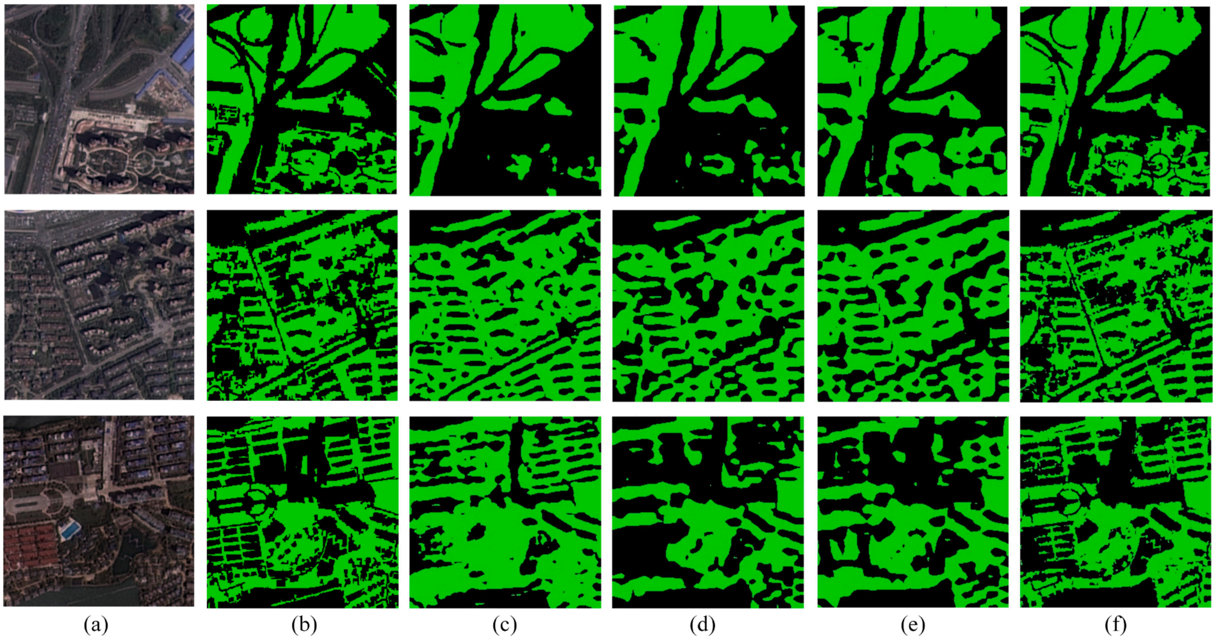

Fig. 9. Comparison experimental visualization results of WHDLD. (a) Input image. (b) Label. (c) DeepLabv3+. (d) DABNet. (e) PSPNet. (f) MFFTNet.

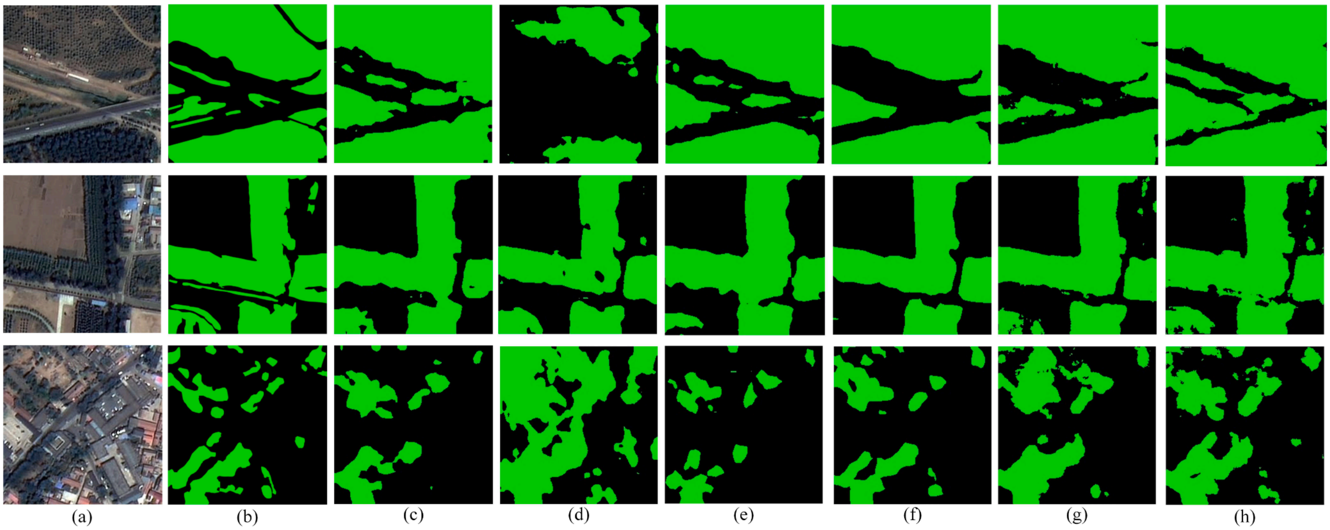

Fig. 10. Comparison experimental visualization results of Greenfield. (a) Input image. (b) Label. (c) DABNet. (d) DeepLabv3+. (e) DenseASPP. (f) PSPNet. (g) SegNet. (h) MFFTNet.

algorithms mistakenly classify some roads as green space. The graphic shows that MFFTNet's categorization outperforms SegNet and DenseASPP, and the segmented edges are more compatible with the actual edge features of green space. The MFFTNet MIOU scores are 1.15 %, 1.39 %, and 4.20 % higher than ESPNetv2, PSPNet, and DFANet, respectively. Experiments employing high-resolution remote data show that the approach has specific advantages in urban green space segmentation. In addition, results show that MFFTNet can detect small-area objects such as street trees more effectively than other networks such as SegNet, DeepLabv3+, and others (Fig. 10).

4.3.3. Comparison experiment of the Greenfield2 dataset

A set of comparative experiments on the Greenfield2 dataset was chosen from current approaches to further test the applicability and rationale of MFFTNet in urban green space segmentation tasks. Table 4 displays the experimental outcomes.

When compared to the Greenfield dataset, the Greenfield2 dataset highlights the greenfield portion more clearly. The scores of the five indicators of MFFTNet, the network suggested in this paper, in the Greenfield2 dataset are as follows: PA, 92.90 %; MPA, 92.73 %; F1, 89.53 %; MIOU, 86.50 %; and FWIOU, 86.75 % (Table 4). When compared to the performance of MFFTNet in the Greenfield dataset, MIOU achieved 86.50 % in the Greenfield2 dataset, a 21.92 % improvement. The results show that using conventional false color image in conjunction with 4-3-2 bands for urban Greenfield extraction can improve the representation of urban Greenfield features. Greenfield areas may be efficiently detected in both lush and sparse Greenfield coverage areas, as illustrated in Fig. 11, and the use of standard false color image effectively helps the urban Greenfield classification work. When compared to other advanced deep learning networks, the MFFTNet has a greater segmentation effect than SegNet and DABNet, and the MIOU is 0.37 %, 0.56 %, 0.86 %, and 2.21 % higher than SegNet, DABNet, PSPNet, and DFANet. The urban surface is complex, with numerous feature element categories and large object scale variations. In the second row of Fig. 11, PSPNet and SegNet are unable to recognize the occluded area due to the influence of building occlusion, whereas MFFTNet effectively segments the small green area in the occluded area via a local–global-local information connection; additionally, it is smoother for segmented edges, avoiding roughness like that which occurs with SegNet segmentation.

The accuracy of green space segmentation is increased by incorporating vegetation feature, with MIOU improving 0.26 %, PA increasing 0.13 %, MPA improving 0.17 %, F1 improving 0.19 %, and FWIOU improving 0.24 % (Table 4). Fig. 12(d)(e) shows that when vegetation feature is not incorporated, there are some areas with incorrect and missing scores, and inaccurate extraction of boundaries. When vegetation feature is incorporated, the incorrect scores are obviously reduced and the results are more consistent with the real surface conditions. Because of the similarity in the spectrum between red buildings and green regions, MFFTNet incorrectly recognizes buildings as green space in the second row of Fig. 12. When vegetation feature is incorporated, this erroneous segmentation is effectively prevented. The correct classification of green space in vast areas is essentially attained, while edge information is improved. The use of GF-2 remote data in conjunction with vegetation feature can considerably improve the segmentation of green space.

Table 4
Experimental results in the Greenfield2.

Method	PA(%)	MPA(%)	F1(%)	MIOU(%)	FWIOU(%)
DFN	89.82	88.93	85.28	80.94	81.73
DensASPP(mobilenet)	89.99	89.99	85.40	81.52	81.79
FCN8s	90.45	90.04	86.07	82.18	82.64
DensASPP	91.02	90.86	86.84	83.22	83.54
ShuffleNetV2	91.08	90.87	86.93	83.31	83.65
ESPNetv2	91.56	91.40	87.60	84.15	84.46
DFANet	91.64	91.50	87.72	84.29	84.59
PSPNet	92.46	92.05	88.93	85.64	86.04
DABNet	92.61	92.30	89.12	85.94	86.29
SegNet	92.71	92.51	89.23	86.13	86.43
DeepLabv3+	92.79	92.45	89.38	86.24	86.59
MFFTNet	92.90	92.73	89.53	86.50	86.75
MFFTNet + NDVI	93.03	92.90	89.72	86.76	86.99

5. Discussion

The ablation experimental investigation validates the efficacy of each improvement module in MFFTNet (Section 4.2). The ablation experimental analysis is carried out on the Greenfield dataset, starting with Res2Net as the baseline and gradually adding modules for the experiments. The performance of the model feature extraction is improved further by employing two 5 × 5 convolutional operations and introducing the activation function Relu6. The effectiveness of the improved backbone is demonstrated by increasing MIOU by 0.34 % (Table 1). The results show that in the decoder part, with the Fusion module, the MIOU is improved by 0.50 %, which can effectively fuse the multi-level features. In addition, compared with selective kernel attention, the EdgeViT

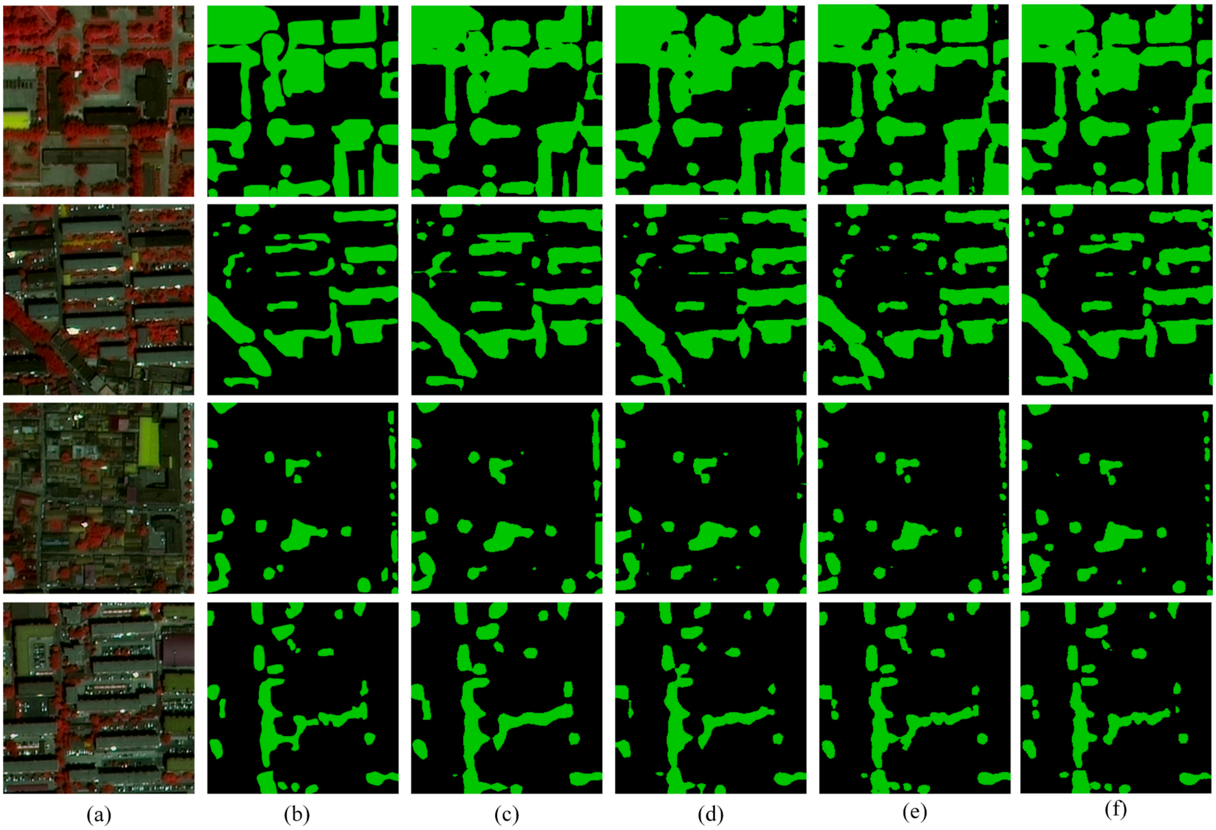

Fig. 11. Visualization results of comparison experiments on Greenfield2. (a) Input image. (b) Label. (c) PSPNet. (d) DABNet. (e) SegNet. (f) MFFTNet.

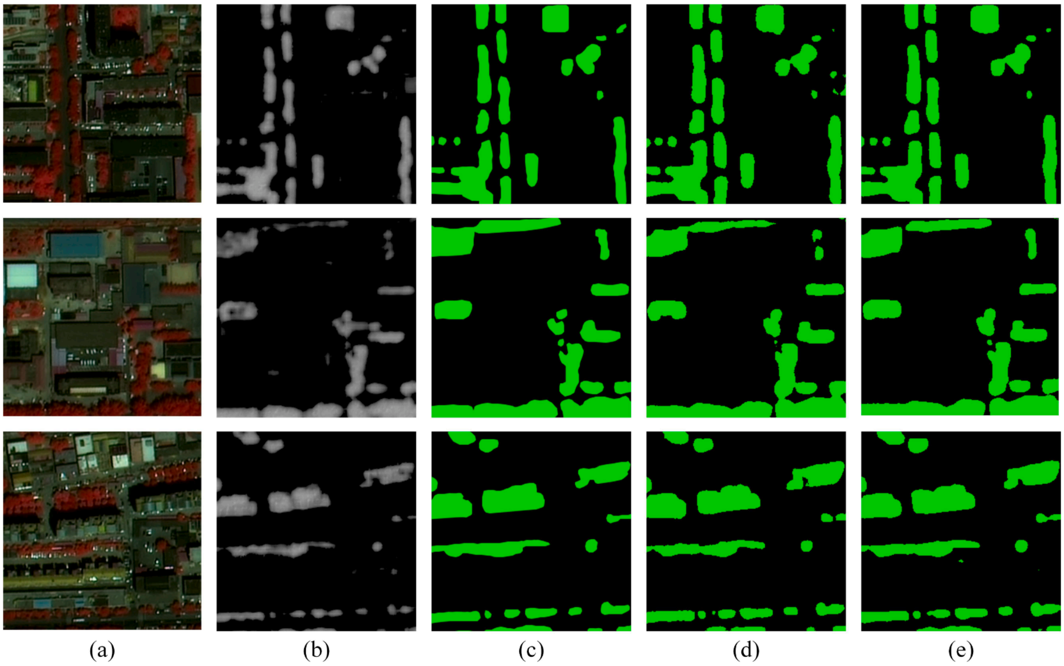

Fig. 12. Visualization results of comparison experiments on Greenfield2. (a) Input image. (b) NDVI. (c) Label. (d) MFFTNet. (e) MFFTNet + NDVI.

module performs better, with a 0.11 % improvement in MIOU. Compared with the single adaptive selection attention, the local–global-local information connection in the EdgeViT module is more applicable to the urban green space segmentation task, which can detect large green space while taking into account small areas such as street trees, thus improving the overall green space segmentation effect (Fig. 10). This series of enhancements demonstrates the effective segmentation of urban green space targets by each module of the model in this paper.

Large-scale urban green space segmentation requires fast construction and good transformation of the model. To verify the spatial generalization capability of MFFTNet, experiments were conducted on WHDLD, the Greenfield dataset, and the Greenfield2 dataset. To adapt to the urban Greenfield classification task, the public dataset WHDLD is optimized; MFFTNet achieves 84.07 % MIOU and can identify Greenfield areas in complex urban environments with clearer segmentation, which is better than other networks. In the Greenfield dataset for comparison experiments, MFFTNet performs better compared with other networks, and its MIOU scores are 1.15 %, 1.35 %, 1.39 %, and 4.2 % higher than ESPNetv2, DensASPP, PSPNet, and DFANet, respectively. Meanwhile, comparison experiments were conducted in the Greenfield2 dataset; the experiments proved that the extraction of urban green space using false color image could better represent the urban green space features, and the MIOU of MFFTNet reached 86.50 %. Comparing with Fig. 12, we found that the MFFTNet effectively segmented the small green space in the area affected by building shading and performed better than PSPNet and SegNet. After incorporating vegetation feature (NDVI), the vegetation information was more obvious, with a 0.26 % increase in MIOU (Table 4). In WHDLD for the public dataset and Greenfield and Greenfield2 datasets for an experimental area in Beijing, the highest MIOU is reached in all of them, which proves the effectiveness and robustness of the MFFTNet to partition the greenfield.

While MFFTNet achieves the highest segmentation accuracy, further optimization is possible in terms of model parameters. The combination of the fusion module and transformer effectively extracts green space feature information from the image. However, compared to the base network, MFFTNet has 50.95 G more floating points (Flops) and 21.27 M more parameters (Params), which affects the segmentation speed. The objects in remote sensing images also have substantial differences in geometric shape features, which results in the problem of scale variation of objects, so the model should have multi-scale segmentation capability. Meanwhile, the target can be clustered by linear discrimination to correctly present the internal sample's structure and thus improve the segmentation accuracy; the marginal distribution can be used to determine the category by testing the instances, which can improve the instance detection and generalization performance under multi-class supervision of the object (Zhu et al., 2022; Zhu et al., 2023). In future research, it is proposed to consider further deepening the research: (1) Commit to minimizing the weights; (2) The distribution of green space in different cities is different, and the green space extracted in this study are biased toward the Changping area, and several different cities can be added in the follow-up to carry out further research; (3) This study only considers the vegetation feature (NDVI) of GF-2 images, and subsequent studies can add texture features and elevation information to make the classification of urban green space more accurate.

6. Conclusions

In this study, we developed a novel hybrid method (MFFTNet) incorporated vegetation feature (NDVI) for urban green space segmentation based on high-resolution remote sensing image, which tackles the challenges caused by complex urban landscape structure. Our method integrated a multi-scale feature fusion module and transformer network, which can provide a multi-scale urban green space segmentation field of view, thus enhancing the recovery of green space edge information. Experiments conducted on the WHDLD, Greenfield, and Greenfield2 datasets reveal that MFFTNet outperformed in accuracy and generalization. In particular, we found that the incorporation of vegetation feature (NDVI) enabled MFFTNet to identify vegetation boundaries more accurately, thereby improving the accuracy of urban green space segmentation. More vegetation features such as phenology information and Sun/Solar-induced Chlorophyll Fluorescence could be considered to incorporate into MFFTNet in the future.

Funding

This research was supported by the National Natural Science Foundation of China (grant numbers 41975183, 42071377, 41875184, 42201053) and supported by a grant from State Key Laboratory of Resources and Environmental Information System (grant number X).

CRediT authorship contribution statement

Yong Cheng: Conceptualization, Methodology, Supervision, Funding acquisition. **Wei Wang:** Conceptualization, Methodology, Supervision, Investigation, Writing – review & editing. **Zhoupeng Ren:** Conceptualization, Methodology, Investigation, Writing – review & editing, Project administration, Funding acquisition. **Yingfen Zhao:** Data curation, Resources. **Yilan Liao:** Writing – review & editing. **Yong Ge:** Conceptualization, Funding acquisition, Writing – review & editing, Project administration. **Jun Wang:** Supervision, Funding acquisition. **Jiaxin He:** Data curation, Visualization. **Yakang Gu:** Data curation, Visualization. **Yixuan Wang:** Data curation, Visualization. **Wenjie Zhang:** Methodology, Funding acquisition, Writing – review & editing. **Ce Zhang:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions that improved this paper.

References

Ardila, J., Bijker, W., Tolpekin, V., Tolpekin, V., Stein, A., 2012. Context-sensitive extraction of tree crown objects in urban areas using VHR satellite images. Int. J. Appl. Earth Obs. Geoinf. 15, 57–69. https://doi.org/10.1016/j.jag.2011.06.005.

Astell-Burt, T., Hartig, T., Putra, I.G.N.E., Walsan, R., Dendup, T., Feng, X., 2022. Green space and loneliness: A systematic review with theoretical and methodological guidance for future research. Sci. Total Environ. 847, 157521 https://doi.org/10.1016/j.scitotenv.2022.157521.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmenta tion. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Bertram, C., Rehdanz, K., 2015. The role of urban green space for human well-being. Ecol. Econ. 120, 139–152. https://doi.org/10.1016/j.ecolecon.2015.10.013.

Cao, J., Li, Y., Sun, M., Chen, Y., Lischinski, D., Cohen-Or, D., Chen, B., Tu, C., 2022. Do-conv: Depthwise over-parameterized con volutional layer. IEEE Trans. Image Process. 31, 3726–3736. https://doi.org/10.1109/TIP.2022.3175432.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818. https://doi.org/10.48550/arXiv.1802.02611.

Chen, Z., Wang, C., Li, J., Fan, W., Du, J., Zhong, B., 2021. Adaboost-like End-to-End multiple lightweight U-nets for road extraction from optical remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 100, 102341 https://doi.org/10.1016/j.jag.2021.102341.

Gao, S., Cheng, M., Zhao, K., Zhang, X., Yang, M., Torr, P., 2019. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell. 43 (2), 652–662. https://doi.org/10.1109/TPAMI.2019.2938758.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

He, D., Shi, Q., Liu, X., Zhong, Y., Zhang, L., 2022. Generating 2m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network. Int. J. Appl. Earth Obs. Geoinf. 106, 102667 https://doi.org/10.1016/j.jag.2021.102667.

Hills, A.P., Farpour-Lambert, N.J., Byrne, N.M., 2019. Precision medicine and healthy living: the importance of the built environment. Prog. Cardiovasc. Dis. 62 (1), 34–38. https://doi.org/10.1016/j.pcad.2018.12.013.

Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Jiang, C., Ren, H., Ye, X., Zhu, J., Zeng, H., Nan, Y., Sun, M., Ren, X., Huo, H., 2022. Object detection from UAV thermal infrared images and videos using YOLO models. Int. J. Appl. Earth Obs. Geoinf. 112, 102912 https://doi.org/10.1016/j.jag.2022.102912.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010.

Kuai, Y., Wang, B., Wu, Y., Chen, B., Chen, X., Xue, W., 2022. Urban vegetation classification based on multi-scale feature perception network for UAV images. J. Geo Inf. Sci. 24 (5), 962–980.

Li, H., Xiong, P., Fan, H., Sun, J., 2019a. Dfanet: Deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9522–9531.

Li, G., Yun, I., Kim, J., Kim, J., 2019b. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. arXiv preprint arXiv:1907.11357. https://doi.org/10.48550/arXiv.1907.11357.

Li, Y., Li, X., Zhang, Y., Peng, D., Bruzzone, L., 2023. Cost-efficient information extraction from massive remote sensing data: When weakly supervised deep learning meets remote sensing big data. Int. J. Appl. Earth Obs. Geoinf. 120, 103345 https://doi.org/10.1016/j.jag.2023.103345.

Li, X., Wang, W., Hu, X., Yang, J., 2019c. Selective kernel networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 510–519.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986. https://doi.org/10.48550/arXiv. 2201.03545.

Liu, D., Han, L., Han, X., 2016. High Spatial Resolution Remote Sensing Image Classification Based on Deep Learning. Acta Opt. Sin. 36 (4), 298–306.

Liu, Y., Yue, W., 2010. Estimation of urban vegetation fraction by image fusion and spectral unmixing. Acta Ecol. Sin. 1, 93–99.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Ma, N., Zhang, X., Zheng, H.T., Sun, J., 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: Proceedings of the European conference on computer vision (ECCV), pp. 116–131. https://doi.org/10.48550/arXiv.1807.11164.

Mehta, S., Rastegari, M., Shapiro, L., Hajishirzi, H., 2019. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9190–9200.

Men, G., He, G., Wang, G., 2021. Concatenated Residual Attention UNet for Semantic Segmentation of Urban Green Space. Forests 12 (11), 1441. https://doi.org/10.3390/f12111441.

Myeong, S., Nowak, D., Duggin, M., 2006. A temporal analysis of urban forest carbon storage using remote sensing. Remote Sens. Environ. 101 (2), 277–282. https://doi.org/10.1016/j.rse.2005.12.001.

Neyns, R., Canters, F., 2022. Mapping of urban vegetation with high-resolution remote sensing: A review. Remote Sens. 14 (4), 1031. https://doi.org/10.3390/rs14041031.

Pan, J., Bulat, A., Tan, F., Zhu, X., Dudziak, L., Li, H., Tzimiropoulos, G., Martinez, B., 2022. Edgevits: Competing light-weight cnns on mobile devices with vision transformers. In: European Conference on Computer Vision. Springer Nature Switzerland, Cham, pp. 294–311. https://doi.org/10.1007/978-3-031-20083-0_18.

Portillo-Quintero, C., Sanchez, A., Valbuena, C., Gonzalez, Y., Larreal, J., 2012. Forest cover and deforestation patterns in the Northern Andes (Lake Maracaibo Basin): a synoptic assessment using MODIS and Landsat imagery. Appl. Geogr. 35 (1–2), 152–163. https://doi.org/10.1016/j.apgeog.2012.06.015.

Räsänen, A., Virtanen, T., 2019. Data and resolution requirements in mapping vegetation in spatially heterogeneous landscapes. Remote Sens. Environ. 230, 111207 https://doi.org/10.1016/j.rse.2019.05.026.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv 2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510–4520. https://doi.org/10.48550/arXiv.1801.04381.

Shao, Z., Zhou, W., Deng, X., Zhang, M., Cheng, Q., 2020. Multilabel remote sensing image retrieval based on fully convolutional network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 318–328. https://doi.org/10.1109/JSTARS.2019.2961634.

Spiering, D., Larsen, C., Potts, D., 2020. Modelling vegetation succession in post-industrial ecosystems using vegetation classification in aerial photographs, Buffalo, New York. Landsc. Urban Plan. 198, 103792 https://doi.org/10.1016/j.landurbplan.2020.103792.

Su, Y., Huang, G., Chen, X., Chen, S., Li, Z., 2011. Research progress in the eco-environmental effects of urban green spaces. Acta Ecol. Sin. 31 (23), 7287–7300.

Thompson, C., Roe, J., Aspinall, P., Mitchell, R., Clow, A., Miller, D., 2012. More green space is linked to less stress in deprived communities: Evidence from salivary cortisol patterns. Landsc. Urban Plan. 105 (3), 221–229. https://doi.org/10.1016/j.landurbplan.2011.12.015.

Tu, X., Huang, G., Wu, J., 2019. Review of the relationship between urban greenspace accessibility and human well-being. Acta Ecol. Sin. 39 (02), 421–431.

Tucker, C., Pinzon, J., Brown, M., Slayback, D., Park, E., Mahoney, R., Vermote, E., Saleous, N., 2005. An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. Int. J. Remote Sens. 26 (20), 4485–4498. https://doi.org/10.1080/01431160500168686.

Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G., 2018. Understanding convolution for semantic segmentation. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, pp. 1451–1460. https://doi.org/10.1109/WACV.2018.00163.

Wang, X., Meng, Q., Zhao, S., Li, J., Zhang, L., Chen, X., 2020. Urban green space classification and landscape pattern measurement based on GF-2 image. J. Geo Inf. Sci. 22 (10), 1971–1982.

Wang, X., Meng, Q., Zhang, L., Hu, D., 2021. Evaluation of urban green space in terms of thermal environmental benefits using geographical detector analysis. Int. J. Appl. Earth Obs. Geoinf. 105, 102610 https://doi.org/10.1016/j.jag.2021.102610.

Wu, Y., He, K., 2018. Group normalization. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19. https://doi.org/10.48550/arXiv.1803.08494.

Xu, Z., Zhou, Y., Wang, S., Wang, L., Li, F., Wang, S., Wang, Z., 2020. A novel intelligent classification method for urban green space based on high-resolution remote sensing images. Remote Sens. 12 (22), 3845. https://doi.org/10.3390/rs12223845.

Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3684–3692.

Yin, J., Dong, J., Hamm, N.A.S., Li, Z., Wang, J., Xing, H., Fu, P., 2021. Integrating remote sensing and geospatial big data for urban land use mapping: A review. Int. J. Appl. Earth Obs. Geoinf. 103, 102514 https://doi.org/10.1016/j.jag.2021.102514.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018. Learning a discriminative feature network for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1857–1866. https://doi.org/10.48550/arXiv.1804.09337.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P., 2019. Joint Deep Learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187. https://doi.org/10.1016/j.rse.2018.11.014.

Zhang, Y., Yin, P., Li, X., Niu, Q., Wang, Y., Cao, W., Huang, J., Chen, H., Yao, X., Yu, L., Li, B., 2022. The divergent response of vegetation phenology to urbanization: A case study of Beijing city, China. Sci. Total Environ. 803, 150079 https://doi.org/10.1016/j.scitotenv.2021.150079.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890. https://doi.org/10.48550/arXiv.1612.01105.

Zhu, F., Gao, J., Yang, J., Ye, N., 2022. Neighborhood linear discriminant analysis. Pattern Recogn. 123, 108422 https://doi.org/10.1016/j.patcog.2021.108422.

Zhu, F., Zhang, W., Chen, X., Gao, X., Ye, N., 2023. Large margin distribution multi-class supervised novelty detection. Expert Syst. Appl. 224, 119937 https://doi.org/10.1016/j.eswa.2023.119937.