




# STMNet: Scene Classification-Assisted and Texture Feature-Enhanced Multiscale Network for Large-Scale Urban Informal Settlement Extraction From Remote Sensing Images

Shouhang Du , Jianghe Xing , Shaoyu Wang, Liguang Wei, and Yirui Zhang 

**Abstract**—Automatic urban informal settlement (UIS) extraction based on high-resolution remote sensing image (HRI) is of great significance for urban planning and management. This study proposes a scene classification-assisted and texture feature-enhanced multiscale network (STMNet) for UIS extraction. First, STMNet takes HRI and computed handcrafted texture feature (HTF) as input. Second, it employs a pseudo-siamese network to extract multidimensional deep features from HRI and HTF, respectively. In addition, a feature attention fusion module is constructed to fuse the aforementioned features. Finally, skip connection and feature decoder are utilized to obtain UIS extraction results. In detail, considering the sparse and dispersed distribution of UIS, a scene information aggregation and classification module is constructed to determine whether the input image patch contains UIS. For the characteristics of high spatial heterogeneity and various shapes and scales of UIS, an improved atrous spatial pyramid pooling is presented to extract multiscale and multireceptive field features. An edge loss function is applied during network training to minimize errors in the edge regions of UIS. The effectiveness of STMNet is tested on a self-produced UIS extraction dataset and the publicly available UIS-Shenzhen dataset. Quantitative results demonstrate that STMNet achieved the best performance in terms of  $f_a$ ,  $F1$ , and  $IoU$ . The  $mr$  is slightly higher than that of MResU-Net and UisNet. In addition, STMNet achieved the best visual interpretation results and the fastest inference speed on the self-produced UIS extraction dataset.

**Index Terms**—Handcrafted texture feature (HTF), high-resolution remote sensing image (HRI), scene classification, semantic segmentation, urban informal settlement (UIS).

## I. INTRODUCTION

URBAN informal settlements (UISs) refer to concentrations of housing in urban areas that lack proper planning, exhibit

Manuscript received 14 May 2024; revised 26 June 2024; accepted 15 July 2024. Date of publication 23 July 2024; date of current version 5 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 42201512 and in part by China Postdoctoral Science Foundation under Grant 2023T160691. (Corresponding author: Jianghe Xing.)

Shouhang Du, Jianghe Xing, Shaoyu Wang, and Yirui Zhang are with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing 100083, China (e-mail: dush@cumt.edu.cn; bqt2200204046@student.cumt.edu.cn; wsy531735430@gmail.com; zhongluck@163.com).

Liguang Wei is with the PIESAT Information Technology Company Limited, Beijing 100195, China (e-mail: weiliguang@piesat.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3432200

poor living conditions, and are characterized by high population density [1], [2], [3]. On one hand, UIS plays a positive role by providing housing for low-income urban residents. However, on the other hand, the existence of UIS has also given rise to significant health and safety hazards, impeding the expansion and modernization of urban functions [1], [4], [5]. Therefore, accurately mapping the spatial extent of UIS is crucial in guiding the relevant authorities to adopt appropriate policies and plans to adjust urban planning.

Before the widespread application of remote sensing, field investigation is a common way for obtaining the spatial extent of UIS [4]. With the continuous improvement in the spatial and temporal resolution of images, remote sensing has become a crucial tool for urban monitoring and planning [6]. Previous studies have largely focused on using traditional machine learning algorithms to extract UIS from high-resolution remote sensing image (HRI). For instance, Duque et al. [7] evaluated the ability of three traditional machine learning algorithms including logistic regression, support vector machine, and random forest to extract UIS in Buenos Aires (Argentina), Medellin (Colombia) and Recife (Brazil), and the results showed that support vector machine with radial basis kernel delivers the best performance. Gevaert et al. [8] utilized the support vector machine with a radial basis kernel to extract UIS in Kigali (Rwanda), and Maldonado (Uruguay). Matarira et al. [9] employed the random forest on the Google Earth engine platform to extract UIS in Durban (South Africa). Although the aforementioned studies have achieved satisfactory results in various study areas, traditional machine learning algorithms still rely on manually designed and selected features, the process is time-consuming and limits the generalizability of the models. Moreover, as the spatial resolution of images continue to improve, the structural and textural information of ground objects becomes more detailed. Traditional machine learning algorithms face challenges in effectively analyzing and utilizing this complex spatial structural context [10], [11]. This poses a disadvantageous impact on the accuracy and generalization capability of UIS extraction.

In recent years, deep learning, particularly convolutional neural networks (CNNs) and fully convolutional networks (FCNs), has made remarkable progress in the field of remote sensing image processing, achieving higher accuracy compared to

traditional machine learning algorithms [11], [12], [13], [14]. Many cutting-edge issues have been alleviated. For instance, a spatial-spectral feature extraction network with a patch attention module was proposed to address the efficient extraction of local and global features in the field of hyperspectral image classification [15]. In addition, a W-shaped hierarchical network was introduced for change detection [16]. To tackle the problem of limited labeled samples in transfer learning, a unified multiscale learning framework was proposed in hyperspectral image classification [17]. To address the high cost of sample annotation, a novel semi-supervised image semantic segmentation network was introduced in the field of land use classification [18]. Moreover, in the domain of scene classification, where dense distribution of objects often results in severe feature mixing, a location-aware multicode generator network was introduced and applied [19].

Currently, in the direction of UIS extraction, there are also a few applications of deep learning. For instance, Verma et al. [20] applied CNN to extract UIS from high and medium-resolution images, demonstrating promising results. More related studies involve the use of FCN for UIS extraction. For instance, Persello and Stein [21] constructed an FCN with dilated convolutions for UIS extraction, which showed the best performance on the self-made dataset. Pan et al. [4] proposed a UIS extraction paradigm based on the U-Net deep learning architecture, demonstrating its superiority through comparison with random forest and object-based image analysis. Fan et al. [11] constructed UisNet, which can utilize multimodal data to extract UIS. Experiments conducted in Shenzhen demonstrate its superior performance. Wei et al. [12] tested the performance of three methods, FCN, UNet, and ResUNet, in extracting UIS from HRI. The results demonstrate that ResUNet is superior to FCN and UNet as it effectively avoids fragmentation and overfitting of UIS.

While the aforementioned studies have made remarkable progress, several challenges still remain. First, unlike widely distributed ground objects such as buildings and roads, the construction of UIS lacks planning, resulting in overall sparse and scattered distribution. The strategy of processing image patches one by one leads to a large amount of redundant computation, which diminishes accuracy and efficiency, as well as may lead to unnecessary false extraction. Second, the high internal spatial heterogeneity of UIS, along with their crowded living spaces and chaotic distribution, complicates the extraction process. Furthermore, most current studies rely solely on remote sensing images for UIS extraction, the strategy limits the extraction capabilities. Although building floors and area data have been applied for UIS extraction, this approach introduces significant data acquisition challenges [11].

To address the aforementioned issues, this study proposes a scene classification-assisted and texture feature-enhanced multiscale network (STMNet) for extracting UIS. The network tackles these issues from three main perspectives.

- 1) *Scene Classification Assisted UIS Extraction*: Both scene classification and ground object extraction are hot topics in the remote sensing community. The former targets identifying the scene category represented by the entire image [19], whereas the latter focuses on the

detailed categorization of each pixel within an image [22]. Particularly, scene classification methods are also commonly used in object-level extraction tasks [23]. The proposed STMNet effectively integrates scene classification with UIS extraction by first determining the presence of UIS in the input image and subsequently deciding whether to proceed with UIS extraction. This approach significantly reduces unnecessary computations due to sparse UIS distribution, decreases errors, and improves extraction efficiency.

- 2) *Multiscale and Multireceptive Field Feature Extraction*: Due to a lack of planning for UIS construction, these areas exhibit high internal spatial heterogeneity, chaotic distribution, and varied shapes and scales. To address these challenges, the proposed STMNet incorporates a multiscale and multireceptive field feature extraction strategy. The multiscale feature helps address challenges arising from the inconsistent shapes and scales of UIS. The multireceptive field configuration is designed to capture multiresolution features. Images at different spatial resolutions present varied representations of ground objects. While high-resolution images capture detailed aspects of UIS, they also magnify internal inconsistencies. Conversely, lower-resolution images can diminish these negative effects, effectively counteracting the complex and heterogeneous scene characteristics associated with UIS.
- 3) *Texture Features Enhance UIS Information*: Handcrafted features are derived directly through mathematical statistical analysis of images. Compared to images rich in spatial information, handcrafted features can specifically highlight regions of interest [24]. Therefore, we not only input images into STMNet but also include handcrafted texture features (HTFs) meticulously designed for UIS characteristics. The incorporation of these texture features significantly enhances the expressive capacity of the scene and improves the accuracy of UIS extraction.

In summary, the proposed STMNet takes HRI and HTF as inputs and utilizes the pseudo-siamese backbone network to extract multidimensional deep features respectively. A feature attention fusion module (FAFM) is constructed to fuse the above-mentioned features. A scene information aggregation and classification module (SIAC) is proposed to introduce scene classification within STMNet. An improved atrous spatial pyramid pooling (ASPP) is introduced to extract multiscale and multireceptive field features. Finally, an additional edge loss function is included to alleviate extraction errors along the edges of UIS. The proposed STMNet is tested on a self-produced dataset for extracting UIS, showing that it effectively improves the extraction of UIS from both visualization and quantitative evaluations.

The innovations of this study can be summarized as follows.

- 1) A novel network named STMNet is proposed for extracting UIS. The network utilizes the FAFM to ingeniously integrate HTF with HRI for high-accuracy UIS extraction.
- 2) This study combines scene classification with ground object extraction tasks through the SIAC, granting STMNet significant advantages in terms of operational speed and reduced false extractions. The constructed improved

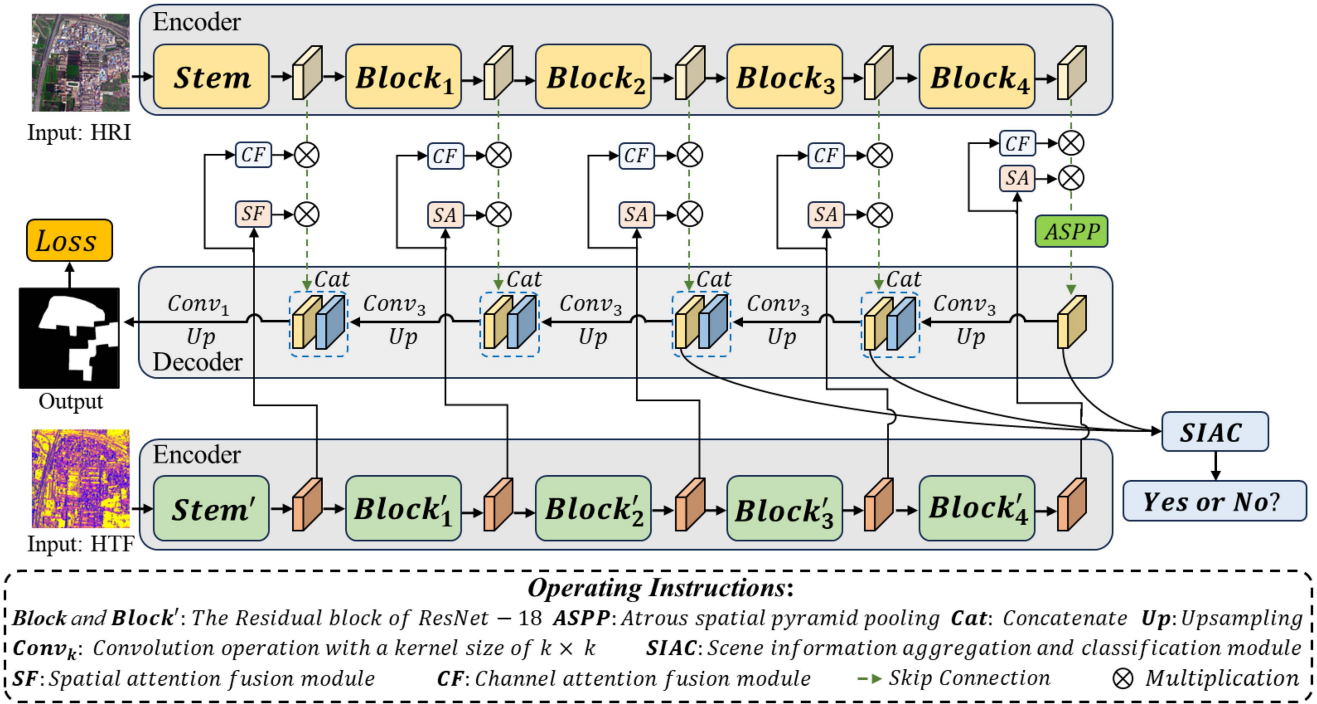


Fig. 1. Illustration of STMNet.

ASPP effectively mitigates the complex and heterogeneous characteristics of UIS. In addition, the use of the edge loss function significantly reduces extraction errors at the edges.

- 3) The UIS extraction dataset is constructed and made public to facilitate the progress of related research. Notably, this dataset is particularly suitable for testing models aimed at extracting sparsely distributed and divergent ground objects in large-scale remote sensing application scenarios. Furthermore, the proposed STMNet achieves optimal performance on this dataset.

## II. METHODOLOGY

### A. Overall Architecture of STMNet

The structure of the proposed STMNet is shown in Fig. 1. It mainly contains four components.

- 1) *Input Data and Feature Encoding*: In this study, both HRI and HTF are input data for STMNet. A pseudo-siamese backbone network is then constructed to extract multidimensional deep features from HRI and HTF, respectively. In addition, an FAFM is designed to fuse the features.
- 2) *SIAC*: Given the sparse and dispersed distribution of UIS, the SIAC is constructed to aggregate the features of the encoder paths and determine whether the input image patch contains UIS or not.
- 3) *Skip Connection and Feature Decoding*: Skip connection is utilized to fuse the features from the encoder path and decoder path. In addition, an improved ASPP is introduced to extract multiscale and multireceptive field features. The features from the decoder path are gradually restored in

size through operations including upsampling, skip connection, concatenation, and convolution, resulting in the extraction results of UIS.

- 4) *Loss Function*: The loss function used in the study consists of the loss for scene classification and the loss for extraction. Moreover, an edge loss function is formulated to impose additional penalties for extraction errors occurring at the UIS edges, with the goal of enhancing the accuracy at these regions.

The whole detailed process of training and testing for STMNet is given in Table I. In the subsequent part sections, we present the key network components.

### B. Feature Extraction Utilizing Pseudo-Siamese Backbone

Before the widespread application of deep learning technologies, the design and utilization of handcrafted features play a key role in ground object extraction tasks [25], [26], [27]. With the evolution of deep learning technologies, an increasing number of CNNs have been proposed [28], [29], [30]. Unlike traditional handcrafted feature extraction, CNNs can automatically extract effective features for tasks like scene classification and ground object extraction [31]. Furthermore, some studies have shown that combining handcrafted features with images as inputs to CNNs can further improve the accuracy of tasks such as scene classification or ground object extraction [32], [33], [34], [35], [36]. This highlights the importance of handcrafted features in complementing and enhancing task performance.

We noticed that UIS exhibits significant texture characteristics, specifically high compactness in distribution and high internal disorder. To strengthen these characteristics, three texture features, energy, homogeneity, and entropy, are calculated in this

TABLE I  
TRAINING AND TESTING PROCEDURE OF THE PROPOSED STMNET

---

**Input:**  
 $imgae_r = HRI$   
 $imgae_n = HTF$   
The initialized or updated learnable parameter:  $\theta$

**Function Encoder**( $imgae_r, imgae_n$ ):  
 $f_{ri} = \mathbf{Block}(imgae_r), i \in (1,2,3,4,5)$   
 $f_{hi} = \mathbf{Block}'(imgae_n), i \in (1,2,3,4,5)$   
 $f_{ei} = \mathbf{FAFM}(f_{ri}, f_{hi}), i \in (1,2,3,4,5)$   
**yes or no** =  $\mathbf{SIAC}(f_{ei}), i \in (3,4,5)$

**If training:**  
 $f_{d5} = \mathbf{ASPP}(f_{e5})$   
 $f_{di} = \mathbf{Conv}_3(\mathbf{cat}(\mathbf{up}(f_{di+1}), f_{ei})), i \in (4,3,2,1)$   
 $loss = \mathbf{Equations} (9, 11, 12, \text{and } 13)$   
 $\hat{\theta} = \mathbf{Equation} (8)$

**If testing:**  
**if yes:**  
 $f_{d5} = \mathbf{ASPP}(f_{e5})$   
 $f_{di} = \mathbf{Conv}_3(\mathbf{cat}(\mathbf{up}(f_{di+1}), f_{ei})), i \in (4,3,2,1)$   
**return**  $\mathbf{Conv}_{1 \times 1}(f_{d1})$   
**else:** pass

---

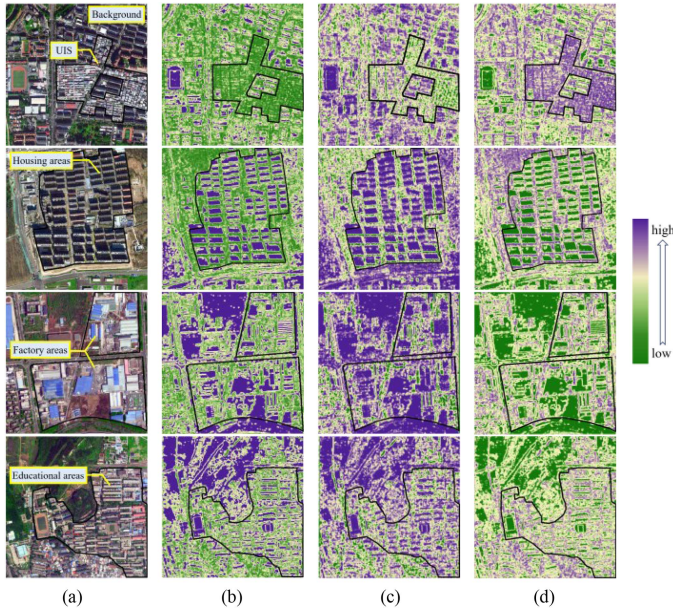


Fig. 2. UIS in gray-scale HRI and texture features. (a) HRI. (b) Energy. (c) Homogeneity. (d) Entropy.

study through the gray-level co-occurrence matrix [37]. Among them, energy reflects the uniformity of gray-scale values in images, similarly, homogeneity reflects the similarity of grayscale values in images. Entropy reflects the complexity of gray-scale values in images. To highlight the distinctive texture features of UIS more prominently, we present a comparison between UIS and backgrounds, as well as other urban functional areas including housing areas, factory areas, and educational areas in Fig. 2. From the perspectives of energy and homogeneity, the values in UIS are low due to the internal disorder of UIS, while entropy exhibits relatively high.

The three aforementioned texture features are concatenated along the channel dimension to form the network input HTF in

this study. To validate the effectiveness of HTF, we visualized the gray-scale histograms of UIS and other regions in the gray-scale HRI, as well as HTFs. As can be seen in Fig. 3, each texture feature contributes to increasing the peak distance in grayscale values between UIS and other regions, thereby enhancing the separability of UIS. For instance, the energy shows that UIS has a higher frequency of occurrences at lower pixel values, while the background tends to have relatively higher pixel values. This contrast is less pronounced in HRI. The distributions of homogeneity and entropy exhibit similar patterns. These differences indicate that the texture features play a significant positive role in extracting UIS.

This study utilizes the pseudo-siamese backbone network to extract deep features from HRI and HTF, respectively. The pseudo-siamese network is similar to the siamese network, but the weights are not shared [24], [38]. Compared to siamese network, pseudo-siamese network can extract more independent and representative deep features [24], [39]. The backbone used in this study is ResNet-18, proposed by Microsoft Research in 2015 [40]. Notably, as shown in Fig. 4, for the UIS extraction task, we have removed the classifier and retained the Stem layer and Residual Blocks. The Stem layer is the initial part of the network that extracts basic features from the input, including a convolution layer, batch normalization, and ReLU activation function. The Residual block, the core of ResNet, contains several convolution layers, batch normalization, and activation functions, as well as residual connections that help address the vanishing gradient problem in deep networks. The pseudo-siamese ResNet-18 configuration can extract five layers of deep features from both HRI and HTF, denoted as  $f_{ri}$  and  $f_{hi}$ , where  $i \in \{1, 2, 3, 4, 5\}$ . The FAFM is constructed in the encoder path to fuse  $f_{ri}$  and  $f_{hi}$ . This module includes a channel attention fusion module (CAFM) and a spatial attention fusion module (SAFM) that enhance  $f_{ri}$  by capturing useful information from  $f_{hi}$  in both channel and spatial dimensions. Spatial attention amplifies responses in key areas, while channel attention optimizes contributions across channels, focusing the model on crucial information. As illustrated in Fig. 5,  $f_{hi}$  is treated as an auxiliary feature, with CAFM and SAFM are utilized to calculate channel attention weight and spatial attention weight, respectively [41]. In the CAFM, two features are obtained through parallel operations of a global average pooling function and a global max pooling function in the spatial dimension. These features are then processed through a multilayer perceptron (MLP), and the resulting features are concatenated. The channel attention weights are subsequently calculated using a sigmoid activation function. For the SAFM, two features are obtained through two parallel operations using the global average pooling and global max pooling function in the channel dimension. The spatial attention weights are then calculated through channel concatenation, a  $1 \times 1$  convolution, and sigmoid activation. Subsequently,  $f_{ri}$  are multiplied by these attention weights to produce enhanced features, labeled as  $f_{ei}$ , where  $i \in \{1, 2, 3, 4, 5\}$ . This process can be defined as follows:

$$w_{ci} = \sigma(\text{MLP}(\text{GAP}(f_{hi})) + \text{MLP}(\text{GMP}(f_{hi}))) \quad (1)$$

$$i \in \{1, 2, 3, 4, 5\}$$

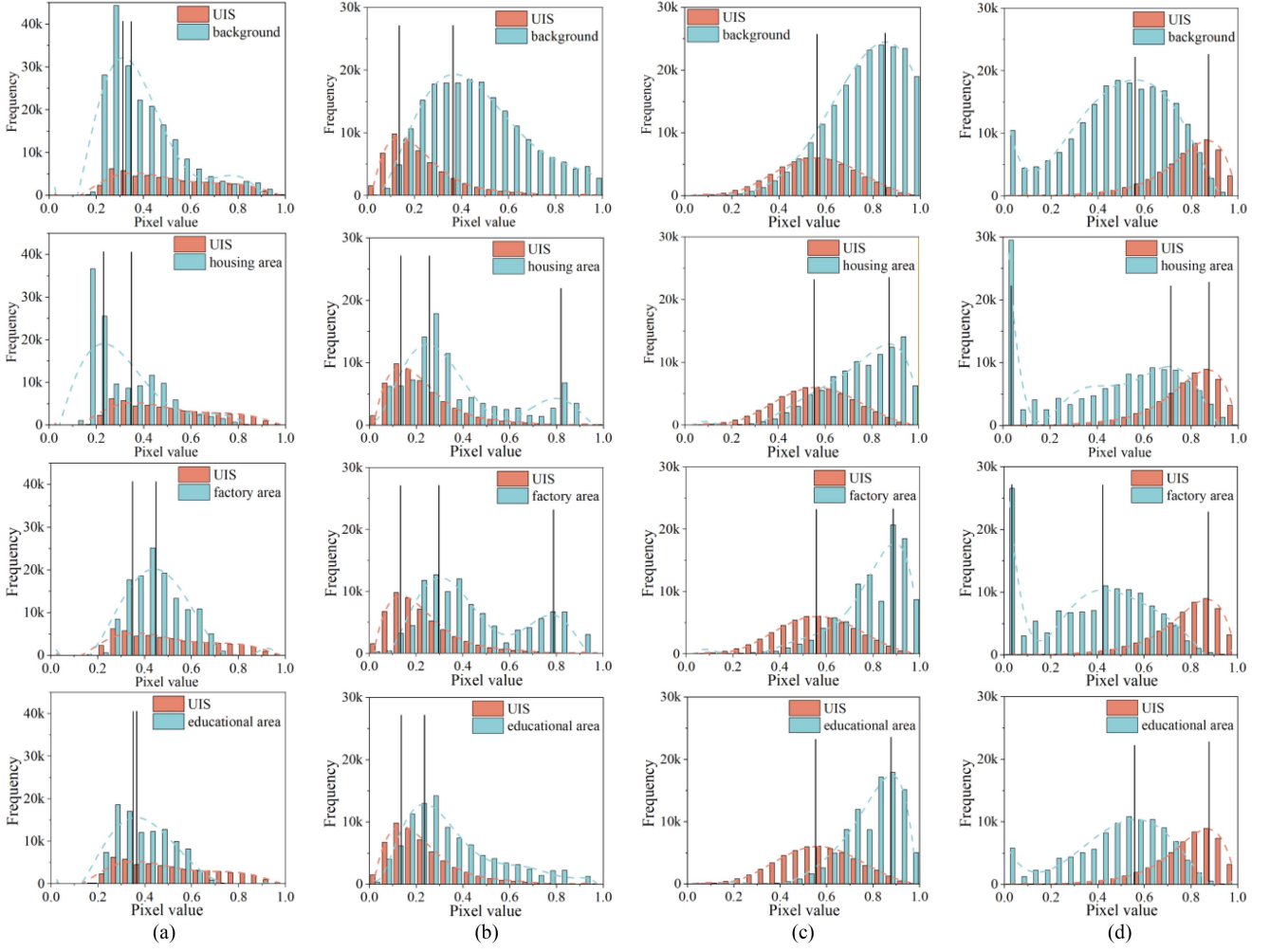


Fig. 3. Gray-scale histograms of UIS and other regions in the gray-scale HRI, as well as HTF. (a) HRI. (b) Energy. (c) Homogeneity. (d) Entropy.

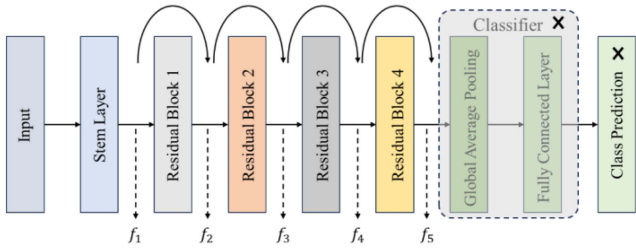


Fig. 4. Modified architecture of ResNet-18 for UIS extraction.

$$w_{si} = \sigma(\text{Conv}_1(\text{Cat}(\text{GAP}(f_{hi}), \text{GMP}(f_{hi})))) \quad (2)$$

$$i \in \{1, 2, 3, 4, 5\}$$

$$f_{ei} = f_{ri} \times w_{ci} \times w_{si}, i \in \{1, 2, 3, 4, 5\} \quad (3)$$

where  $\sigma(\cdot)$  denotes sigmoid function;  $\text{MLP}(\cdot)$  denotes MLP function;  $\text{GAP}(\cdot)$  denotes global average pooling function;  $\text{GMP}(\cdot)$  denotes global max pooling function;  $\text{Cat}(\cdot)$  denotes the feature concatenated in the channel dimension;  $w_c$  is the channel attention weight;  $w_s$  is the spatial attention weight.

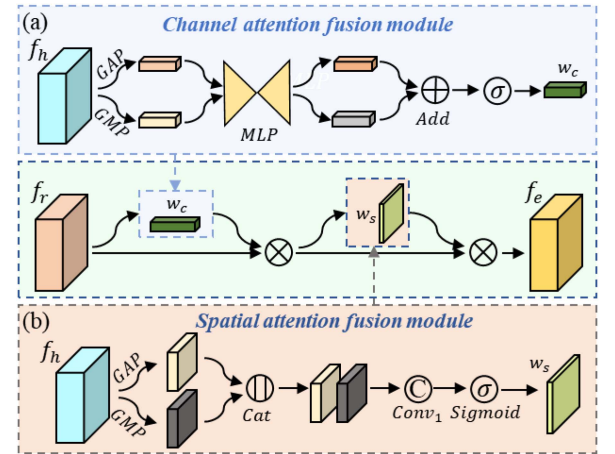


Fig. 5. Illustration of FAFM. (a) CAFM. (b) SAFM.

### C. Scene Information Aggregation and Classification

The task of ground object extraction often requires cropping large images into smaller patches, which are then inputted into the network for extraction, especially when computational

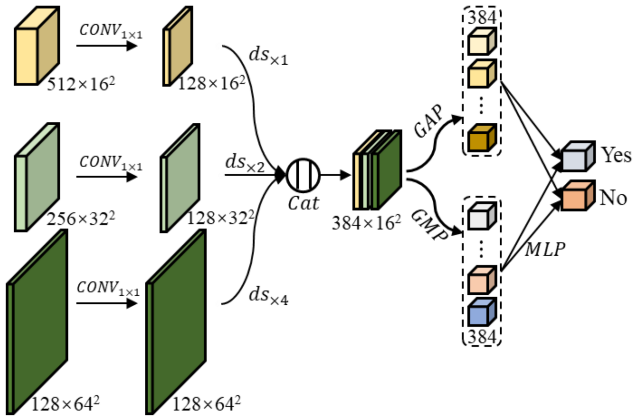


Fig. 6. Illustration of SIAC.

resources are limited [24]. However, when the target ground object occupies a small proportion of the study area, most image patches do not contain the target ground object. This can result in significant redundant computation and extraction errors. Therefore, the SIAC is constructed in this study to perform scene classification, which involves aggregating the features of the encoder paths to determine whether the input image patch contains UIS or not (see Fig. 6). This module utilizes the last three deep features from the encoder path, i.e.,  $f_{e3}$ ,  $f_{e4}$ , and  $f_{e5}$ , to determine the presence of UIS, providing a reference for subsequent computation. Specifically,  $1 \times 1$  convolution and downsampling operations with different strides are employed to process the aforementioned three features to ensure they have the same number of channels and spatial dimensions. Subsequently, these three features are concatenated, and global features are extracted using GAP and GMP operations. Finally, an MLP is utilized to obtain the scene classification result. The computational process can be defined as

$$f' = \text{Cat}(ds_{\times 4}(\text{Conv}_{1 \times 1}(f_{e3})), ds_{\times 2}(\text{Conv}_1(f_{e4})), \text{Conv}_1(f_{e5})) \quad (4)$$

$$\text{result}_{\text{yes or no}} = \text{MLP}(\text{Cat}(\text{GAP}(f'), \text{GMP}(f'))) \quad (5)$$

where  $ds_{\times i}$  represents  $i$  times downsampling operation;  $\text{Conv}_k$  represents the convolution operation with a kernel size of  $k \times k$ ;  $\text{result}_{\text{yes or no}}$  is the scene classification result.

During the network training, UIS extraction operations are executed regardless of whether the scene classification result is “yes” or “no.” This approach is essential as the training of the decoder path requires both positive and negative samples to ensure model generalizability. However, during the network testing, the model parameters are fixed, UIS extraction will not proceed if the scene classification result is “no.” This strategy is particularly effective for UIS that are sparsely and scatteredly distributed, as it significantly reduces unnecessary computational time and decreases the false alarm. Furthermore, the accuracy of scene classification will influence the effectiveness of UIS extraction. Specifically, a false negative, where a

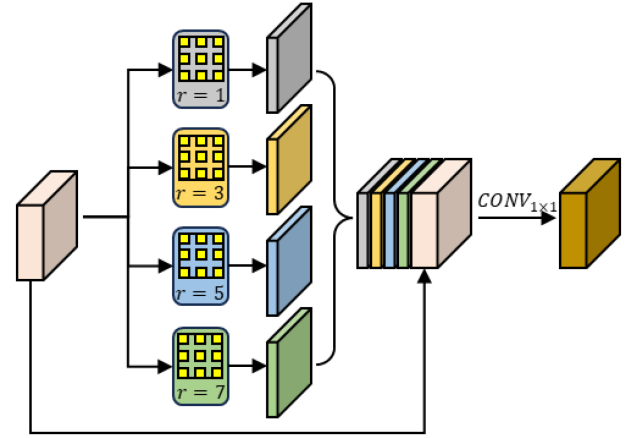


Fig. 7. Illustration of the improved ASPP.

UIS-containing image patch is incorrectly classified as “no,” leads to missed extraction, thereby increasing the  $mr$ . As analyzed earlier, accurate scene classification helps to decrease the  $fa$  and significantly shortens the runtime. Furthermore, the loss of scene classification, serving as an auxiliary loss for STMNet, improves gradient flow and mitigates potential issues of gradient vanishing or exploding during training. It also provides early supervisory signals, helping the network learn more useful features, enhances training efficiency, and reduces the likelihood of overfitting. Consequently, achieving accurate scene classification contributes to the accuracy of UIS extraction.

#### D. Multiscale and Multireceptive Field Features Extraction

Skip connection is an important component in CNNs with an encoder–decoder structure, which serves to fuse the encoder features containing rich spatial information with the decoder features containing rich semantic information. It also helps to alleviate the problem of gradient explosion or vanishing that may occur during the network training [29], [40].

Due to the unplanned construction, UIS exhibits high internal spatial heterogeneity, crowded living spaces, chaotic distribution, and varying shapes and sizes [5]. To further improve the extraction accuracy of UIS, an improved ASPP is proposed in the fifth layer of the skip connection [30]. As shown in Fig. 7, unlike ASPP, we modify the global pooling to identity mapping. The improved ASPP can use convolutions with different dilation rates to extract multiscale features while using different receptive fields to capture multireceptive field features. This module can alleviate the difficulty of UIS extraction caused by the varying shapes and scales of UIS, as well as the spatial heterogeneity within UIS. The improved ASPP can be defined as

$$f_{\text{ASPP}} = \text{Conv}_1(\text{Cat}(f_{\text{ASPP1}}, f_{\text{ASPP2}}, f_{\text{ASPP3}}, f_{\text{ASPP4}}, f_{\text{ASPP5}})) \quad (6)$$

$$f_{\text{ASPP1}} = \text{Conv}_3^1(f_{e5})$$

$$\begin{aligned}
f_{\text{ASPP2}} &= \text{Conv}_3^3 (f_{e5}) \\
f_{\text{ASPP3}} &= \text{Conv}_3^5 (f_{e5}) \\
f_{\text{ASPP4}} &= \text{Conv}_3^7 (f_{e5}) \\
f_{\text{ASPP5}} &= f_{e5}
\end{aligned} \tag{7}$$

where  $\text{Conv}_k^d$  represents the convolution operation with a kernel size of  $k \times k$  and a dilation rate of  $d$ .

### E. Joint Global and Edge Loss Function Supervised Network Training

Back propagation is an optimization technique widely used in optimizing CNNs [31]. The back propagation involves calculating the error gradients for each network layer based on the computed loss and updating the parameters of the network layers using the gradient descent algorithm [42]. The expression for parameter updating can be defined as

$$\hat{\theta} = \theta - lr \times (\partial \text{Loss} / \partial \theta) \tag{8}$$

where  $\hat{\theta}$  is the updated weight;  $\theta$  is the weight before the update;  $lr$  is the learning rate;  $\partial \text{Loss} / \partial \theta$  represents the error gradient of the loss function.

The calculation of the loss is key to the effective operation of the back propagation algorithm. Overall, the loss in this study contains two components: the loss of UIS scene classification and the loss of UIS extraction. In detail, an edge loss function is implemented in the extraction task to improve the accuracy of edge regions of UIS, by adding a penalty to the extraction error in those regions. The loss function used in this study can be defined as

$$\begin{aligned}
\text{Loss} &= \text{Loss}_{\text{scene}} + \text{Loss}_{\text{extraction}} \\
&= \text{Loss}_{\text{scene}} + \text{Loss}_{\text{global}} \\
&\quad + \text{Loss}_{\text{edge}}
\end{aligned} \tag{9}$$

where  $\text{Loss}_{\text{scene}}$  denotes the loss of UIS scene classification;  $\text{Loss}_{\text{extraction}}$  denotes the loss of UIS extraction;  $\text{Loss}_{\text{global}}$  denotes the global loss function;  $\text{Loss}_{\text{edge}}$  denotes the edge loss function, and  $\text{Loss}_{\text{extraction}} = \text{Loss}_{\text{global}} + \text{Loss}_{\text{edge}}$ .

In this study, the binary cross-entropy loss function is employed to measure the differences between predictions and ground truth in both scene classification and extraction tasks [43], [44]. The binary cross-entropy loss function can be defined as follows:

$$\begin{aligned}
\text{CELoss} &= \frac{1}{N} \sum - [y_i \times \log(p_i) + (1 - y_i) \\
&\quad \times \log(1 - p_i)]
\end{aligned} \tag{10}$$

where  $N$  denotes the total number of samples; If the sample is positive,  $y_i = 1$ ; otherwise,  $y_i = 0$ .  $p_i$  denotes the result predicted by the method. Therefore,  $\text{Loss}_{\text{scene}}$  can be defined as

$$\begin{aligned}
\text{Loss}_{\text{scene}} &= \frac{1}{N^s} \sum - [y_i^s \times \log(p_i^s) + (1 - y_i^s) \\
&\quad \times \log(1 - p_i^s)]
\end{aligned} \tag{11}$$

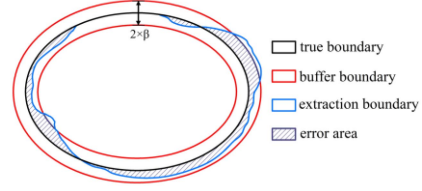


Fig. 8. Illustration of extraction errors within the buffer area.

where  $N^s$  represents the number of images. If the input image contains UIS,  $y_i^s = 1$ ; otherwise,  $y_i^s = 0$ .  $p_i^s$  denotes the result predicted by the SIAC. And  $\text{Loss}_{\text{global}}$  can be defined as

$$\begin{aligned}
\text{Loss}_{\text{global}} &= \frac{1}{N^g} \sum - [y_i^g \times \log(p_i^g) + (1 - y_i^g) \\
&\quad \times \log(1 - p_i^g)]
\end{aligned} \tag{12}$$

where  $N^g$  represents the number of pixels in the image. If the prediction result for a pixel is UIS,  $y_i^g = 1$ ; otherwise,  $y_i^g = 0$ .  $p_i^g$  denotes the result predicted by the STMNet.

As illustrated in Fig. 8, the calculation of  $\text{Loss}_{\text{edge}}$  first involves using the sobel operator to compute and identify the edges of UIS. Then, a buffer area is constructed both internally and externally around each pixel  $\beta$ , and additional penalties are applied to the extraction errors within this buffer area. The relationship between the edge loss function and the global loss function can be defined as

$$\text{Loss}_{\text{edge}} = \text{buffer}(\text{sobel}(\text{mask})) \times \text{Loss}_{\text{global}} \tag{13}$$

where  $\text{mask}$  represents the binary mask of UIS;  $\text{buffer}(\cdot)$  denotes the function for building buffer area.

## III. EXPERIMENTS

### A. Study Area and Dataset

The capital of China, Beijing, is selected as the study area. Beijing faces challenges like uneven urbanization of social spaces due to its rapid urban expansion. This problem is reflected in the fact that the city contains a large number of modern commercial, residential, and educational areas, as well as many UIS to be developed [45]. Consequently, the extraction and analysis of UIS hold significant reference value for subsequent urban development and planning research.

The HRI used in this study was downloaded from the AMap platform, with a resolution of 2.38 m. The image was captured by satellites from DigitalGlobe's WorldView series and SpaceView's Gaofen satellite series in the year 2022. The image includes three bands: red, green, and blue. Considering that UIS is primarily located within the physical urban areas of Beijing, this study specifically concentrates on the physical urban areas of Beijing (the physical urban areas of Beijing can be found online).<sup>1</sup> The location and image coverage of the study area are shown in Fig. 9. We randomly delineate 30 rectangular boxes with 5000 m sides evenly spaced across the study area. The boxes

<sup>1</sup>[Online]. Available at: [http://geoscape.pku.edu.cn/otherdata\\_en.html](http://geoscape.pku.edu.cn/otherdata_en.html).

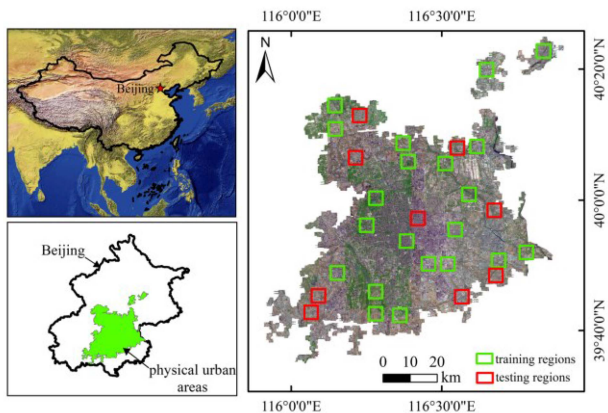


Fig. 9. Schematic diagram of the study area.

are randomly divided into training and testing areas in a 7:3 ratio, assigning 21 boxes for training and 9 boxes for testing. These processes do not involve any manual intervention. The area of the study area is 3341.67 km<sup>2</sup>, with 525 km<sup>2</sup> used for training and 225 km<sup>2</sup> for testing. The key to the accuracy of the mask lies in the correctness of the UIS boundaries. Due to the boundaries between UIS and other regions not being distinctly visible in HRI, we adopted a strategy of field investigation, visual interpretation, and cross-validation to construct the dataset. Specifically, during the field investigation, we accurately recorded the boundaries between UIS and other regions. These boundaries are usually narrow and winding paths, which are difficult to distinguish in HRI. Subsequently, we conducted visual interpretation using the data from the surveys and HRI. Finally, different researchers conducted on-site verification of the interpreted boundaries. Due to limitations in computer processing performance, images need to be cropped into patches before being fed into the network. During this operation, we did not impose any regulations. Although this may result in incomplete UIS structures and the loss of global UIS information, it is consistent with practical applications, where the distribution of UIS in the test regions is unknown. In the study, an overlapping cropping strategy with a 256-pixel overlap is applied when cropping the images and masks. This method is widely used in large-scale ground object extraction to minimize the negative influence of incomplete structures. The images and masks are cropped to patches of 512×512 pixels. The data from the training regions is then divided into the training dataset and validation dataset in the ratio of 7:3. After cropping, the training dataset contains 720 pairs of samples, the validation dataset contains 309 pairs, and the test dataset contains 441 pairs. To enhance the generalization of the method, data augmentation is performed before training by conducting various operations on image patches, including horizontal flipping, vertical flipping, and diagonal flipping [46].

The dataset constructed in this study is publicly available. Compared to other public UIS extraction datasets such as UIS-Shenzhen [11] and Xi'an dataset [47], our dataset is distinctive in that it is randomly selected and retains negative samples that do not contain UIS. This strategy is commonly employed in datasets focused on large-scale applications, such as the

WHU-Building dataset (After cropping into 256 × 256 image patches, the proportion of negative samples is 28.09%) [48]. Retaining negative samples allows the model to encounter more nontarget scenes during training, which aids in learning to distinguish between targets and nontargets, alleviates overfitting issues, and enhances the model's robustness and generalization capabilities in unknown application scenarios. Furthermore, since STMNet includes the SIAC, removing negative samples would render the SIAC training infeasible.

The dataset constructed in this study is publicly available and provides a more accurate reflection of real-world applications compared to other public UIS extraction datasets, such as UIS-Shenzhen and Xi'an datasets. Specifically, our dataset involves random selection for training and testing sets, whereas other datasets involve human intervention, retaining only image blocks that contain UIS and discarding those without, which distorts the dataset away from actual application scenarios.

### B. Experimental Configuration

The experiments are implemented using a desktop computer with Intel Xeon Gold 5118, 32GB memory, and NVIDIA GeForce RTX2080Ti. All algorithms are performed using Python language (v3.6.5) on the PyCharm platform (Community Edition 2020.2.1 × 64). The deep learning framework is Pytorch (v1.7.0+cu110).

The method of manual parameter tuning is employed to search for a suitable learning rate and  $\beta$ , which are set to 0.0005 and 10, respectively. Due to the limitations of computer memory and data volume, the batch size is set at 4. The Adam function is selected as the parameter optimizer since it is robust and well-suited to a wide range of non-convex optimization problems in deep learning. To train the model until the loss curve tends to be smooth, the maximum number of training iteration epochs is set to 100.

The proposed method is compared with several advanced and classical methods, including two transformer-structured networks, namely Shunted Dual Skip Connection UNet (SDSC-UNet) [49], UNet-like transformer (abbreviated as UNetFormer) [50], and UIS semantic segmentation network (abbreviated as UisNet) [11]; two CNN-structured networks, namely multiattention network (abbreviated as MANet) [51] and multistage attention ResU-Net (abbreviated as MAREsU-Net) [52], and two classical CNNs, namely DeepLabv3+ [30], and pyramid scene parsing network (abbreviated as PSPNet) [53]. Notably, since UisNet requires multimodal data inputs, the comparative experiment with UisNet is not conducted on the self-produced Beijing UIS extraction dataset. Instead, it is carried out on the publicly available UIS-Shenzhen dataset.

The performance of STMN is evaluated from both qualitative and quantitative aspects. Qualitative evaluation is to visualize the extraction results to visually compare the performance of different methods. Quantitative evaluation is conducted from three perspectives: UIS extraction, UIS scene classification, and method efficiency. Four evaluation metrics are adopted to quantitatively evaluate the performance of extraction, including falsealarm (fa) :  $FP/(TP + FP) \times 100\%$ , missratemr :  $FN/$



TABLE II  
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT METHODS

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
STMNet	<b>10.60</b>	12.17	<b>88.61</b>	<b>79.55</b>
SDSC-UNet	12.65	13.72	86.81	76.70
UNetFormer	12.45	12.55	87.50	77.78
MAResU-Net	14.07	<b>11.49</b>	87.20	77.31
MANet	12.25	12.95	87.40	77.62
DeepLabv3+	14.60	18.31	83.50	71.68
PSPNet	16.52	16.52	83.48	71.64

The best results are highlighted in bold.

$(TP + FN) \times 100\%$ ,  $F1 - \text{score} (F1) : 2 \times TP / (2 \times TP + FP + FN) \times 100\%$ , and Intersection over Union (IoU) :  $TP / (TP + FP + FN) \times 100\%$ , where TP, FP, and FN represent the true positive, false positive, and false negative, respectively, in the confusion matrix for the extraction task. Three evaluation metrics are adopted to quantitatively evaluate the performance of scene classification, including falsealarm/(fa):  $FP / (TP + FP) \times 100\%$ , missrate/(mr):  $FN / (TP + FN) \times 100\%$ ,  $F1 - \text{score} (F1) : 2 \times TP / (2 \times TP + FP + FN) \times 100\%$ , where TP, FP, and FN represent the true positive, false positive, and false negative, respectively, in the confusion matrix for the scene classification task. These metrics directly reflect the effectiveness of UIS scene classification and extraction, avoiding the interference of the background. Furthermore, metrics including params size (measured in MB), floating point operations (FLOPs, measured in Giga), and the inference time (measured in seconds for the physical urban areas of Beijing) are calculated to evaluate the complexity and efficiency [54]. Here, params size reflects the model's storage requirements, when the network parameters are stored as float32 type,  $\text{params size} = \text{parameters} \times 4 \text{bytes}$ . FLOPs measure the computational complexity, and the inference time indicates the model's efficiency in practical applications. To ensure reliable results and minimize random errors, all experiments in this study are repeated nine times and the results are averaged.

### C. Comparison of Different Methods

The quantitative evaluation results are shown in Table II. It is evident that STMNet achieves the best performance in the metrics of  $fa$ ,  $F1$  and IoU. Specifically, compared to the suboptimal method (MANet), STMNet reduces the  $fa$  by 1.65 percentage points, and compared to the worst method (PSPNet), STMNet reduces the  $fa$  by 5.92 percentage points. In terms of the  $F1$ , STMNet outperforms the suboptimal method (UNetFormer) by 1.11 percentage points, and outperforms the worst method (PSPNet) by 5.13 percentage points. In terms of IoU, STMNet outperforms the suboptimal method (UNetFormer) by 1.77 percentage points and outperforms the worst method (DeepLabv3+) by 7.87 percentage points. Although STMNet does not achieve the lowest  $mr$ , it closely follows the optimal method (MAResU-Net), with only a difference of 0.68 percentage points. In summary, STMNet exhibits a significant

advantage in the task of extracting UIS. It is worth noting that the two classical CNNs (DeepLabv3+ and PSPNet) perform poorly.

The Quantitative evaluation results of the nine rectangular boxes are shown in Fig. 10. Taking  $fa$  as an example, STMNet achieved the optimal results in four of the nine rectangular boxes (1, 6, 7, and 9). In the other five rectangular boxes, different methods achieved optimal results; SDSC-UNet achieves optimal results in rectangle box 4, MAResU-Net achieves optimal results in rectangle box 8, DeepLabv3+ achieves optimal results in rectangle boxes 3 and 5, and PSPNet achieves optimal results in rectangle box 2. Regarding other evaluation metrics, STMNet has the lowest  $mr$  in two of the nine rectangular boxes (2 and 4). In terms of  $F1$  and IoU, STMNet achieves optimal results in five of the nine rectangular boxes (1, 2, 4, 6, and 9).

Fig. 11 illustrates the extraction results of UIS in some image patches. It is evident that all methods can extract UIS from HRI. However, it can be seen from some details that there is a significant difference in the extraction ability of each method. STMNet has the best visualization, characterized by the least number of false and missed extracted pixels, as well as extracted boundaries that closely match the true boundaries (e.g., regions 1, 2, 3, 4, 5, 6, and 8). Although the proposed STMNet in some regions such as regions 7 and 9 has significant missed extractions, it is also surpassed only by a few methods such as UNetFormer. We have also counted the number of pixels of TP, FN, FP, and TN in each image patch as a percentage of the total number of pixels in Fig. 12. It can be seen in Fig. 12(a)–(e) that only the fifth image patch where the sum of the percentage of FN and FP is 4.4% is not the least. It can also be seen in Fig. 12(f) that the percentage of FN and FP is the least among all the five image patches, proving that the area of false and missed extracted pixels is minimized.

Fig. 13 illustrates the UIS extraction results in two rectangular boxes. In the first example, despite the obvious false and missed extractions in MANet's results, this is the only one of the seven methods that extracts all UIS. It is evident that other methods have significant missed extractions, such as UNetFormer, MAResU-Net, DeepLabv3+, and PSPNet, which failed to extract the UIS in Region 1. SDSC-UNet does not completely extract the UIS in Region 4. The proposed STMNet also failed to extract the UIS in Region 6, due to an incorrect judgment by the SIAC about the presence of UIS in that image patch. However, it can be seen from many details that the proposed STMNet has the least number of false and missed extracted pixels, such as in regions 2 and 7. In the second example, all methods successfully extract all UIS. In Region 8, MANet has the fewest miss-extraction pixels, while UNetFormer has the most. In Region 11, the proposed STMNet has the fewest false-extraction pixels. Additionally, it is noteworthy that due to the use of the SIAC, STMNet has no false-extraction pixels in regions 3, 5, 9, 10, and 12. Fig. 14 quantifies the proportions of TP, FP, FN, and TN in the two aforementioned examples. STMNet has the lowest proportion of FP, further confirming the advantage of the SIAC. The combined ratio of FP and FN is also the lowest in both the two examples, indicating that

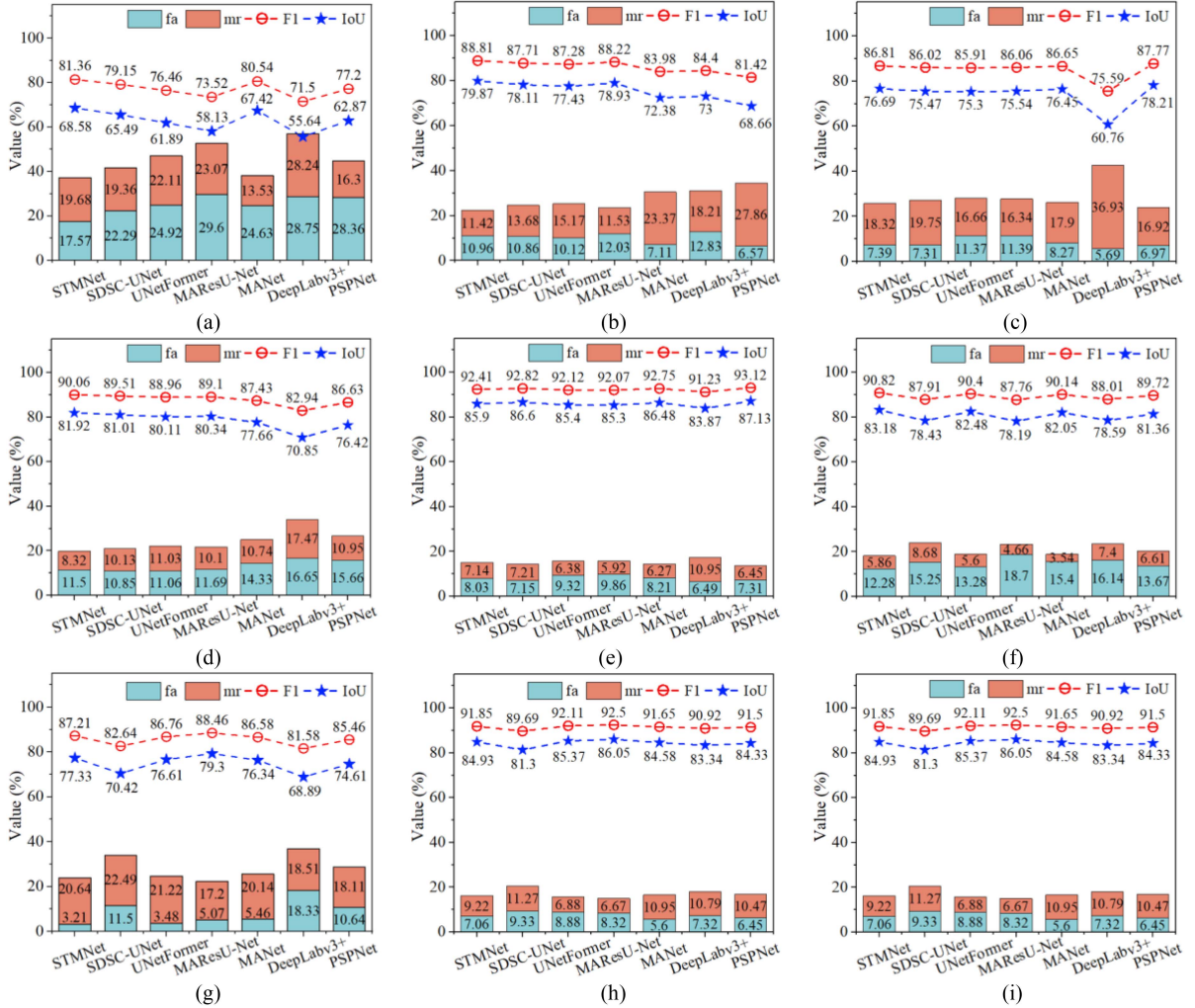


Fig. 10. Quantitative evaluation results of the nine rectangular boxes. (a) Rectangular box 1. (b) Rectangular box 2. (c) Rectangular box 3. (d) Rectangular box 4. (e) Rectangular box 5. (f) Rectangular box 6. (g) Rectangular box 7. (h) Rectangular box 8. (i) Rectangular box 9.

the STMNet has the fewest miss-extraction and false-extraction pixels.

#### IV. DISCUSSION

The aforementioned studies have validated the effectiveness of the proposed STMNet from both visual interpretation and quantitative evaluation. However, some issues still need to be further discussed. For example, the importance and performance analysis of each module, and how well the method generalizes to other datasets. Whether the method has advantages in terms of computational complexity and efficiency? Last but not least, what are the shortcomings of the proposed STMNet and the outlook for future study? These will be discussed in detail in this section.

##### A. Benefits of HTF

In this study, comparative experiments are designed to evaluate whether HTF is beneficial for UIS extraction, as well as to assess whether all three computed texture features are

TABLE III  
QUANTITATIVE EVALUATION RESULTS OF THE HTF EFFECT

Method	Energy	Homogeneity	Entropy	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
STMNet							
w/o HTF	—	—	—	11.97	14.93	86.52	76.25
w/ HTF	✓	—	—	12.69	13.18	87.06	77.08
STMNet							
w/o HTF	—	—	—	11.95	12.42	87.82	78.28
w/ HTF	✓	—	✓	11.69	13.21	87.55	77.85
	✓	✓	✓	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>

The best results are highlighted in bold. w/ means with and w/o means without.

advantageous. The results in Table III confirm that the three texture features, whether used individually or in combination, enhance UIS extraction performance, with the simultaneous use of all three features yielding the optimal performance.

The aforementioned experiment validates the effectiveness of HTF in the UIS extraction task. How to integrate HTF into neural networks is also a widely studied topic. Existing

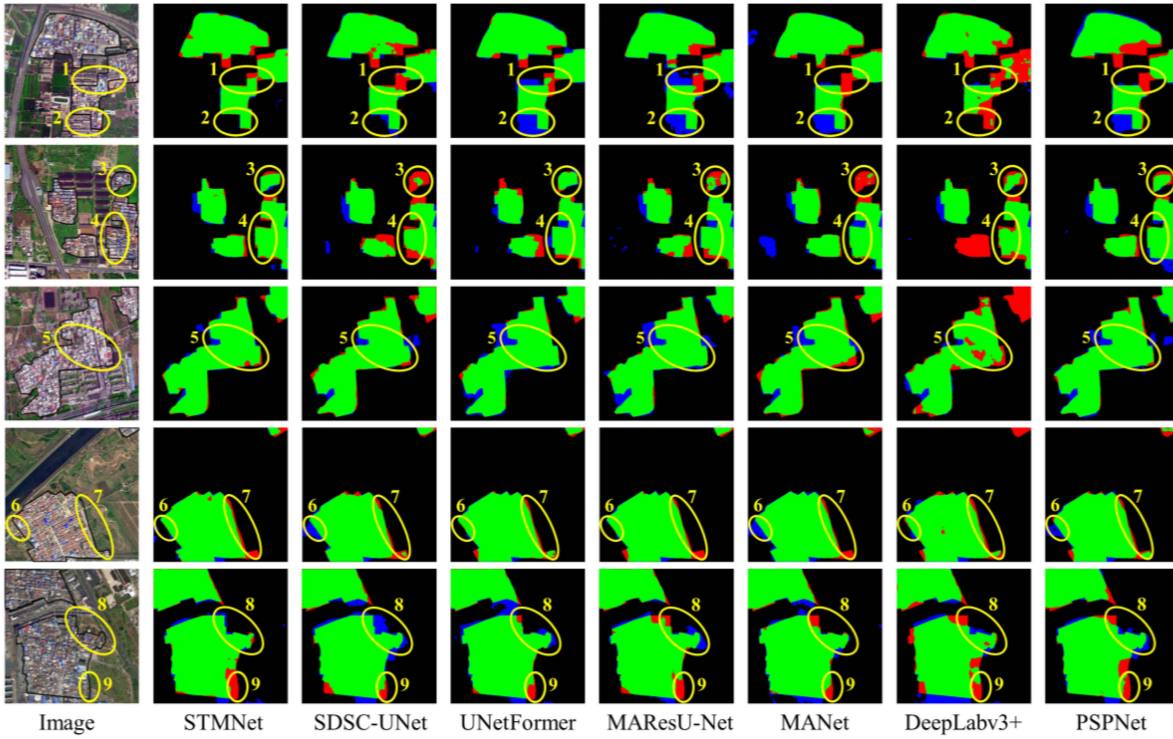


Fig. 11. Extraction results of UIS in some image patches. Different colors represent the correctness or incorrectness of the results, including TP (green), TN (black), FP (blue), and FN (red).

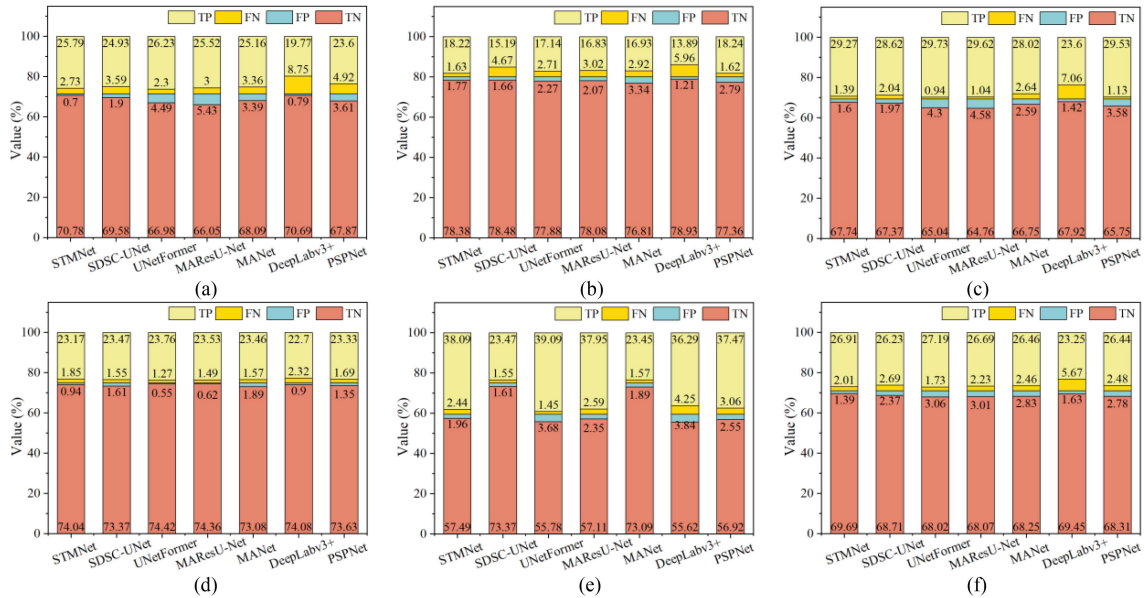


Fig. 12. The percentage of TP, FP, FN, and TN. (a) Image patch 1. (b) Image patch 2. (c) Image patch 3. (d) Image patch 4. (e) Image patch 5. (f) All image patches.

studies are usually conducted with two methods: 1) Early fusion, where HTF are concatenated with HRI and then inputted into a feature extractor to obtain multidimensional deep features [55], [56]. 2) late fusion, which involves directly concatenating HRI with deep features extracted from HRI [34]. In this study, we compared the pseudo-siamese feature extraction method

used in STMNet with the two methods mentioned above. The experimental results, as shown in Table IV, reveal that STMNet achieved the optimal results in terms of fa, mr,  $F1$  and IoU.

In this study, the FAFM is utilized to integrate the deep features extracted from HTF with those extracted from HRI. Other common feature fusion strategies, including addition and

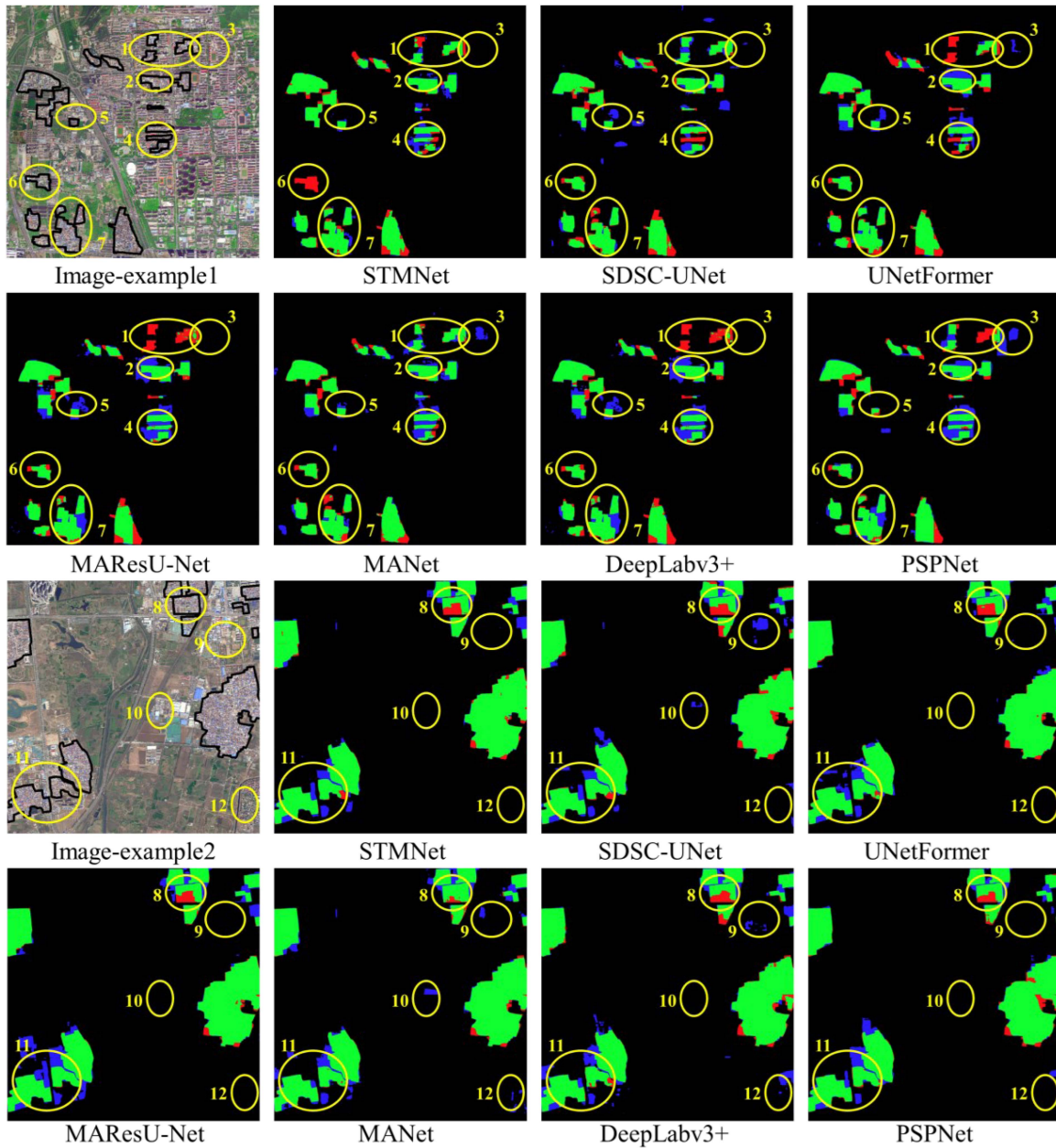


Fig. 13. UIS extraction results in two rectangular boxes. Different colors represent the correctness or incorrectness of the results, including TP (green), TN (black), FP (blue), and FN (red).

TABLE IV  
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT HTF UTILIZATION METHODS

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
early-fusion	11.21	14.91	86.90	76.84
late-fusion	13.06	12.46	87.24	77.37
STMNet	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>

The best results are highlighted in bold.

TABLE V  
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT FEATURE FUSION STRATEGIES

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
Addition	11.62	12.82	88.05	78.65
Concatenation	10.80	12.80	88.19	78.87
FAFM	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>

The best results are highlighted in bold.

concatenation, are also considered. Comparative experiments of these three strategies are conducted. It can be seen from Table V that STMNet with FAFM obtains the optimal  $fa$ ,  $mr$ ,  $F1$ , and  $IoU$ .

### B. Benefits of SIAC

Due to the lack of planning in the construction of UIS, they are usually dispersed within physical urban areas. In this

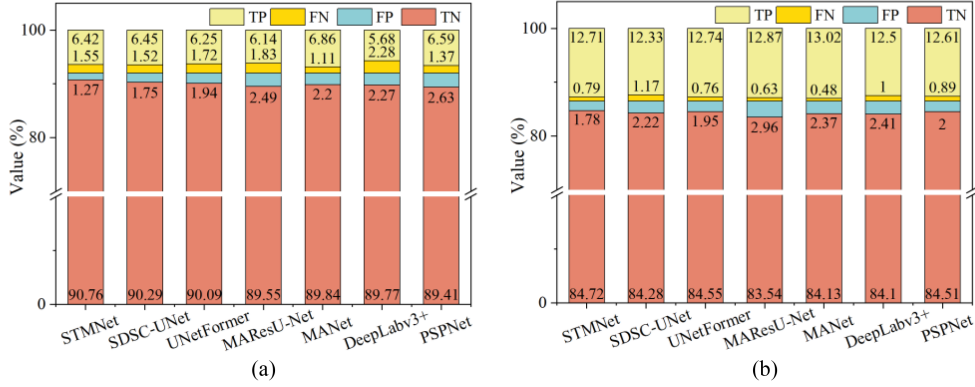


Fig. 14. Percentage of TP, FP, FN and TN. (a) Example 1. (b) Example 2.

TABLE VI  
QUANTITATIVE EVALUATION RESULTS OF SCENE CLASSIFICATION FOR  
DIFFERENT INPUT FEATURES

Features	$fa' \downarrow$	$mr' \downarrow$	$F1 \uparrow$
$\{f_{e5}\}$	2.61	2.30	97.54
$\{f_{e4} \text{ and } f_{e5}\}$	2.28	1.64	98.04
$\{f_{e3}, f_{e4}, \text{ and } f_{e5}\}$	<b>1.30</b>	<b>0.66</b>	<b>99.02</b>
$\{f_{e2}, f_{e3}, f_{e4}, \text{ and } f_{e5}\}$	1.96	1.64	98.20
$\{f_{e1}, f_{e2}, f_{e3}, f_{e4}, \text{ and } f_{e5}\}$	1.32	1.97	98.35

The best results are highlighted in bold.

study, SIAC is introduced to determine whether the image patch contains UIS or not. Lower level features contain rich spatial information, whereas higher level features contain rich semantic information. By fusing these low-level and high-level features, we can enhance the spatial and semantic information of the features used for classification. A comparative experiment on feature selection is conducted to illustrate how the combination of these features can achieve optimal scene classification results. As shown in Table VI, the optimal results are achieved using three features, specifically  $f_{e3}$ ,  $f_{e4}$ , and  $f_{e5}$ . The  $fa'$ ,  $mr'$ , and  $F1$  are 1.30%, 0.66%, and 99.02%, respectively.

In addition, comparative experiments are conducted to assess the effectiveness of SIAC and to investigate the effect of the scene classification threshold ( $\theta$ ) on scene classification and UIS extraction accuracy. The results, presented in Table VII, can be summarized in the following three points.

- 1) *Appropriate  $\theta$  Setting*: It is evident that a lower  $\theta$  leads to higher rates of false classifications and false extractions, whereas a higher  $\theta$  results in greater rates of missed classifications and extractions. When  $\theta$  is set to 0.5, a good balance is achieved. This value is commonly used in scene classification and ground object extraction tasks, and we believe it provides the most accurate balance for a wide range of tasks. In addition, we recommend adjusting  $\theta$  based on experimental performance. For instance, for tasks such as fire detection, we suggest setting a lower  $\theta$  to enhance the model's sensitivity to early fire warnings, as a timely response is far more critical than missing any potential fire spots.

- 2) *Advantages of SIAC*: Although the use of SIAC ( $\theta = 0.5$ ) increases  $mr$  by 1.44 percentage points, it significantly reduces  $fa$  by 3.50 percentage points. In terms of comprehensive accuracy metrics, the  $F1$  and IoU improve by 1.06 and 1.69 percentage points, respectively. Therefore, the use of SIAC offers significant advantages.

- 3) *Computational Complexity and Efficiency*: From this perspective, although SIAC requires an additional 0.82 MB of memory and increases computational demand by 394.00 Mega FLOPs, it significantly reduces the processing time by 329.23 s when analyzing the physical urban areas of Beijing.

Fig. 15 further demonstrates the effectiveness of SIAC. From the first two image patches, it is evident that classification errors in STMNet have led to an increase in missed extracted pixels (regions 1 and 2). However, the correct scene classification of the latter two image patches has prevented false extracted pixels (regions 3 and 4).

### C. Benefits of Improved ASPP

In this study, the improved ASPP is used to extract multiscale and multireceptive field features, and the effectiveness of the improved ASPP is verified in Table VIII and Fig. 16. As shown in Table VIII, the application of ASPP decreases the  $fa$  and  $mr$  by 1.03 and 0.64 percentage points, respectively. It also improves the  $F1$  and IoU, with increases of 0.83 and 1.33. Fig. 16 demonstrates that the extraction of multiscale and multireceptive features using improved ASPP leads to results that are more in line with the ground truth, with fewer false-extraction and miss-extraction. In addition, it noticeably reduces the holes in the extraction results, resulting in a higher compactness.

### D. Benefits of Edge Loss Function

In this study, the edge loss function is utilized to penalize errors at edges, with  $\beta$  controlling the size of the edge buffer zone. The appropriate  $\beta$  is crucial for achieving optimal extraction performance. A larger  $\beta$  may dilute the emphasis on edge loss, thereby reducing the training focus and increasing computational complexity. Conversely, a smaller  $\beta$  may be insufficient to adequately cover the necessary boundary transition areas. The

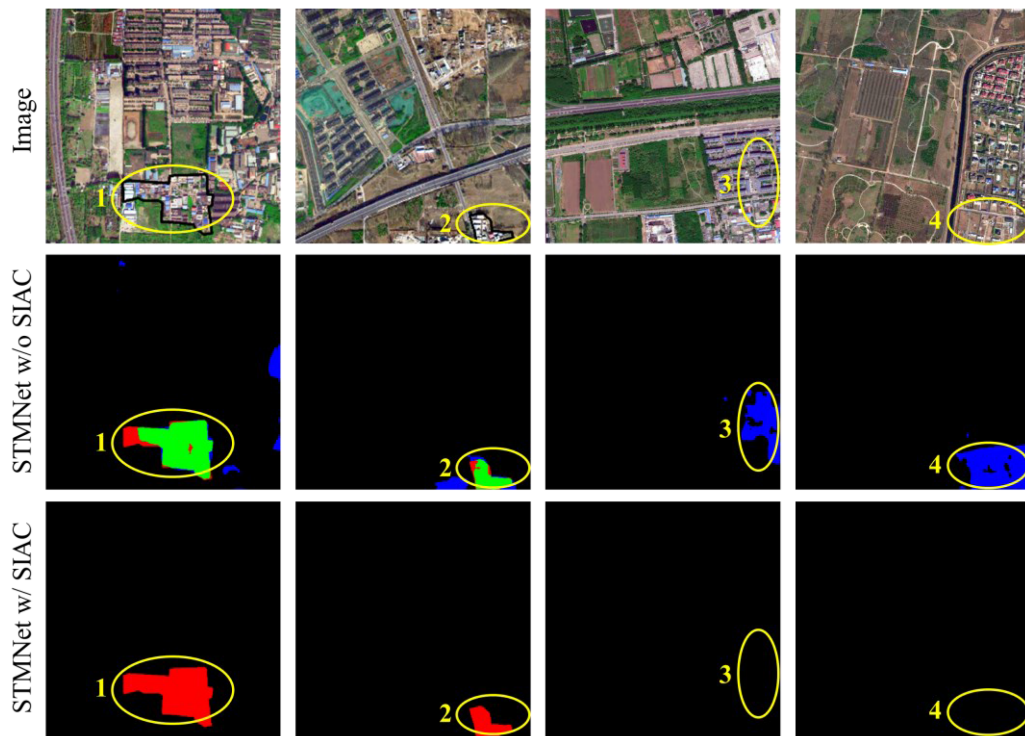


Fig. 15. Impact of SIAC on the UIS extraction results. Different colors represent the correctness or incorrectness of the results, including TP (green), TN (black), FP (blue), and FN (red).

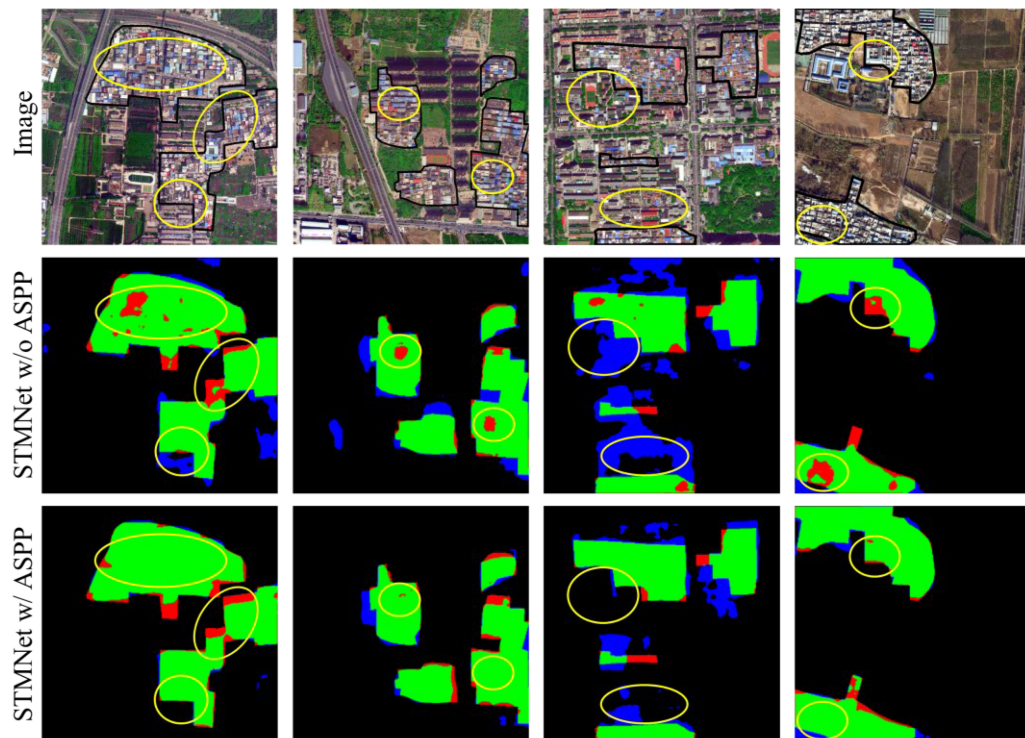


Fig. 16. Impact of improved ASPP on the UIS extraction results. Different colors represent the correctness or incorrectness of the results, including TP (green), TN (black), FP (blue), and FN (red).

TABLE VII  
QUANTITATIVE EVALUATION RESULTS OF THE SIAC EFFECT

Method	$\theta$	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$	Params size	FLOPs	Inference Time
STMNet w/o SIAC	—	—	—	—	14.10	<b>10.73</b>	87.55	77.86	—	—	896.76
	0.3	32.64	<b>0.00</b>	80.50	14.10	<b>10.73</b>	87.55	77.86			
STMNet w/ SIAC	0.4	29.20	<b>0.00</b>	83.26	12.57	<b>10.73</b>	88.34	79.12	0.82	394.00 (Mega)	567.53
	0.5	1.30	0.66	<b>99.02</b>	10.60	12.17	<b>88.61</b>	<b>79.55</b>			
	0.6	<b>0.34</b>	2.62	98.51	<b>10.18</b>	13.73	88.01	78.59			

The best results are highlighted in bold. w/ means with and w/o means without.

TABLE VIII  
QUANTITATIVE EVALUATION RESULTS OF THE IMPROVED ASPP EFFECT

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
STMNet w/o improved ASPP	11.63	12.81	87.78	78.22
STMNet w/ improved ASPP	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>

The best results are highlighted in bold. w/ means with and w/o means without.

TABLE IX  
QUANTITATIVE EVALUATION RESULTS OF THE EDGE LOSS FUNCTION EFFECT

Method	$\beta$	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
STMNet w/o $loss_{edge}$	0	12.23	11.48	88.15	78.81
	5	12.01	11.66	88.16	78.83
STMNet w/ $loss_{edge}$	10	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>
	15	11.60	<b>11.34</b>	88.53	79.42
	20	12.11	12.66	87.61	77.96

The best results are highlighted in bold. w/ means with and w/o means without.

effect of the edge loss function and  $\beta$  on the extraction accuracy is presented in Table IX. Notably,  $\beta = 0$  means the edge loss function is not applied. It can be seen that when  $\beta$  is set to 10, the performance is the best, achieving the best  $fa$ ,  $F1$ , and  $IoU$  of 10.60%, 88.61%, and 79.55%, respectively. Compared to not using the edge loss function,  $fa$  decreases by 1.63 percentage points, and  $F1$  and  $IoU$  increase by 0.46 and 0.74 percentage points, respectively, with only a slight decrease in  $mr$ . When  $\beta$  is set to 15, the optimal  $mr$  of 11.34 is achieved.

Figs. 17 and 18 further validate the effectiveness of the edge loss function, both from visualization and quantitative evaluation. Fig. 17 illustrates that there are fewer false extracted and missed extracted pixels at the edges of UIS, and the extracted boundaries are more consistent with the ground truth. Fig. 18 quantifies the number of FP and FN pixels within the buffer area around the edges of UIS. It demonstrates that the use of edge loss function effectively reduces the number of FP and FN pixels within the buffer area.

### E. Ablation Experiments With Multiple Modules

In this study, both HRI and HTF are utilized for UIS extraction. In addition, SIAC is introduced to determine whether the image patch contains UIS or not. The improved ASPP is used to extract multiscale and multireceptive field features, and the edge loss function is employed to penalize errors at boundaries. In this section, ablation experiments are conducted to evaluate the effectiveness of these modules for UIS extraction. The experimental results, as shown in Table X, indicate a gradual improvement in the  $fa$ ,  $F1$ , and  $IoU$  with the introduction of each

TABLE X  
QUANTITATIVE EVALUATION RESULTS OF ABLATION STUDY

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
baseline	15.85	13.57	85.27	74.33
Baseline w/ HTF	16.12	12.33	85.73	75.03
Baseline w/ (HTF and SIAC)	12.46	12.52	87.51	77.79
Baseline w/ (HTF, SIAC, and Improved ASPP)	11.44	12.40	88.08	78.69
STMNet	<b>10.60</b>	<b>12.17</b>	<b>88.61</b>	<b>79.55</b>

The best results are highlighted in bold. w/ means with and w/o means without.

TABLE XI  
QUANTITATIVE EVALUATION RESULTS OF THE NEGATIVE SAMPLE ABLATION STUDY

Dataset	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
Dataset w/o negative samples	14.83	12.52	86.31	75.92
Dataset w/ negative samples	<b>14.10</b>	<b>10.73</b>	<b>87.55</b>	<b>77.86</b>

The best results are highlighted in bold. w/ means with and w/o means without.

module. However, the  $mr$  slightly increases with the introduction of SIAC. Nevertheless, overall, there is still a decreasing trend. STMNet, which integrates all modules, exhibits the best performance across all evaluation metrics.

Furthermore, the UIS dataset constructed in this study retains negative samples that do not contain UIS, primarily to aid in the training of SIAC and enhance the model's generalization capabilities. To explore the effectiveness of this strategy, we conducted comparative experiments, specifically applying STMNet without SIAC. The results, as shown in Table XI, demonstrate that the removal of negative samples leads to an increase in errors, particularly in terms of false extractions. Specifically, the  $fa$  increased by 0.73 percentage points, the  $mr$  by 1.79 percentage points, the  $F1$  decreased by 1.24 percentage points, and the  $IoU$  decreased by 1.94 percentage points.

### F. Generalizability Analysis With Additional Datasets

To further prove the generalization performance of the proposed STMNet, we conduct comparative experiments on the publicly available UIS-Shenzhen dataset [11]. This dataset is constructed using HRI of Shenzhen, Guangdong Province, downloaded from Google Earth in 2020, with a resolution of 1.19 m, along with building polygons data obtained from Baidu Maps.

The results of the comparative experiments are shown in Table XII. STMNet still achieves the best performance in the metrics of  $fa$ ,  $F1$  and  $IoU$ . It has a 0.005 percentage point lower  $fa$  than the suboptimal method (PSPNet), a 0.86 percentage point higher  $F1$  than the suboptimal method (UNetFormer), and a 1.24 percentage points higher  $IoU$  than the suboptimal method, also

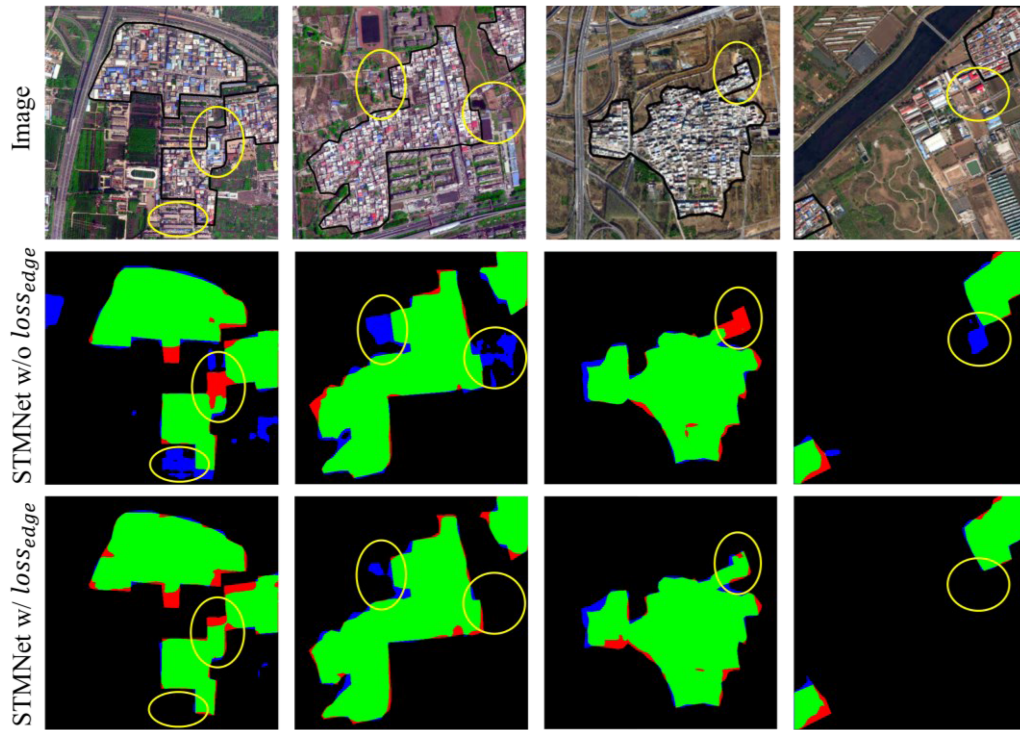


Fig. 17. Impact of the edge loss function on the UIS extraction results. Different colors represent the correctness or incorrectness of the results, including TP (green), TN (black), FP (blue), and FN (red).

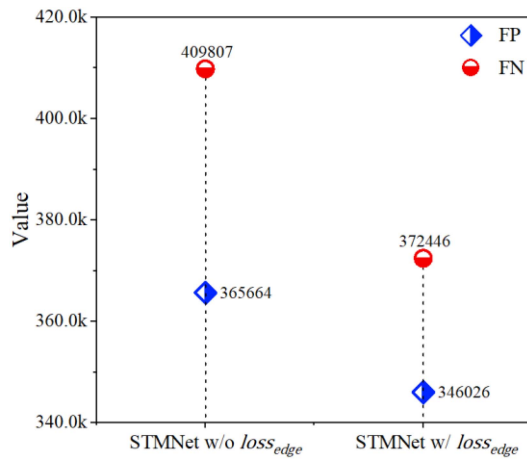


Fig. 18. Number of FP and FN pixels within the buffer area.

TABLE XII  
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT METHODS ON THE  
UIS-SHENZHEN DATASET

Method	$fa \downarrow$	$mr \downarrow$	$F1 \uparrow$	$IoU \uparrow$
STMNet	<b>12.04</b>	12.67	<b>87.64</b>	<b>78.01</b>
SDSC-UNet	14.79	12.98	86.11	75.60
UNetFormer	12.46	13.81	86.86	76.77
UisNet	14.27	<b>12.15</b>	86.78	76.64
MAResU-Net	14.49	12.44	86.52	76.25
MANet	14.42	14.10	85.74	75.04
DeepLabv3+	12.20	16.77	84.98	73.88
PSPNet	12.09	18.04	84.83	73.66

The best results are highlighted in bold.

TABLE XIII  
COMPUTATIONAL COMPLEXITY AND EFFICIENCY OF DIFFERENT METHODS

Method	Params size	FLOPs	Inference Time
STMNet	109.03	87.645	567.53
SDSC-UNet	81.33	64.18	807.28
UNetFormer	44.87	35.12	856.58
UisNet	106.82	377.51	—
MAResU-Net	100.31	104.35	915.55
MANet	137.03	231.95	887.64
DeepLabv3+	209.39	248.95	1072.05
PSPNet	9.19	29.19	833.58

UNetFormer. In terms of  $mr$ , STMNet ranked third, only behind UisNet and MAResU-Net by 0.52 and 0.23 percentage points, respectively. Although not the best in  $mr$ , STMNet's ability to balance  $fa$  and  $mr$  still demonstrates its excellent performance.

### G. Computational Complexity and Efficiency

To evaluate the computational complexity and efficiency of STMNet and the other seven methods, the params size, FLOPs, and inference time for the physical urban areas of Beijing are calculated. The results are listed in Table XIII. It should be noted that since UisNet requires multimodal data inputs, it is not tested on the self-produced dataset, and thus, no inference times for UisNet are reported. The analysis reveals that PSPNet has the smallest params size, while DeepLabv3+ has the largest. STMNet's params size is only better than DeepLabv3+ and MANet. This is due to its use of the pseudo-siamese backbone with nonshared parameters, resulting in a larger params size of



85.34 MB. In terms of FLOPs, SDSC-UNet registers the lowest, and MANet registers the highest. STMNet is at an average level among the various methods. However, when considering the inference time, STMNet consumes much less time than other methods, which is mainly attributed to the application of SIAC.

#### H. Limitations and Possible Improvements

The proposed STMNet has achieved higher accuracy in UIS extraction, which has been proved in the experiments, but there is still content worth studying to improve the accuracy further.

- 1) *Improvements to the Current Model:* The application of SIAC decreases false-extraction, but scene classification errors lead to an increase in miss-extraction. Therefore, future studies could implement measures such as multi-level scene classification to enhance accuracy. In addition, although the use of SIAC has significantly reduced inference time, techniques such as model pruning and reparameterization still need to be further explored to reduce computation.
- 2) *Application of New Technologies:* To fully exploit the information in HRI and HTF, we will explore the integration of effective Transformer or Mamba structures with CNNs. To address the pressure of obtaining training samples, we plan to combine pretrained large computer vision models (such as segment anything model) with unsupervised, weakly supervised, and semi-supervised methods for UIS extraction. For challenging samples in UIS extraction tasks, we will adopt effective hard sample mining and constraint modules to improve accuracy. These methods will be attempted in future research.
- 3) *Expanding the Research Area:* Despite efforts in most developing countries to transform UIS, cities with rapid urbanization rates, like Guangzhou in China and Dar es Salaam in Tanzania, still have many such areas. Therefore, future studies will focus on studying the distribution and scale of UIS in these regions. This will provide reference data for urbanization development in developing countries.

#### V. CONCLUSION

In this article, an STMNet is proposed for extracting large-scale UIS. The proposed STMNet takes HRI and HTF as inputs and incorporates FAFM, SIAC, improved ASPP, and the edge loss function to tackle the complex characteristics of UIS. Comparative experiments are conducted on a self-produced UIS extraction dataset and the publicly available UIS-Shenzhen dataset. The results are analyzed for both visualization and quantitative evaluation, with the STMNet achieving remarkable results. The experiments also demonstrate that each module incorporated into the proposed STMNet contributes significantly to the UIS extraction. From the perspective of method efficiency, the application of SIAC has significantly decreased the inference time. Future studies will focus on trying new segmentation methods for UIS extraction to improve accuracy. In addition, we aim to expand the study area for extracting UIS, especially in developing countries, to promote the achievement of global sustainable development goals.

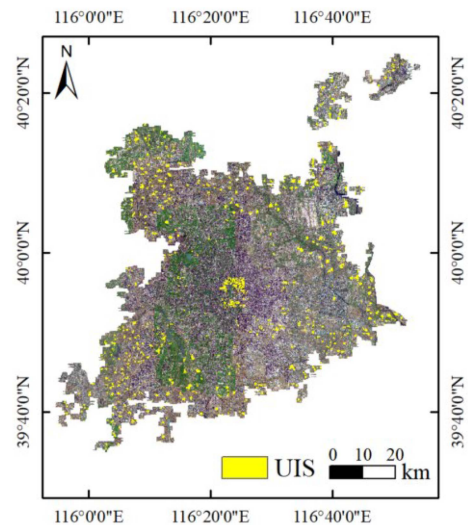


Fig. 19. UIS extraction results within the physical urban areas of Beijing.

#### APPENDIX

The extraction results of UIS within the physical urban areas of Beijing can be seen in Fig. 19. All the data and results in this study are available online.<sup>2</sup>

#### REFERENCES

- [1] Q. Peng et al., "Identification of densely populated-informal settlements and their role in Chinese urban sustainability assessment," *Gisci. Remote Sens.*, vol. 60, no. 1, Dec. 2023, Art. no. 2249748.
- [2] J. Wang, M. Kuffer, and K. Pfeffer, "The role of spatial heterogeneity in detecting urban slums," *Comput. Environ. Urban Syst.*, vol. 73, pp. 95–107, Jan. 2019.
- [3] R. A. Ansari and K. M. Buddhiraju, "Textural segmentation of remotely sensed images using multiresolution analysis for slum area identification," *Eur. J. Remote Sens.*, vol. 52, pp. 74–88, Aug. 2019.
- [4] Z. K. Pan, J. S. Xu, Y. B. Guo, Y. M. Hu, and G. X. Wang, "Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net," *Remote Sens.*, vol. 12, no. 10, May 2020, Art. no. 1574.
- [5] H. B. Wei, Y. Cao, and W. Qi, "The vanishing and renewal landscape of urban villages using high-resolution remote sensing: The case of Haidian district in Beijing," *Remote Sens.*, vol. 15, no. 7, Apr. 2023, Art. no. 1835.
- [6] S. Ouyang, S. Du, X. Zhang, S. Du, and L. Bai, "MDFF: A method for fine-grained ufs mapping with multimodal geographic data and deep network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9951–9966, 2023.
- [7] J. C. Duque, J. E. Patino, and A. Betancourt, "Exploring the potential of machine learning for automatic slum identification from VHR imagery," *Remote Sens.*, vol. 9, no. 9, Sep. 2017, Art. no. 895.
- [8] C. M. Gevaert, C. Persello, R. Sliuzas, and G. Vosselman, "Informal settlement classification using point-cloud and image-based features from UAV data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 125, pp. 225–236, Mar. 2017.
- [9] D. Matarira, O. Mutanga, and M. Naidu, "Google earth engine for informal settlement mapping: A random forest classification using spectral and textural information," *Remote Sens.*, vol. 14, no. 20, Oct. 2022, Art. no. 5130.
- [10] S. H. Du, J. H. Xing, J. Li, S. H. Du, C. Y. Zhang, and Y. Q. Sun, "Open-pit mine extraction from very high-resolution remote sensing images using OM-DeepLab," *Natural Resour. Res.*, vol. 31, no. 6, pp. 3173–3194, Dec. 2022.
- [11] R. Y. Fan, F. P. Li, W. Han, J. N. Yan, J. Li, and L. Z. Wang, "Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630316.

<sup>2</sup>[Online]. Available at: <https://doi.org/10.6084/m9.figshare.25009637.v3>.

- [12] C. Z. Wei et al., "Gaofen-2 satellite image-based characterization of urban villages using multiple convolutional neural networks," *Int. J. Remote Sens.*, vol. 44, no. 24, pp. 7808–7826, Dec. 2023.
- [13] Y. X. Chen, F. F. Peng, S. Yao, and Y. X. Xie, "Lightweight multilevel feature-fusion network for built-up area mapping from Gaofen-2 satellite images," *Remote Sens.*, vol. 16, no. 4, Feb. 2024, Art. no. 716.
- [14] H. G. Yue, L. B. Qing, Z. X. Zhang, Z. Y. Wang, L. Guo, and Y. H. Peng, "MSE-Net: A novel master-slave encoding network for remote sensing scene classification," *Eng. Appl. Artif. Intell.*, vol. 132, Jun. 2024, Art. no. 107909.
- [15] R. Ji, K. Tan, X. Wang, C. Pan, and L. Xin, "PASSNet: A spatial-spectral feature extraction network with patch attention module for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5510405.
- [16] X. Tang, T. Zhang, J. Ma, X. Zhang, F. Liu, and L. Jiao, "WNet: W-shaped hierarchical network for remote-sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615814.
- [17] X. Wang, K. Tan, P. Du, C. Pan, and J. Ding, "A unified multiscale learning framework for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4508319.
- [18] Z. Li, H. Chen, J. J. Wu, J. Li, and N. Jing, "SegMind: Semisupervised remote sensing image semantic segmentation with masked image modeling and contrastive learning method," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4408917.
- [19] X. Y. Bian, Z. F. Yang, C. S. Hu, M. Peng, and J. S. Tang, "Location-aware multi-code generator for remote sensing scene classification," *Int. J. Remote Sens.*, vol. 45, no. 8, pp. 2519–2547, Apr. 2024.
- [20] D. Verma, A. Jana, and K. Ramamritham, "Transfer learning approach to map urban slums using high and medium resolution satellite imagery," *Habitat Int.*, vol. 88, p. 12, Jun. 2019.
- [21] C. Persello and A. Stein, "Deep fully convolutional networks for the detection of informal settlements in VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2325–2329, Dec. 2017.
- [22] X. Zhang, S. Xiong, X. Dong, and S. Du, "Synergistic classification of multilevel land patches (SC-MLPs): Reducing conflicts and improving mapping results for land uses and functional spaces with very-high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4410217.
- [23] Q. Q. Zhu et al., "Knowledge-guided land pattern depiction for urban land use mapping: A case study of Chinese cities," *Remote Sens. Environ.*, vol. 272, Apr. 2022, Art. no. 112916.
- [24] C. Y. Zhang, J. H. Xing, J. Li, S. H. Du, and Q. M. Qin, "A new method for the extraction of tailing ponds from very high-resolution remotely sensed images: PSVED," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2681–2703, Dec. 2023.
- [25] J. M. Peña, P. A. Gutiérrez, C. Hervás-Martínez, J. Six, R. E. Plant, and F. López-Granados, "Object-based image classification of summer crops with machine learning methods," *Remote Sens.*, vol. 6, no. 6, pp. 5019–5041, Jun. 2014.
- [26] J. Meng et al., "Urban ecological land extraction from Chinese Gaofen-1 data using object-oriented classification techniques," in *Proc. IEEE Int. Symp. Geosci. Remote Sens.*, 2015, pp. 3076–3079.
- [27] L. N. Hao, Z. Zhang, and X. X. Yang, "Mine tailing extraction indexes and model using remote-sensing images in southeast Hubei province," *Environ. Earth Sci.*, vol. 78, no. 15, Aug. 2019, Art. no. 493.
- [28] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Lecture Notes Comput. Sci.*, 2015, pp. 234–241.
- [30] L. C. E. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Lecture Notes Comput. Sci.*, 2018, pp. 833–851.
- [31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [32] J. Wang, F. Chen, M. Zhang, and B. Yu, "NAU-Net: A new deep learning framework in glacial lake detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2000905.
- [33] C. J. Ge, W. J. Xie, and L. K. Meng, "Extracting lakes and reservoirs from GF-1 satellite imagery over China using improved U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 1504105.
- [34] L. H. Wang et al., "Dynamic inversion of inland aquaculture water quality based on UAVs-WSN spectral analysis," *Remote Sens.*, vol. 12, no. 3, Feb. 2020, Art. no. 402.
- [35] H. Wang and Y. L. Yu, "Deep feature fusion for high-resolution aerial scene classification," *Neural Process. Lett.*, vol. 51, no. 1, pp. 853–865, Feb. 2020.
- [36] Z. Guo, H. Liu, Z. Zheng, X. Chen, and Y. Liang, "Accurate extraction of mountain grassland from remote sensing image using a capsule network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 6, pp. 964–968, Jun. 2021.
- [37] H. Lu, C. Liu, N. W. Li, X. Fu, and L. G. Li, "Optimal segmentation scale selection and evaluation of cultivated land objects based on high-resolution remote sensing images with spectral and texture features," *Environ. Sci. Pollut. Res.*, vol. 28, no. 21, pp. 27067–27083, Jun. 2021.
- [38] W. Li, C. Yang, Y. Peng, and J. Du, "A pseudo-siamese deep convolutional neural network for spatiotemporal satellite image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1205–1220, 2022.
- [39] C. Wu, L. Chen, Z. He, and J. Jiang, "Pseudo-siamese graph matching network for textureless objects' 6-D pose estimation," *IEEE Trans. Ind. Electron.*, vol. 69, no. 3, pp. 2718–2727, Mar. 2022.
- [40] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [41] S. H. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Lecture Notes Comput. Sci.*, 2018, pp. 3–19.
- [42] L. Yang, W. Chen, P. S. Bi, H. Z. Tang, F. J. Zhang, and Z. Wang, "Improving vegetation segmentation with shadow effects based on double input networks using polarization images," *Comput. Electron. Agriculture*, vol. 199, Aug. 2022, Art. no. 107123.
- [43] R. X. Zhu, L. Yan, N. Mo, and Y. Liu, "Attention-based deep feature fusion for the scene classification of high-resolution remote sensing images," *Remote Sens.*, vol. 11, no. 17, Sep. 2019, Art. no. 1996.
- [44] H. Wu, P. Huang, M. Zhang, W. Tang, and X. Yu, "CMTFNet: CNN and multiscale transformer fusion network for remote-sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2004612.
- [45] Y. N. Feng, S. H. Du, S. W. Myint, and M. Shu, "Do urban functional zones affect land surface temperature differently? A case study of Beijing, China," *Remote Sens.*, vol. 11, no. 15, Aug. 2019, Art. no. 1802.
- [46] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.
- [47] X. Zhang, Y. Liu, Y. Lin, Q. Liao, and Y. Li, "UV-SAM: Adapting segment anything model for urban village identification," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 22520–22528.
- [48] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
- [49] R. H. Zhang, Q. Zhang, and G. X. Zhang, "SDSC-UNet: Dual skip connection ViT-based U-shaped model for building extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 6005005.
- [50] L. B. Wang et al., "UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 190, pp. 196–214, Aug. 2022.
- [51] R. Li et al., "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607713.
- [52] R. Li, S. Y. Zheng, C. X. Duan, J. L. Su, and C. Zhang, "Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8009205.
- [53] H. S. Zhao, J. P. Shi, X. J. Qi, X. G. Wang, and J. Y. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6230–6239.
- [54] S. H. Du et al., "IMG2HEIGHT: Height estimation from single remote sensing image using a deep convolutional encoder-decoder network," *Int. J. Remote Sens.*, vol. 44, no. 18, pp. 5686–5712, Sep. 2023.
- [55] Y. Wang, L. Gu, X. Li, and R. Ren, "Building extraction in multitemporal high-resolution remote sensing imagery using a multifeature LSTM network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 9, pp. 1645–1649, Sep. 2021.
- [56] K. G. Nambiar, V. I. Morgenshtern, P. Hochreuther, T. Seehaus, and M. H. Braun, "A self-trained model for cloud, shadow and snow detection in Sentinel-2 images of snow- and ice-covered regions," *Remote Sens.*, vol. 14, no. 8, Apr. 2022, Art. no. 1825.



**Shouhang Du** received the Ph.D. degree in cartography and geographic information system from the Peking University, Beijing, China, in 2021.

He is currently a Lecturer with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing. His research interests include intelligent understanding of geospatial data and deep learning.



**Liguang Wei** received the master's degree in surveying and mapping engineering from China University of Mining and Technology, Beijing, China, in 2014.

He is currently a Department Director with PIESAT Information Technology Company, Ltd. His research interest focuses on the application of deep learning in remote sensing.



**Jianghe Xing** is currently working toward the Ph.D. degree with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China.

His current research interests include remote sensing and land use classification.



**Yirui Zhang** is currently working toward the master's degree with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China.

Her current research interests include urban ecology and urban function.



**Shaoyu Wang** is currently working toward the Master's degree with the College of Geoscience and Surveying Engineering, China University of Mining and Technology, Beijing, China.

His current research interests include urban functional zone classification, deep learning, and computer vision.