Fine-Grained Urban Village Extraction by Mask Transformer From High-Resolution Satellite Images in Pearl River Delta

Zhuoqun Chai, Mengxi Liu, Qian Shi, *Senior Member, IEEE*, Yuanyuan Zhang, Minglin Zuo, and Da He, *Member, IEEE*

***Abstract*—Urban renewal has led to the proliferation of informal urban habitats, such as slums, shanty towns, and urban villages (UVs). As an important component of urban renewal, UVs influence urban spatial structure and land use patterns. Therefore, the fine extraction of UV is of great theoretical and practical significance. Existing UV classification techniques mostly employ machine learning and convolutional neural network based models, which struggle to perceive long-range global semantic information. In this article, based on high-resolution remote sensing images, we propose a multiscale mask transformer model for UV (MaskUV). It can extract both local texture features and global features. The multiscale mask transformer module with mask attention can aggregate different levels of pixel and object features, enhancing the model's recognition and generalization abilities. We extracted UV in seven cities in the Pearl River Delta (PRD) using MaskUV and analyzed the spatial pattern and accessibility of UV. Due to the scarcity of fine-grained UV detection datasets, we also provide a novel dataset (UVSet) containing 3415 pairs of 512 × 512 high-resolution UV images and labels, with a spatial resolution of 1 m. Comparative experiments with several UV extraction models demonstrate the effectiveness of MaskUV, achieving an $F1$ score of 84.39% and an IoU of 73.00% on UVSet. Besides, MaskUV achieves highly accurate detection results in seven cities in the PRD, with average $F1$ and IoU values of 84.41% and 72.44%, respectively.**

***Index Terms*—Deep learning, Pearl River Delta (PRD), remote sensing, urban villages (UVs), urbanization.**

I. Introduction

URBANIZATION reshapes the physical, social, and ecological landscapes of cities globally, with significant effects on biodiversity, ecosystem functions, and service provisions [1], [2], [3]. Accelerated urban renewal, while propelling regional economic expansion, concurrently leads to the proliferation of unofficial urban habitats, such as slums, shantytowns, and urban villages (UVs) [4], [5], [6]. The UN-habitat defines UV as densely populated, informal urban areas marked by poverty and subpar standards [7]. These communities often fail to keep pace with the swift advancement of city development, leading to poor living conditions and insufficient infrastructure. According to UN-habitat statistics, nearly 1.1 billion individuals currently live in slums or similar impoverished urban conditions, with projections suggesting an increase to 2 billion over the next three decades [8]. Regularly falling outside conventional urban administrative structures, these locales are typically characterized by inadequate public spaces, uncontrolled land use, overpopulated substandard buildings, and lack of essential amenities, leading to unsanitary conditions [9], [10], [11], [12]. With the transformation of China's urbanization development, the prime focus is high-quality and sustainable development [13], [14], [15]. Timely and accurate data on these informal settlements are crucial for improving urban spaces and living standards [16], [17].

UVs, a prevalent type of informal residential area in China, often emerge when the government circumvents rural settlements to mitigate costs during urban expansion, exemplifying China's localized informal residential spaces [18], [19], [20]. In the Pearl River Delta (PRD), UVs are dense, informal settlements that evolved from rural villages and have been incrementally engulfed by urbanization [21], [22], [23]. Despite their informality, they offer affordable housing for migrant workers, bolstering local economies [24]. Yet, they face issues including poor living conditions, insufficient infrastructure, and high pollution levels. In recent decades, numerous UVs in seven cities of PRD (see Section II-A) have been drastically transformed through urban renewal projects, causing substantial shifts in land use, socioeconomic statuses, and urban ecosystems.

On-site surveys or government statistic data can offer fundamental information on UV, from land and building areas to building density and average floors. However, these labor-intensive and inefficient methods struggle with large-scale implementation, failing to meet current application demands for accessibility and openness. High spatial resolution satellite imagery, with its broad observation range, rich surface data, and ready availability, has become a crucial data source for urban planning and management [25], [26], [27]. Researchers worldwide have utilized high-resolution imagery and associated technologies for extensive identification in UV and slums [7], [11], [28].

Manuscript received 31 December 2023; revised 19 May 2024; accepted 22 July 2024. Date of publication 5 August 2024; date of current version 8 August 2024. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3903402, and in part by the National Natural Science Foundation of China under Grant 42222106 and Grant 42201340. *(Corresponding authors: Da He.)*

The authors are with the Guangdong Provincial Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China (e-mail: chaizhq@mail2.sysu.edu.cn; liumx23@mail2.sysu.edu.cn; zhangyy388@mail2.sysu.edu.cn; zuomlin@mail2.sysu.edu.cn; heda@mail.sysu.edu.cn).

The dataset in the article will be available for download at https://github.com/xls111/UVSet.

Digital Object Identifier 10.1109/JSTARS.2024.3434487

Numerous studies have leveraged spectral and textural features from imagery, object segmentation, and machine learning to identify and analyze UV, given their distinct differences from conventional urban areas in terms of building density, roof materials, and living conditions. Specific methods include [29]'s semiautomatic identification of informal residences from QuickBird high-resolution imagery in Delhi, India, using multiresolution segmentation and object-oriented classification methods. Zhu et al. [30] identified informal settlements in developing countries using a decision tree model based on extracted features of buildings, roads, and spatial patterns. In China, Huang et al. [11] studied the spatiotemporal distribution of UV in Shenzhen and Wuhan using machine learning and multi-index scene models. Nevertheless, the unplanned growth of UV results in complicated spectral and spatiotemporal building patterns, exposing issues in traditional feature extraction methods, such as inadequate feature representation, complex extraction processes, and challenges in adapting to dynamic environments.

Relying on the robust automatic feature learning capabilities, deep learning has progressively become a focal method for fine-grained building scene classification and mapping in remote sensing imagery [31], [32], [33], [34], [35]. Deep learning offers an accurate, automated, and scalable method for extracting fine-scale architectural features from high-resolution imagery [36], [37], [38], which reduces manual labor and time consumption by learning complex patterns from extensive data. Furthermore, deep-learning models can be trained to perform in various urban contexts and landscapes, enabling efficient large-scale mapping of UV. Shi et al. [39] employed deep fully convolutional networks (FCNs) to automatically learn a hierarchy of informative features for detecting informal urban settlements. Vaswani et al. [40] explored the potential of FCNs for UV semantic segmentation in QuickBird high-resolution and Sentinel2 imagery via transfer learning. Wang et al. [41] tackled the data domain shift issue in large-scale UV identification through adversarial learning, achieving finer large-scale UV mapping.

Transformer [42] has been broadly applied in image classification, semantic segmentation, object detection, image generation, image captioning, and super-resolution tasks [43], [44], [45]. Current research on UV classification using deep learning primarily focuses on convolutional neural networks (CNNs) [7], [46], [47], with few studies employing vision transformer (ViT) for end-to-end semantic segmentation of UV. While CNNs excel at capturing local patterns within their receptive fields, transformers are designed to model long-range, global dependencies in the data, which is particularly beneficial for high-resolution remote sensing images where contextual information from a larger area can be crucial for semantic understanding. Moreover, unlike CNNs, transformers do not rely on spatially local, shift-invariant filters, and therefore, can learn to accommodate the diverse and complex spatial patterns of UV in remote sensing imagery [44], [45].

To better recognize the fine-grained features of UV and break through the constraints of a single sensor, high-resolution imagery, and multisource data fusion methods are employed to acquire granular socioeconomic attribute information. Chen et al. [17] combined remote sensing imagery with social sensing data, such as nighttime lights, points of interest (POI), and taxi trajectories to generate a fine-grained map of UV in Shenzhen. Fan et al. [7] validated the robustness of their fusion method in both custom and public datasets by extracting multiscale spatial fusion features from satellite and street view images based on a CNN. Hu et al. [48] designed network branches for satellite and street view images, using a gating module for multimodal feature fusion, effectively detecting UV in the Beijing–Tianjin–Hebei region. Although the effectiveness of multimodal data has been proven by multiple studies, it is challenging to acquire sufficient multimodal data for large-scale mapping in the PRD region. ~~Therefore, we plan to use high-resolution remote~~ sensing ~~imagery exclusively to develop an end-to-end urban village extraction framework based on CNN-Transformer architecture.~~

This article aims to leverage the power of deep-learning and high-resolution remote sensing to identify the accurate spatial distribution of UV in the PRD region. Traditional semantic segmentation methods designed primarily for natural images, often fall short in UV extraction due to their reliance on large, labeled datasets, and their general-purpose feature extraction approaches. These models typically struggle with the fine-grained differentiation required to distinguish UVs from other high-density urban areas, leading to reduced accuracy and higher misclassification rates. Specifically, we propose the mask transformer for UV extraction (MaskUV) to address these challenges by incorporating a novel multiscale mask transformer module (MMTM). This module enhances the model's ability to accurately capture the unique instance-level features of UVs, and leverages multiscale feature fusion to integrate both local and global contextual information, thereby improving segmentation performance in complex urban environments. Additionally, a high-resolution UV dataset (UVSet) is made available for model training and validation. Our findings could enrich existing single-building datasets, bolster data support for studies on building management and heat island effects, and offer valuable insights into the sustainable development of the PRD region. The contributions of this article can be summarized in three points as follows.

1) A multiscale MaskUV is proposed for UV, in which an MMTM is designed to encode multiscale semantic information, and the mask attention is integrated to capture local context and improve model efficiency.
2) A high-resolution UVSet in PRD is provided for relevant studies, which contains 3415 pairs of 512 $\times$ 512 images with spatial resolutions of 1.1 m.
3) The first UV map of seven cities in PRD with a 1-m spatial resolution for the year 2021 is produced. The spatial pattern of UV in the PRD and the potential application of the UV map are also analyzed.

The rest of this article is organized as follows. Section II introduces the study area and our UVSet. Section III presents the methodology. Section IV presents the experimental settings. Section V demonstrates and analyzes the accuracy results. Comparative experiments, ablation studies, and model efficiency are discussed in Section VI. Finally, Section VII concludes this article.

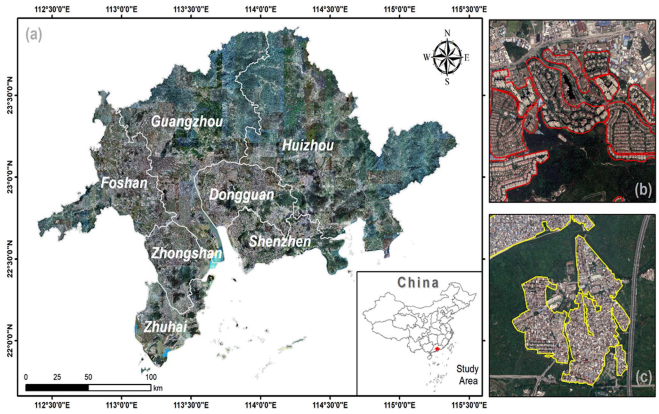

Fig. 1. Study area and study objects. (a) Study area of seven cities in Pearl River Delta, China, including Guangzhou, Shenzhen, Foshan, Dongguan, Zhongshan, Zhuhai, and Huizhou. (b) Examples of apartment complexes and villas. (c) Examples of UVs.

II. Study Area and Materials

A. Study Area

The PRD is located in the south-central part of Guangdong Province in China, which is now one of the pivotal megacity regions in the world. Over the past three decades, expeditious urbanization in the PRD has aroused the expansion and agglomeration of UV [49]. As UVs are formed in the process of rapid urban development, there is no unified planning and management and low living standards as residential areas [50]. Therefore, in order to explore and analyze the housing inequality phenomenon in China, the seven cities in PRD are selected as the study area (21°27′–23°56′N,111°59’–115°26’E), covering a total land area of 54770.21 km^2 (Fig. 1) and with a population of approximately 64.47 million in 2019 [51].

B. Data and Preprocessing

1) Google Earth Imagery: To obtain the distribution of informal housing space, a total number of 329 Google Earth imageries in 2020, covering the whole study area of the seven cities in the PRD, are downloaded for UV mapping. Each image has a 1:25 000 map size and a spatial resolution of about 1 m. All images are clear and cloud-free images obtained through artificial inspection, as demonstrated in Fig. 1.

2) Urban Village Dataset (UVSet): Considering that the specific distribution of informal housing space is difficult to obtain directly, a DL-based approach is introduced to obtain the distribution of UV of the seven studied cities. Nevertheless, while the performance of a DL model relies heavily on adequate datasets, there is no large-scale, publicly available dataset of UV based on HRIs. Therefore, a UVSet of HRIs is constructed to support DL model training and validation, which will also be open source for future scientific research.

The process of constructing the UVSet can be summarized as follows: first, 16 training images and 10 testing images are evenly and randomly selected from the 329 Google Earth images as sample area, as shown in Fig. 2; next, the 26 selected images are annotated through manual visual interpretation to obtain UV in vector, as shown in Fig. 1(b); finally, all vectors of UV annotated in the above steps will be rasterized into pixelwise annotations, where the non-UV area and the UV area are denoted by 0 and 255, respectively.

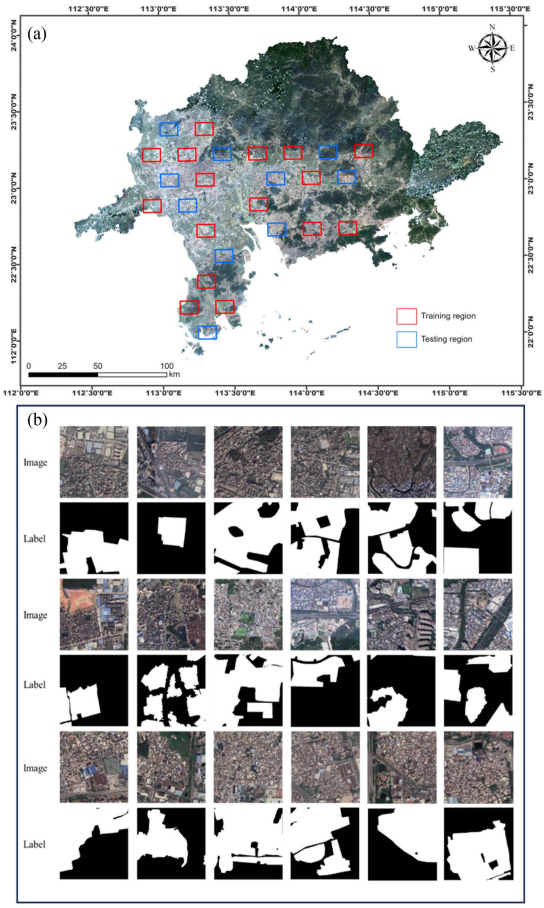

Fig. 2. Introduction of UVSet. (a) Distribution of UVSet training and testing data. (b) Example samples in the UVSet of size 512 × 512. Each sample contains an image and a corresponding label, in which “1” (the white pixels) denotes urban village area, and “0” (the black pixels) denotes nonurban village area.

After obtaining labeled image–sample pairs by the above steps, we need to crop them into patches to meet the requirements of model training on GPU. Therefore, 2102 pairs of 512 × 512 size samples are obtained through nonoverlapping sampling, which are separated into the training set, validation set, and test set in the ratio of 3:1:1. For the sake of making the trained model more robust for large-scale mapping, data augmentation strategies are applied to the training set. Specifically, random rotations of 90°, 180°, and 270° are performed at first to double the size of the training set, after which random left and right flips are made to expand the training set by three times. In addition, random gamma transformation (with a probability of 0.6) and a blurring (with a probability of 0.1) are also utilized in the above two steps. The distribution of samples and some example samples in the UVSet are provided in Fig. 2.

3) Global Urban Boundary: The global urban boundary (GUB) data [52] are multitemporal GUB data for 7 years (1990, 1995, 2000, 2005, 2010, 2015, and 2018) constructed based on the global high resolution (30 m) artificial impervious area

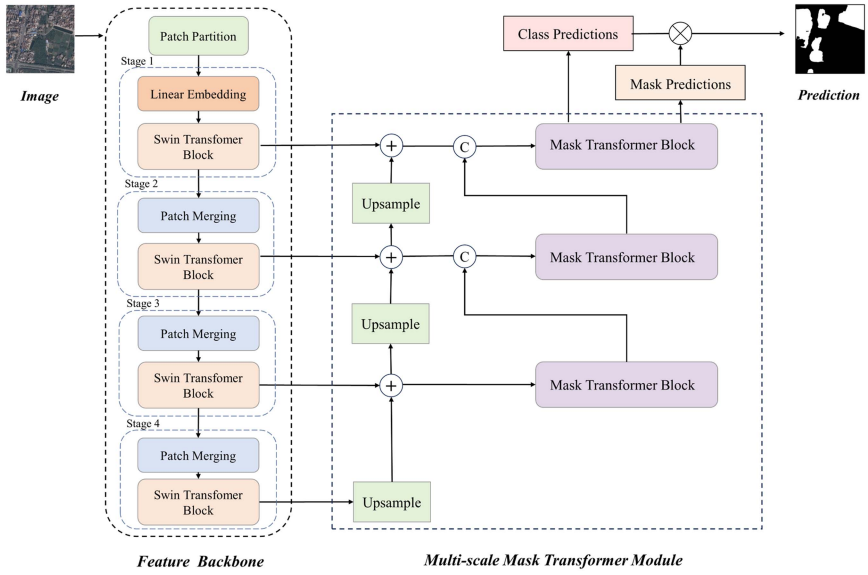

Fig. 3. Overview of the proposed mask transformer network.

data. The GUB dataset delineates the boundaries of all cities and surrounding settlements with an area of more than 1 square kilometer in the world, which can well capture the complicated shapes and boundary features of urban fringe areas. In this article, the GUB dataset is used to remove the aggregated villages outside the city to obtain the UV in the study area.

III. Methodology

A. Overview

As shown in Fig. 3, the proposed MaskUV mainly comprises two components: a swin transformer feature backbone and an MMTM. The backbone, made up of four swin transformer blocks, is designed to extract multiscale spatial and textural features. To address the complex architectural structures and multiscale characteristics of UV, we propose an MMTM that learns multilevel local and global contextual semantic features from satellite imagery, where a novel mask classification mechanism is utilized. The MMTM is composed of three mask transformer blocks, each consisting of a pixel-level module, a transformer module, and a segmentation module that handles features at different scales and semantic levels.

B. Feature Backbone

The backbone of the MaskUV initially uses the tiny version model of the swin transformer [53] as a feature backbone to seize multiscale image features effectively. To be specific, the backbone comprises a patch partition process and four stages of swin transformer modules. The swin transformer architecture adopts a patch-based approach similar to the ViT, dividing the input RGB image into nonoverlapping patches and projecting their raw pixel RGB values into an embedding space. These patches are then processed by multiple swin transformer blocks, featuring a unique shifted window-based multihead self-attention module followed by a two-layer multilayer perceptron (MLP) with GELU [54] nonlinearity. Hierarchical representation is achieved through patch merging layers, downsampling the token count by a factor of 2×2, and preserving the resolutions of the feature maps across successive stages. LayerNorm [55] layers and residual connections ensure stable training and information flow within each swin transformer block.

The hierarchical structure of swin transformer enables our backbone to capture multiscale information while maintaining computational efficiency, making it well-suited for semantic segmentation tasks.

C. Multiscale Mask Transformer Module

The MMTM is designed to effectively capture multilevel contextual semantic features for UV extraction. Comprising three mask transformer blocks, MMTM leverages a novel mask classification mechanism within each block to handle features at various scales and semantic levels. Each block within MMTM consists of three interconnected modules: a pixel-level module, a transformer module, and a segmentation module.

Fig. 4 illustrates the structure of the mask transformer block. The pixel-level module serves as the foundation of MaskFormer, extracting per-pixel embeddings to facilitate binary mask predictions. Initially, an image of size H $\times$ W is inputted into the model. A backbone network generates a low-resolution image feature map$F \in R^{C_F \times H/S \times W/S}$, where C_F represents the number of channels and Sdenotes the stride of the feature map. Subsequently, a pixel decoder gradually upsamples the features to generate per-pixel embeddings $\varepsilon_{\text{pixel}} \in R^{C_\varepsilon \times H \times W}$,where C_ε represents the embedding dimension. This module accommodates various per-pixel classification-based segmentation models including recent transformer-based architectures. In our module, the pixel decoder is seamlessly composed of 3×3 convolutions.

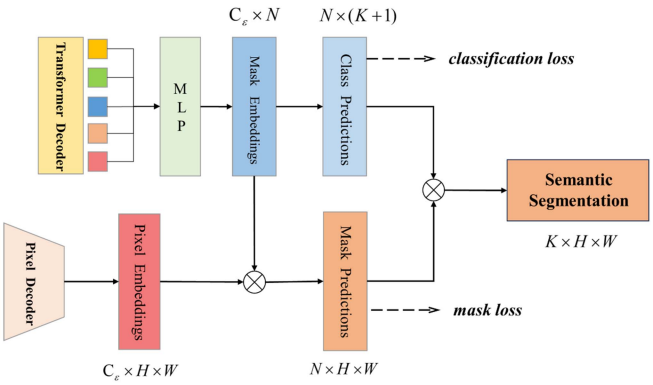

Fig. 4. Structure of the mask transformer block.

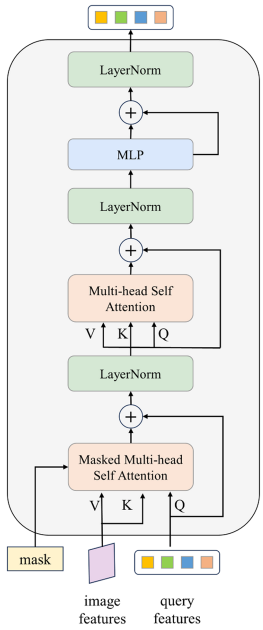

Fig. 5. Structure of the mask attention block.

The transformer module leverages a stack of transformer decoder layers to compute Nper-segment embeddings. By processing the image features $F$and learnable positional embeddings, the module outputs Nper-segment embeddings $Q \in R^{C_Q \times N}$. These embeddings encode global information about each segment. The decoder yields all predictions in parallel [56], enabling efficient and simultaneous processing of multiple segments.

While context features are crucial for image segmentation, recent findings [57], [58] suggest that the slow convergence of transformer-based models may result from the global context in the cross-attention layer. This is because it often requires extensive training epochs for cross-attention to effectively focus on localized object regions. To address this, we replace the standard cross-attention with the mask attention, a modification of cross-attention, which exclusively attends to the foreground regions of the predicted mask for each query. This approach aims to leverage local features for updating query features, while relying on self-attention mechanisms to gather contextual information efficiently. The structure of the mask attention block is depicted in Fig. 5.

The masked attention mechanism updates the attention matrix as

$$X_l = \text{softmax}(M_{l-1} + Q_l K_l^T) V_l + X_{l-1}. \tag{1}$$

Here, $l$is the layer index. $X_l \in R^{N \times C}$represents query features at layer l, and $Q_l = f_Q(X_{l-1}) \in R^{N \times C}$ is the linear transformation of input query features. X_0denotes the input query features to the transformer decoder. $K_l, V_l \in R^{H_l W_l \times C}$ are query and value from image features, respectively. $H_l$and W_l represent the spatial resolution of the image features.

The attention mask $M_{l-1}(x, y)$ at feature location (x, y) is defined as

$$M_{l-1}(x,y) = \begin{cases} 0 & \text{if } M_{l-1}(x,y) = 1 \\ -\infty & \text{otherwise.} \end{cases} \tag{2}$$

Here, $M_{l-1} \in 0, 1^{N \times H_l W_l}$ is the binarized output (thresholded at 0.5) of the resized mask prediction from the previous $(l-1)$thtransformer decoder layer, resized to match the resolution of K_l. M_0 represents the binary mask prediction obtained from X_0, prior to feeding the query features into the transformer decoder.

Finally, the segmentation module synthesizes the per-segment embeddings to generate class probability predictions for each segment. A linear classifier followed by a softmax activation is applied to the per-segment embeddings Q to yield class probability predictions $\{p_i \in \Delta^{K+1}\}_{i=1}^N$ for each segment. Utilizing a novel mask classification mechanism, the module produces binary mask predictions by applying an MLP to the embeddings. These predictions are then assembled into the final segmentation output, providing comprehensive insights into the spatial distribution and semantic characteristics of UVs in the satellite imagery.

To train the mask classification model, we require a matching σ between the set of predictions z and the set of N_{gt} truth segments $z_{gt} = \{(c_i^{gt}, m_i^{gt}) | c_i^{gt} \in \{1, \ldots, K\}, m_i^{gt} \in \{0,1\}^{H \times W}\}_{i=1}^{N^{gt}}$. Here, c_i^{gt} represents the ground truth class of the$i$th ground truth segment. Given that the sizes of the prediction set $|z| = N$ and the ground truth set $|z^{gt}| = N^{gt}$ are typically unequal, it is assumed $N \geq N^{gt}$ and the ground truth labels are padded with "no object" tokens $\varnothing$ to facilitate one-to-one matching. For extraction of UV, the ith prediction is matched to a ground truth region with class label i, and to $\varnothing$ if a region with class label i is not present in the ground truth. Recent research has shown that a bipartite matching-based assignment outperforms fixed matching. The approach involves utilizing class and mask predictions directly, where $\mathcal{L}_{\text{mask}}$ denotes a binary mask loss.

During training, MaskUV minimizes the combination of a cross-entropy classification loss and a binary mask loss for each predicted segment. The mask loss is a linear combination of focal loss and dice loss. This combined loss function ensures effective training of the model parameters, enhancing segmentation

accuracy and robustness

$$\mathcal{L}_{\text{mask-cls}}(z, z^{gt}) = \sum_{j=1}^{N} \Big[-\log p_{\sigma(j)}(c_j^{gt}) + 1_{c_j^{gt} \neq \varnothing} \mathcal{L}_{\text{mask}}(m_{\sigma(j)}, m_{gt}^{j}) \Big]. \quad (3)$$

D. Semantic Inference

The semantic inference adopted by MaskUV involves a simple matrix multiplication to compute the per-pixel class probabilities for UV and non-UV. Specifically, the most likely semantic label for each pixel is computed from the argmax over the sum of class probabilities weighted by their corresponding mask values

$$Y[h, w] = \text{argmax}_{c \in \{1,\ldots,K\}} \sum_{i=1}^{N} p_i(c) \cdot m_i[h, w]. \quad (4)$$

This effectively considers the contributions of multiple masks in determining the final semantic label for each pixel. It is important to note that the argmax operation does not include the "no object" category $\varnothing$, as standard semantic segmentation requires each output pixel to be assigned a label.

IV. Experimental Settings

A. Model Training Settings

All experiments are configured using PyTorch and trained on NVIDIA GeForce 2080ti. The model is trained with a batch size of 8, the initial learning rate is set to 0.0001, and the Adam optimizer is used to optimize the model parameters. The total number of epochs for model training is 150. After training of 100 epochs, the learning rate will be linearly decreased to help the model better reach the optimum. During training, data augmentation on the training set, including random rotation and random flipping, will also be applied.

B. Comparative Methods

To further test the validity of the MaskUV model, we introduce eight advanced semantic segmentation models for comparative experiments, including UNet, ENet, BiSeNet, Deeplab v3+, SegFormer, Segmenter, DDRNet, and PIDNet. A concise overview of the distinct features of each method is as follows.

1) UNet [59] is a CNN that was specifically designed for biomedical image segmentation, known for its symmetric expansive path that recovers the spatial information lost in the contracting path, enabling precise localization.
2) ENet [60] is an efficient CNN proposed by Adam Paszke et al. It is used for real-time semantic segmentation tasks, characterized by its lightweight architecture that carefully balances model complexity and accuracy.
3) BiSeNet [61] is designed for high-resolution image segmentation, featuring a unique architecture that simultaneously processes multiscale features through two parallel paths.
4) DeeplabV3+ [62] is an innovative encoder–decoder structured CNN that excels in semantic image segmentation. It utilizes an enhanced Xception-41 encoder, atrous spatial pyramid pooling, and a decoder module to encode multiscale context information and capture sharper boundaries.
5) SegFormer [63] is a state-of-the-art (SOTA) transformer-based architecture proposed for semantic segmentation tasks. It integrates transformers with lightweight MLP decoders, featuring a hierarchically structured transformer encoder for multiscale feature extraction and simplified MLP decoders combining local and global attention.
6) Segmenter [64] is an SOTA transformer model to acquire global context throughout the network, enabling global context modeling from the initial layer throughout the network and leveraging ViT to extend semantic segmentation with output embeddings of image patches.
7) DDRNet [65] is an efficient network designed for real-time semantic segmentation, featuring deep dual-resolution networks with multiple bilateral fusions and the deep aggregation pyramid pooling module for enhanced receptive fields and multiscale context integration.
8) PIDNet [66] is a novel three-branch network architecture inspired by the proportional–integral–derivative controllers, designed to parse detailed, context, and boundary information separately and mitigate overshoot issues through boundary attention-guided fusion.

C. Evaluation Metrics

The predicted results of the UV will be compared to the ground truth labels, and the accuracy will be measured by four indicators: precision, recall, $F1$, and IoU. Precision assesses the detection rate of predicted positive pixels, Recall assesses the detection rate of true positive pixels, and the $F1$ score is the summed average of precision and recall, aiming to consider the effects of both. The IoU is the intersection of the "UV" class in this article. They can be expressed by the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (8)$$

where TP, FP, FN, and TN indicate true positive, false positive, false negative, and true negative of the predicted results.

V. Analysis of Results

A. Accuracy Evaluation on UV Results

After obtaining the UV results of the study area using the well-trained MaskUV model on UVSet, an accuracy evaluation should be conducted to verify the results. Several test images from each city in the study area are evenly and randomly selected for validation. All test images will be annotated by expert visual interpretation to obtain corresponding reference maps. Fig. 6

Fig. 6. Comparisons between the reference maps (the second column) and predicted results (the third column) of UVs, cases from (a) Shenzhen City; (b) Foshan City; and (c) Dongguan City.

TABLE I
ACCURACY EVALUATION OF UV IN PRD

City	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Guangzhou	86.93	83.06	84.95	73.84
Shenzhen	73.33	90.71	81.10	70.21
Zhuhai	69.98	87.86	77.68	69.40
Dongguan	84.07	88.10	86.04	75.50
Zhongshan	79.78	87.94	83.66	71.91
Foshan	86.24	89.23	87.71	78.11
Huizhou	83.42	80.37	81.86	74.71
Average	**81.90**	**87.08**	**84.41**	**72.44**

The bold values indicate the highest accuracy values corresponding to the respective performance metrics in different columns.

shows the comparison between the reference maps and the extracted UV results from Shenzhen, Dongguan, and Foshan.

The accuracy evaluation of UV in PRD is shown in Table I. It can be seen that the UV extraction results for the whole study area are generally satisfactory, with the average $F1$ and IoU of 84.41% and 72.44%, respectively. Specifically, Foshan has the highest extraction accuracy, with the highest $F1$ and IoU of 87.71% and 78.11%, respectively. Following closely, Dongguan achieves an $F1$ score of 86.04% and an IoU of 75.50%. Guangzhou also yields notably high results, with an $F1$ score of 84.95% and the highest precision among all cities at 86.93%. Moreover, Shenzhen, Zhongshan, and Huizhou all exhibit $F1$ scores and IoU values surpassing 80% and 70%, respectively. However, Zhuhai presents the lowest verification accuracy, recording an $F1$ score of 77.68%. Overall, these accuracy evaluations underscore the effectiveness and practicality of the proposed MaskUV and UVSet for large-scale UV mapping, affirming the precision of the derived UV delineations for subsequent applications and analyses.

B. Comparative Experiments With DL Models

Fig. 7 illustrates the UV recognition results of several test images using different deep-learning models. Compared with other methods, MaskUV visually achieves more accurate classification results. The multiscale mask transformer architecture of MaskUV facilitates the model to recognize and localize multiscale buildings more accurately while maintaining high precision across different scenarios. In contrast, Deeplab v3+ and BiSeNet perform poorly in UV with low building density, leading to inaccurate boundary predictions of small buildings. Among the SOTA methods, SegFormer and Segmenter yield UV predictions with numerous holes, overlooking the local features of intricate contours of UV buildings and roads. Unet, DDRNet, and PIDNet suffer from many false negative predictions for roads with similar texture and color to UV buildings, indicating that their building extraction results are not robust enough to handle buildings in different complex scenes. Deeplab v3+ also performs poorly in generating predictions with smooth boundaries, suggesting that the pyramid pooling module is not suitable for building extraction from VHR remote sensing

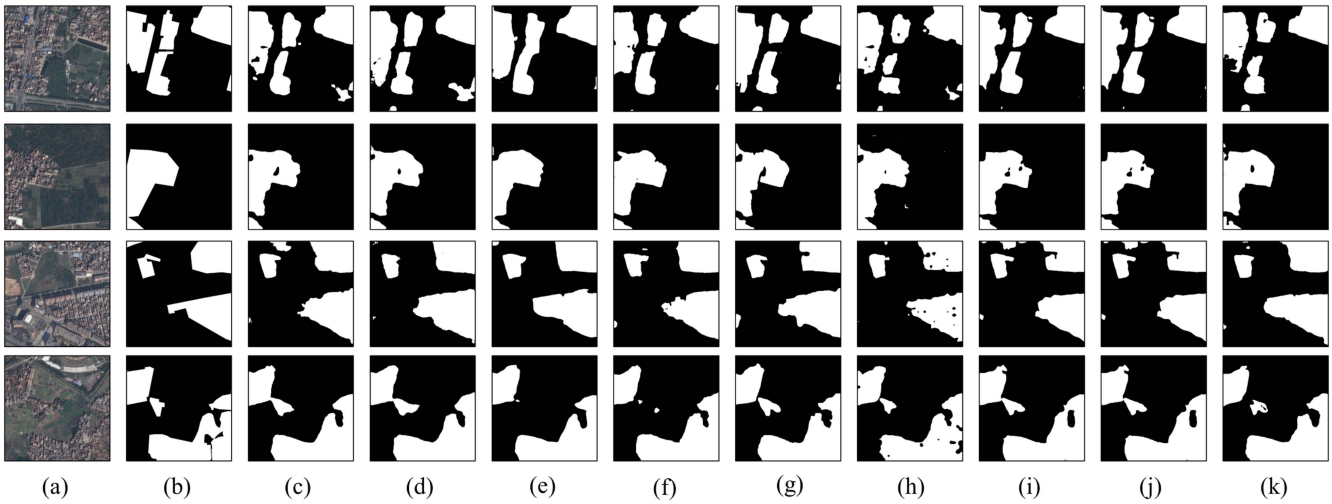

Fig. 7. Visualization of experimental results on UVSet dataset: (a) image; (b) label; (c) UNet; (d) ENet; (e) BiSeNet; (f) deeplab v3+; (g) SegFormer; (h) segmenter; (i) DDRNet; (j) PIDNet; and (k) MaskUV.

TABLE II
PERFORMANCE OF DIFFERENT MODELS ON UVSET

Model	Precision (%)	Recall (%)	F1 (%)	IoU (%)
UNet	81.67	83.99	82.82	70.67
ENet	78.96	**86.40**	82.51	70.23
BiSeNet	82.37	84.21	83.28	71.36
Deeplab v3+	81.99	84.33	83.14	71.15
SegFormer	81.62	85.60	83.56	71.77
Segmenter	79.43	86.01	82.59	70.35
DDRNet	82.58	83.44	83.01	70.95
PIDNet	**85.84**	79.41	82.50	70.21
MaskUV	82.84	86.01	**84.39**	**73.00**

The bold values indicate the highest accuracy values corresponding to the respective performance metrics in different columns.

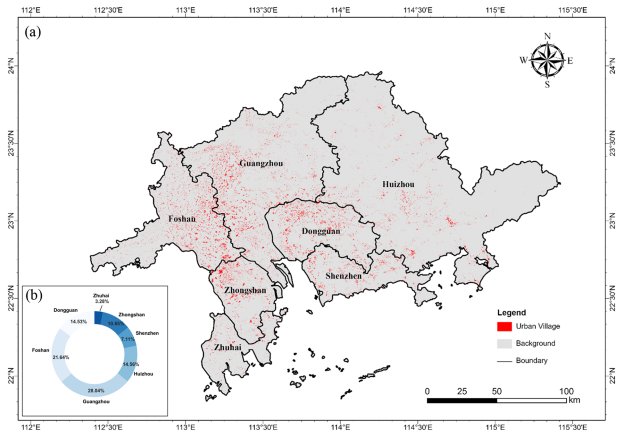

Fig. 8. Distribution and areal statistics of urban villages of PRD in 2020: (a) presents the spatial distribution of urban villages, and (b) shows urban village area statistics for different regions.

images, mainly because the pyramid pooling module aggregates contextual information at various scales, which usually leads to smoother prediction. BiSeNet performs well among the seven compared methods and achieves high accuracy in extracting UV and low-density buildings.

Four indicators including precision, recall, $F1$, and IoU are used for comparisons. As shown in Table II, MaskUV achieves the best performance among the comparative methods, with an $F1$ score and an IoU of 83.68% and 71.94%, respectively. The second-rank model is the SegFormer with an $F1$ of 83.56% and an IoU of 71.77%, which are 0.83% and 1.23% lower than those of MaskUV. The following are BiSeNet and Deeplab v3+, which obtain the $F1$ of 83.28% and 83.14%, respectively. While ENet achieves the highest recall rate of 86.40%, its $F1$ score and mIoU are not as competitive, standing at 82.51% and 70.23%, respectively. It can be seen from the quantitative results that, compared with the existing models, MaskUV has advantages in the extraction of UV.

C. Spatial Pattern of UVs in Seven Cities of PRD

Fig. 8 shows the spatial distribution of UV in the seven cities of PRD in 2020. The rapid urbanization of PRD has led to the widespread distribution of UV, with significant spatial heterogeneity. The total area of UV in the study area was estimated to be 827.64 km^2, with Guangzhou and Foshan having the largest proportions (28.04% and 21.64%, respectively), followed by Huizhou (14.56%), Dongguan (14.53%), and Zhongshan (10.85%), while Zhuhai had the smallest proportion (3.28%). From the kernel density map [67] of UV in Fig. 9, UVs exhibit distinct spatial patterns in different cities, which are related to the cities' own development history and spatial planning. Guangzhou's UVs are mainly concentrated in the west, while Foshan's are concentrated in the east, adjacent to Guangzhou. In contrast, Dongguan and Zhongshan have more UV in the north. The distribution of UV in Shenzhen, Zhuhai, and Huizhou is relatively dispersed, without an obvious center of concentration.

Compared with formal residential areas, most UVs still have significant deficiencies in the supply and quality of public services, such as sewage treatment and sanitation facilities. Since UVs are informal residential areas surrounded by built-up urban

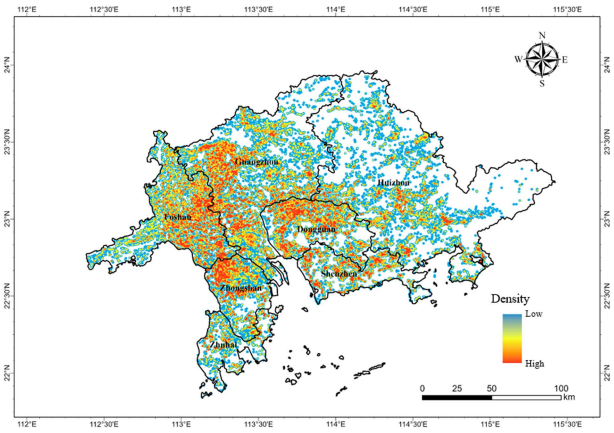

Fig. 9. Kernel density map of urban villages of PRD in 2020.

TABLE III
ABLATION STUDY OF THE PROPOSED MODEL ON UVSET

Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Base	78.48	85.35	81.77	70.28
+MTB	80.19	85.58	82.80	71.84
+MTM	81.78	85.88	83.78	72.24
MaskUV	82.84	86.01	84.39	73.00

areas, the accessibility of their public transportation systems, commercial services, and parks becomes an important indicator of the convenience of living for UV residents. We collected POI data in 2020 for the study area using the Amap API and calculated the Euclidean distance from the center of each UV to the nearest commercial center, park, and trunk road using nearest neighbor analysis. The nearest distances are shown in Fig. 10. Except for a few UVs in Guangzhou, Foshan, and Zhuhai, most UVs are conveniently located within 2 km of a trunk road, making it easy for residents to travel. The distance from UV to parks and commercial centers varies, with commercial centers generally being farther away. In Huizhou, the scattered distribution of commercial centers results in relatively long distances from UVs to these centers.

VI. DISCUSSION

A. Ablation Study

In this section, we conduct an ablation study on MaskUV to further validate the efficacy of the mask transformer block (MTB) and the multiscale transformer module (MTM) integrated into the MaskUV. The "base" model serves as the reference for comparison with the standard transformer and single branch configurations. The models denoted as "+MTB" and "MTM" represent the "base" model augmented with the MTB and MMTM components, respectively. The ablation study results are summarized in Table III.

Compared against the "base" model, which achieves an $F1$ score of 81.77%, the "+MTB" model and the "+MTM" model exhibit improvements of 1.03% and 2.01% in $F1$ score, respectively. This indicates the beneficial impact of integrating these modules into the base architecture. Furthermore, the MaskUV achieves the most favorable results in the ablation experiments,

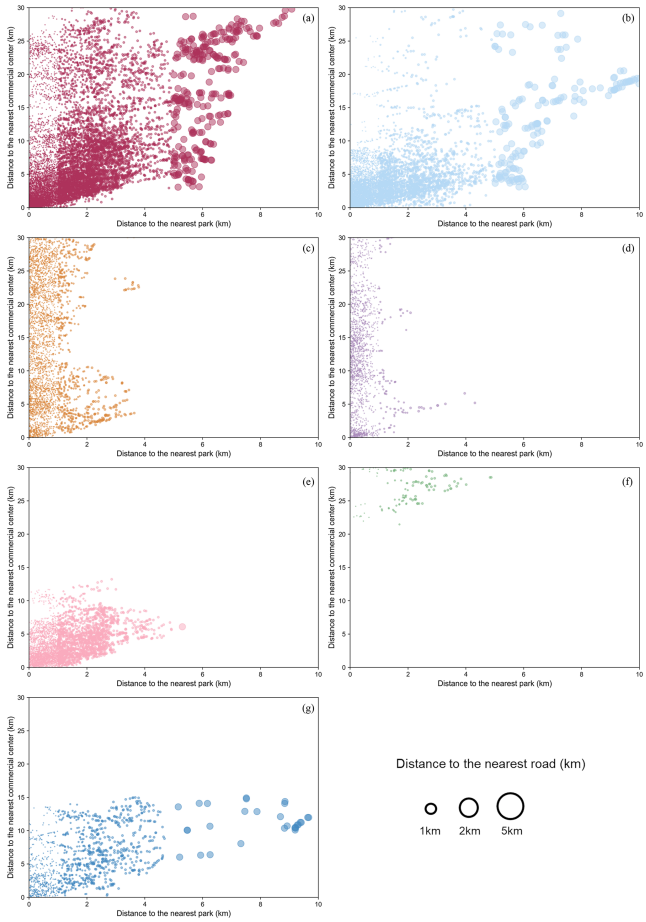

Fig. 10. Relationships between the distances to the nearest public facilities of UVs in (a) Guangzhou, (b) Foshan, (c) Dongguan, (d) Shenzhen, (e) Zhongshan, (f) Huizhou, and (g) Zhuhai.

with a precision of 82.84%, recall of 86.01%, $F1$ score of 84.39%, and IoU of 73.00%. These findings strongly support the feasibility and efficacy of integrating MMTM into the MaskUV architecture for UV mapping.

B. Model Complexity

To compare the computational complexity of different models, three metrics, including floating points of operations (FLOPs) and a number of parameters (Params). The FLOPs can denote the temporal computational complexity of the model, with a unit of 10^9 (G), and the Params represent the spatialwise complexity of the model, with a unit of 10^6 (M). Given two bitemporal inputs of size $1 \times 3 \times 512 \times 512$, the FLOPs and Params of all methods are shown in Table IV. The proposed MaskUV can achieve optimal performance under relatively lower FLOPs of 49.81G and Params of 41.63M, demonstrating its advantages in fast and large-scale UV extraction applications.

C. Limitations and Future Work

The above experiments and analysis have proved the usability of the proposed method in multiscale housing inequality evaluation. However, due to limited open access data, there are still limitations of this article which can be explored in the future.

TABLE IV
MODEL EFFICIENCY OF DIFFERENT METHODS

Model	Params (M)	FLOPs (G)
UNet	7.85	56.35
ENet	0.35	2.18
BiSeNet	90.68	166.47
Deeplab v3+	59.46	95.78
SegFormer	44.59	41.92
Segmenter	26.03	38.45
DDRNet	32.30	35.15
PIDNet	37.30	34.44
MaskUV	41.63	49.81

Specifically, though the experiment has been conducted based on the PRD in China, the usability of the method in other regions has not been tested for the time being. However, since the proposed method only relies on public data, the model is cheap enough to rebuild, validate, and apply for different regions. In addition, due to the poor availability of high-resolution satellite imagery and social media data of long time series, it is still hard to explore the spatial distribution of UV over time. Therefore, future article can be attempted from the following aspects.

1) Applying the UV model to a wider range of regions for better generalizability. In the future, a larger UV dataset could be created to provide ample training data for the UV classification model.
2) Designing a multimodal fusion model to combine high-resolution imagery, street view imagery, and POI data, thereby harnessing the fine-grained visual details from multiple perspectives and socioeconomic attributes of UV.
3) Combining long-time series images with medium spatial resolution to comprehensively investigate the spatiotemporal patterns of UV.

VII. CONCLUSION

This article identifies the UV of seven cities in the PRD in 2020 using high-resolution remote sensing images and analyzes the spatial pattern and accessibility of UV. To address the challenges of UV detection in the rapidly urbanized PRD, this article proposes MaskUV and a new high-resolution dataset (UVSet). The MaskUV integrates the strengths of the mask transformer, enabling the extraction of local texture features and global features. The MTM with mask attention is designed to aggregate features from different levels, and the mask classification enhances feature learning by introducing object-level features through the prediction of different masks.

The experimental results on seven cities in PRD and UVSet manifest that the proposed MaskUV achieves promising performance in UV classification. The ablation study further verifies the effectiveness of the random token masking strategy. In terms of model complexity and computational efficiency, the MaskUV exhibits advantages in memory and computational complexity while maintaining performance, as evidenced by the comparison of FLOPs and Params. Overall, this article suggests that the proposed MaskUV can fully exploit multiscale features from high-resolution remote sensing images to achieve fine-grained UV classification.

REFERENCES

[1] Z. Wei, Y. Liu, S. He, and H. Mo, "Housing differentiation in transitional urban China," *Cities*, vol. 96, Jan. 2020, Art. no. 102469, doi: 10.1016/j.cities.2019.102469.

[2] S. M. Reia, P. S. C. Rao, M. Barthelemy, and S. V. Ukkusuri, "Spatial structure of city population growth," *Nature Commun.*, vol. 13, no. 1, Oct. 2022, Art. no. 5931, doi: 10.1038/s41467-022-33527-y.

[3] M. Alberti, "Eco-evolutionary dynamics in an urbanizing planet," *Trends Ecol. Evol.*, vol. 30, no. 2, pp. 114–126, Feb. 2015, doi: 10.1016/j.tree.2014.11.007.

[4] Y. Liu, S. He, F. Wu, and C. Webster, "Urban villages under China's rapid urbanization: Unregulated assets and transitional neighbourhoods," *Habitat Int.*, vol. 34, no. 2, pp. 135–144, Apr. 2010, doi: 10.1016/j.habitatint.2009.08.003.

[5] Z. Tao et al., "Dying villages to prosperous villages: A perspective from revitalization of idle rural residential land (IRRL)," *J. Rural Stud.*, vol. 84, pp. 45–54, May 2021, doi: 10.1016/j.jrurstud.2021.02.010.

[6] C. Wang, "Urban village on a global scale: Diverse interpretations of one label," *Urban Geogr.*, vol. 43, no. 2, pp. 184–205, Feb. 2022, doi: 10.1080/02723638.2020.1842097.

[7] R. Fan, J. Li, F. Li, W. Han, and L. Wang, "Multilevel spatial-channel feature fusion network for urban village classification by fusing satellite and streetview images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5630813, doi: 10.1109/TGRS.2022.3208166.

[8] "World cities report 2022: Envisaging the future of cities," 2022, [Online]. Available: https://digitallibrary.un.org/record/3984713/files/9789210054386.pdf

[9] A. Ghasempour, "Informal settlement; concept, challenges and intervention approaches," *Specialty J. Architecture Construction*, vol. 1, no. 3, pp. 10–16, 2015.

[10] H. Ren, W. Wu, T. Li, and Z. Yang, "Urban villages as transfer stations for dengue fever epidemic: A case study in the Guangzhou, China," *PLoS Neglected Trop. Dis.*, vol. 13, no. 4, Apr. 2019, Art. no. e0007350, doi: 10.1371/journal.pntd.0007350.

[11] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of China using high-resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3639–3657, Jul. 2015, doi: 10.1109/TGRS.2014.2380779.

[12] K. Zhao, Y. Liu, S. Hao, S. Lu, H. Liu, and L. Zhou, "Bounding boxes are all we need: Street view image classification via context encoding of detected buildings," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602817, doi: 10.1109/TGRS.2021.3064316.

[13] C. Garriga, A. Hedlund, Y. Tang, and P. Wang, "Rural-urban migration, structural transformation, and housing markets in China," *Amer. Econ. J.: Macroecon.*, vol. 15, no. 2, pp. 413–440, Apr. 2023, doi: 10.1257/mac.20160142.

[14] L. Liang, M. Chen, and D. Lu, "Revisiting the relationship between urbanization and economic development in China since the reform and opening-up," *Chin. Geogr. Sci.*, vol. 32, no. 1, pp. 1–15, Feb. 2022, doi: 10.1007/s11769-022-1255-7.

[15] X. Guan, H. Wei, S. Lu, Q. Dai, and H. Su, "Assessment on the urbanization strategy in China: Achievements, challenges and reflections," *Habitat Int.*, vol. 71, pp. 97–109, Jan. 2018, doi: 10.1016/j.habitatint.2017.11.009.

[16] W. Ma, G. Jiang, T. Zhou, and R. Zhang, "Mixed land uses and community decline: Opportunities and challenges for mitigating residential vacancy in peri-urban villages of China," *Front. Environ. Sci.*, vol. 10, Apr. 2022, Art. no. 887988, doi: 10.3389/fenvs.2022.887988.

[17] D. Chen et al., "A hierarchical approach for fine-grained urban villages recognition fusing remote and social sensing data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 106, Feb. 2022, Art. no. 102661, doi: 10.1016/j.jag.2021.102661.

[18] Z. Li and F. Wu, "Residential satisfaction in China's informal settlements: A case study of Beijing, Shanghai, and Guangzhou," *Urban Geogr.*, vol. 34, no. 7, pp. 923–949, Nov. 2013, doi: 10.1080/02723638.2013.778694.

[19] P. Hao, P. Hooimeijer, R. Sliuzas, and S. Geertman, "What drives the spatial development of urban villages in China?," *Urban Stud.*, vol. 50, no. 16, pp. 3394–3411, Dec. 2013, doi: 10.1177/0042098013484534.

[20] L. Jiang, Y. Lai, K. Chen, and X. Tang, "What drives urban village redevelopment in China? A survey of literature based on web of science core collection database," *Land*, vol. 11, no. 4, Apr. 2022, Art. no. 525, doi: 10.3390/land11040525.

[21] L. Lingling, "Urban villages as spaces of cultural identity: Urban migrant writers in the Pearl River Delta," *Int. J. China Stud.*, vol. 4, no. 2, 2013, Art. no. 189.

[22] H. Song, M. Pan, and Y. Chen, "Nightlife and public spaces in urban villages: A case study of the Pearl River Delta in China," *Habitat Int.*, vol. 57, pp. 187–204, Oct. 2016, doi: 10.1016/j.habitatint.2016.07.009.

[23] W. Wu, H. Ren, M. Yu, and Z. Wang, "Distinct influences of urban villages on urban heat islands: A case study in the Pearl River Delta, China," *Int. J. Environ. Res. Public Health*, vol. 15, no. 8, Aug. 2018, Art. no. 1666, doi: 10.3390/ijerph15081666.

[24] X. Yang, "Determinants of migration intentions in Hubei Province, China: Individual versus family migration," *Environ. Plan. A: Econ. Space*, vol. 32, no. 5, pp. 769–787, May 2000, doi: 10.1068/a32114.

[25] N. Mboga, C. Persello, J. Bergado, and A. Stein, "Detection of informal settlements from VHR images using convolutional neural networks," *Remote Sens.*, vol. 9, no. 11, Oct. 2017, Art. no. 1106, doi: 10.3390/rs9111106.

[26] J. Mast, C. Wei, and M. Wurm, "Mapping urban villages using fully convolutional neural networks," *Remote Sens. Lett.*, vol. 11, no. 7, pp. 630–639, Jul. 2020, doi: 10.1080/2150704X.2020.1746857.

[27] Z. Pan, J. Xu, Y. Guo, Y. Hu, and G. 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, doi: 10.3390/rs12101574.

[28] C. 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, doi: 10.1080/01431161.2023.2288948.

[29] S. Niebergall, A. Loew, and W. Mauser, "Integrative assessment of informal settlements using VHR remote sensing data—The Delhi case study," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 3, pp. 193–205, Sep. 2008, doi: 10.1109/JSTARS.2008.2007513.

[30] K. K. Owen and D. W. Wong, "An approach to differentiate informal settlements using spectral, texture, geomorphology and road accessibility metrics," *Appl. Geogr.*, vol. 38, pp. 107–118, Mar. 2013, doi: 10.1016/j.apgeog.2012.11.016.

[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[32] Q. Zhu et al., "A global context-aware and batch-independent network for road extraction from VHR satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 353–365, May 2021, doi: 10.1016/j.isprsjprs.2021.03.016.

[33] X. Yang et al., "An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 238–262, Jul. 2021, doi: 10.1016/j.isprsjprs.2021.05.004.

[34] Q. Feng et al., "Mapping of plastic greenhouses and mulching films from very high resolution remote sensing imagery based on a dilated and non-local convolutional neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102441, doi: 10.1016/j.jag.2021.102441.

[35] Q. Feng et al., "Integrating multitemporal sentinel-1/2 data for coastal land cover classification using a multibranch convolutional neural network: A case of the Yellow River Delta," *Remote Sens.*, vol. 11, no. 9, Apr. 2019, Art. no. 1006, doi: 10.3390/rs11091006.

[36] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019, doi: 10.1016/j.isprsjprs.2019.04.015.

[37] W. Han et al., "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities," *ISPRS J. Photogrammetry Remote Sens.*, vol. 202, pp. 87–113, Aug. 2023, doi: 10.1016/j.isprsjprs.2023.05.032.

[38] J. Guo et al., "GluonCV and gluon NLP: Deep learning in computer vision and natural language processing," *J. Mach. Learn. Res.*, vol. 21, pp. 845–851, Feb. 2020.

[39] 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, doi: 10.1109/LGRS.2017.2763738.

[40] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, Apr. 2019, doi: 10.1016/j.isprsjprs.2019.02.006.

[41] Q. Shi et al., "Domain adaption for fine-grained urban village extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020, doi: 10.1109/LGRS.2019.2947473.

[42] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[43] L. Wang, R. Li, D. Wang, C. Duan, T. Wang, and X. Meng, "Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images," *Remote Sens.*, vol. 13, no. 16, Aug. 2021, Art. no. 3065, doi: 10.3390/rs13163065.

[44] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5615611, doi: 10.1109/TGRS.2021.3136190.

[45] Z. Zhong, Y. Li, L. Ma, J. Li, and W.-S. Zheng, "Spectral–spatial transformer network for hyperspectral image classification: A factorized architecture search framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5514715, doi: 10.1109/TGRS.2021.3115699.

[46] A. Crivellari, H. Wei, C. Wei, and Y. Shi, "Super-resolution GANs for upscaling unplanned urban settlements from remote sensing satellite imagery—The case of Chinese urban village detection," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 2623–2643, Dec. 2023, doi: 10.1080/17538947.2023.2230956.

[47] H. Guo, Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding, "Scene-driven multitask parallel attention network for building extraction in high-resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4287–4306, May 2021, doi: 10.1109/TGRS.2020.3014312.

[48] B. Chen et al., "Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, May 2022, Art. no. 102794, doi: 10.1016/j.jag.2022.102794.

[49] L. Zhang, Q. Weng, and Z. Shao, "An evaluation of monthly impervious surface dynamics by fusing Landsat and MODIS time series in the Pearl River Delta, China, from 2000 to 2015," *Remote Sens. Environ.*, vol. 201, pp. 99–114, Nov. 2017, doi: 10.1016/j.rse.2017.08.036.

[50] L. Hu, J. Yang, T. Yang, Y. Tu, and J. Zhu, "Urban spatial structure and travel in China," *J. Plan. Literature*, vol. 35, no. 1, pp. 6–24, Feb. 2020, doi: 10.1177/0885412219853259.

[51] "Guangdong statistics yearbook," Guangdong Provincial Bureau of Statistics, 2019. [Online]. Available: http://stats.gd.gov.cn/gdtjnj/index.html

[52] X. Li et al., "Mapping global urban boundaries from the global artificial impervious area (GAIA) data," *Environ. Res. Lett.*, vol. 15, no. 9, Sep. 2020, Art. no. 094044, doi: 10.1088/1748-9326/ab9be3.

[53] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[54] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," Jun. 2023, Accessed: May 16, 2024. [Online]. Available: http://arxiv.org/abs/1606.08415

[55] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," Jul. 2016, Accessed: May 16, 2024. [Online]. Available: http://arxiv.org/abs/1607.06450

[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision*, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[57] Z. Sun, S. Cao, Y. Yang, and K. Kitani, "Rethinking transformer-based set prediction for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3591–3600, doi: 10.1109/ICCV48922.2021.00359.

[58] P. Gao, M. Zheng, X. Wang, J. Dai, and H. Li, "Fast convergence of DETR with spatially modulated co-attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 3601–3610, doi: 10.1109/ICCV48922.2021.00360.

[59] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. (Lecture Notes in Computer Science). Berlin, Germany: Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[60] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," Jun. 2016, Accessed: Dec. 30, 2023. [Online]. Available: http://arxiv.org/abs/1606.02147

[61] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 334–349.

[62] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," Aug. 2018, Accessed: Dec. 30, 2023. [Online]. Available: http://arxiv.org/abs/1802.02611

[63] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Neural Inf. Process. Syst.*, Oct. 2021, pp. 12077–12090.

[64] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 7242–7252, doi: 10.1109/ICCV48922.2021.00717.

[65] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023, doi: 10.1109/TITS.2022.3228042.

[66] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 19529–19539, doi: 10.1109/CVPR52729.2023.01871.

[67] S. Węglarczyk, "Kernel density estimation and its application," in *Proc. ITM Web Conf.*, 2018, Art. no. 00037, doi: 10.1051/itmconf/20182300037.

Zhuoqun Chai is currently working toward the M.S. degree in geographic information science with the School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China.

His research interests include environmental spatial analysis, change detection, data fusion, and deep learning.

Mengxi Liu received the B.S. degree in geographic information science and the Ph.D. degree in cartography and geographic information system from the Sun Yat-Sen University, Guangzhou, China, in 2019 and 2024, respectively.

She is currently a Postdoctoral Researcher with the School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China. Her research interests include intelligent understanding of remote sensing images, change detection, and domain adaptation.

Qian Shi (Senior Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing from the Wuhan University, Wuhan, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, in 2015.

She is currently a Professor with the School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China. Her research interests include remote sensing image classification, including deep learning, active learning, and transfer learning.

Yuanyuan Zhang received the B.S. degree in geographic information science in 2024 from the Sun Yat-Sen University, Guangzhou, China, where she is currently working toward the M.S. degree in environmental engineering.

Her research interests include environmental monitoring using remote sensing and environmental remediation.

Minglin Zuo received the B.S. degree in geographic information science from the Sun Yat-Sen University, Guangzhou, China, in 2024.

Her research interests include intelligent understanding of remote sensing images and spatial analysis.

Da He (Member, IEEE) received the B.S. degree in remote sensing science and technology and the Ph.D. degree in photogrammetry and remote sensing from the Wuhan University, Wuhan, China, in 2015 and 2020, respectively.

He is currently an Associate Professor with the School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China. His research interests include multi and hyperspectral remote sensing image classification, deep learning, data fusion, and change detection.