A Cross-Angle Propagation Network for Built-Up Area Extraction by Fusing Spatial–Spectral-Angular Features From the ZY-3 Multiview Satellite Imagery: Dataset and Analysis of China's 41 Major Cities

Renxiang Zuo, Xin Huang, *Senior Member, IEEE*, Jiayi Li, *Senior Member, IEEE*, and Xiaofeng Pan

***Abstract*— Obtaining timely and reliable built-up area (BUA) information across extensive geographical zones holds crucial significance for understanding environmental change and human activities. BUAs often exhibit detailed textures and structures in high-resolution imagery but also present strong heterogeneity. Current methods for BUA extraction primarily relied on planar information from single-view imagery, struggling to effectively capture the 3-D attributes of urban landscapes. Therefore, to address this challenge, this article proposes a cross-angle propagation network (CAPNet) based on multiview remote sensing stereo observation imagery. Our contributions are threefold: 1) we propose the cross-angle fusion module (CAFM) to exploit BUA's complementary spatial–spectral-angular context across different viewing angles. This module leverages attention mechanisms for the automated acquisition of multiangle feature representation learning from diverse angle combinations. 2) We propose a multiangular propagation decoder (MAPD) that pioneers the exploration of gradually propagating multiangle disparity information through bidirectional-adjacent feature fusion across hierarchical levels. 3) We construct a large-scale, high-resolution multiview BUA (MVBA) dataset over China's 41 major cities based on the ZY-3 satellites. Extensive experiment results on MVBA and the public WV-3 multiview semantic stereo datasets verify CAPNet's superiority to existing state-of-the-art (SOTA) models, on preserving overall BUA shape, edge, and internal structures. The dataset and the source code of CAPNet will be publicly available at https://github.com/zuo-ux/Cross-Angle-Propagation-Network.**

***Index Terms*— Attention mechanism, built-up area (BUA), dataset, high resolution, multiview.**

I. Introduction

ACCORDING to the latest statistics from the United Nations, the urban population is projected to reach 68% of the global population by 2050 [1]. The urbanization process leads to an expansion of cities in both the horizontal and vertical dimensions. On one hand, cities encroach upon other types of land resources as they expand spatially. On the other hand, they exhibit vertical growth with the emergence of increasingly tall buildings, resulting in a more intricate and diverse vertical structure of urban features. Cities significantly alter the climate, ecological environments, regional and global water cycles, and gas circulation. According to existing literature [2], [3], [4], [5], [6], [7], in this study, built-up area (BUA) is defined as the spatial extent covered by continuous building structures, but excluding main roads, parks, and large open spaces. To better understand human activities and their interactions with the environment, it is paramount to timely and reliably acquire spatial distribution and extent of BUA. This information is critical in numerous domains, including urban development and planning, national land dynamics monitoring and assessment, ecological environments, climate change, public safety, and sustainable development [8], [9].

However, existing research primarily relies on medium to low-resolution imagery for large-scale BUA extraction, such as MODIS 1 km Map of Global Urban Extent (MOD1 K, 927 m) [10], Global Human Settlement Layer (GHSL-Landsat, 20∼38 m) [11], [12], Global Impervious Surface Area (GISA, 30 m) [13], Global Urban Footprint (GUF, 12 m) [3], and GISA-10 m [14]. These coarse-resolution products may fall short of meeting the need for precise information regarding the distribution of BUA. Moreover, the accuracy of these BUA products often varies significantly across different regions. For instance, GUF derived from synthetic aperture radar (SAR) data presents challenges in distinguishing between buildings, trees, and elevated bridges, often resulting in sparse and low-rise BUA omissions. GUF is also susceptible to coherence noise, especially in complex urban environments. Similarly, GHSL may encounter difficulties in distinguishing roads from bare ground and overlooking residential areas with lower brightness [5]. Therefore, conducting further research into high-quality and high-resolution BUA extraction is necessary.

High-resolution imagery can provide clearer details of spatial textures and geometric structures, visually representing

Manuscript received 22 March 2024; revised 5 August 2024; accepted 29 August 2024. Date of publication 3 September 2024; date of current version 19 September 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903701 and in part by the National Natural Science Foundation of China under Grant 42271328 and Grant 42071311. *(Corresponding author: Xin Huang.)*

Renxiang Zuo, Xin Huang, and Jiayi Li are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: zrxshimilywhu@whu.edu.cn; xhuang@whu.edu.cn; zjjercia@whu.edu.cn).

Xiaofeng Pan is with Shenzhen Ecological and Environmental Monitoring Center of Guangdong Province, Shenzhen 518049, China.

Digital Object Identifier 10.1109/TGRS.2024.3453868

the spatial relationships between target and neighboring elements. This is beneficial for distinguishing different land cover types within urban areas, facilitating accurate extraction of BUA, and potentially mitigating challenges associated with mixed pixels in medium to low-resolution imagery. Nevertheless, although higher spatial resolution allows for more detailed information acquisition, it accentuates challenges such as high intraclass heterogeneity and low interclass homogeneity. Moreover, considering the diversity of land cover categories within BUA and the complexity of the neighborhoods, achieving accurate large-scale BUA extraction still presents numerous challenges. To be specific, influenced by historical, cultural, environmental, and economic development levels, BUA in different regions also exhibits distinct regional characteristics. Furthermore, due to their similar planar spatial characteristics, artificial features exhibit lower separability, making it difficult to accurately distinguish complex urban artificial land cover categories solely based on single-view remote sensing imagery. This issue can be addressed by considering 3-D spatial information [15]. For example, point cloud data acquired through SAR and light detection and ranging (LiDAR) contain 3-D information. However, SAR data is susceptible to coherent noise in complex urban environments, and LiDAR data's high costs limit their applicability to large-scale urban areas. Compared to SAR and LiDAR data, utilizing multiview high-resolution (MVHR) imagery to capture 3-D spatial information offers a series of advantages, e.g., efficiency, timeliness, cost-effectiveness, and broad observation coverage.

BUA primarily consists of clustered structures that rise above the Earth's surface, displaying diverse spatial distributions in high-resolution imagery when observed from various viewing angles. For instance, the roofs of buildings are prominently visible in nadir (NAD)-view imagery, while their sides are typically captured in forward (FWD)- and backward (BWD)-view imagery. Consequently, multiview information has the potential for differentiating BUA from other land cover types, resulting in more accurate and detailed urban scene interpretation [16].

In recent years, some high-resolution optical satellites have been capable of acquiring MVHR imagery through observation modes such as constellations, co-orbiting, or cross-orbiting configurations. For instance, the ZiYuan3 (ZY3) constellation, consisting of ZY3-01, ZY3-02, and ZY3-03 satellites, successfully launched on January 9, 2012, May 30, 2016, and July 25, 2020, respectively. ZY3 are China's first civilian high-resolution stereo mapping satellites designed to acquire stereo images along its orbital path. ZY3 satellites employ the three-line-array charge-coupled device (CCD) equipped with multispectral and three panchromatic cameras. It captures nearly simultaneous imagery of the same area from multiple angles, including the NAD-view multispectral imagery, FWD-, NAD-, and BWD-view panchromatic imagery, all at fixed observation angles ($\pm 22^{\circ}$). As illustrated in Fig. 1, the NAD, FWD, and BWD cameras are stitched by three or four pieces of the time-delay-integration CCD sensors. The detailed specifications parameters of the ZY3 satellites are listed in Table I.

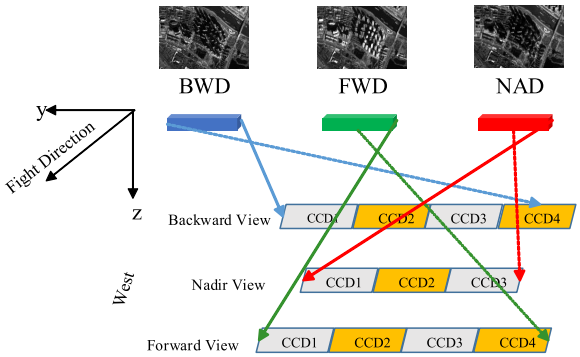

Fig. 1. Relationship between the panchromatic camera focal and data output of the ZY3-03.

TABLE I
DETAILED SPECIFICATIONS PARAMETERS OF THE ZY3 SATELLITES

Specification	Value	
Orbit parameters		
Launch Site	Taiyuan Satellite Launch Center, China	
Average Orbit Altitude	505.983 km	
Orbit Period/ Orbital type	97.7 min/ Sun-Synchronous	
Orbit Inclination/ Time	97.421° /10:30 am	
Revisit cycle	5 days	
Cycle Duration	59 days	
Detailed payload information		
Payload information	Panchromatic Camera	Multispectral Camera
Spectral range	0.5∼0.8 μm	Blue:0.45∼0.52 μm Green:0.52∼0.59 μm Red:0.63∼0.69 μm NIR:0.77∼0.89 μm
Ground spatial distance	NAD: 2.1 m FWD, BWD (01): 3.5 m FWD, BWD (02,03): 2.5 m	5.8 m
Focal length	1700 mm	1750 mm
CCD array information	NAD: 24576× 7 μm FWD, BWD: 16384×10 μm	9216×20 μm
Inclinations from nadir	FWD: $+23.5^{\circ}$ BWD: -23.5°	$+6^{\circ}$
Swath width	50 km	52 km

The MVHR imagery is acquired with minimal time intervals, ensuring that the observed area's land use, atmospheric, and illumination conditions remain unchanged. In this scenario, the variations in grayscale values in the MVHR imagery primarily result from different viewing angles. Moreover, suitable tilt angles are beneficial for detecting artificial features such as buildings. There have been a few studies utilizing MVHR imagery for urban area extraction [17], building extraction [18], and building height estimation [19].

As illustrated in Fig. 2, vertical features, such as staggered high-rise buildings, exhibit distinct angular variations under varying observation angles due to the side views, surface anisotropy, and shadows [20]. In contrast, ground-level objects, such as bare land, maintain higher consistency in different viewing angles. Based on this phenomenon, the fundamental idea of the proposed cross-angle fusion module (CAFM) is able to emphasize the angular variation manifestations of

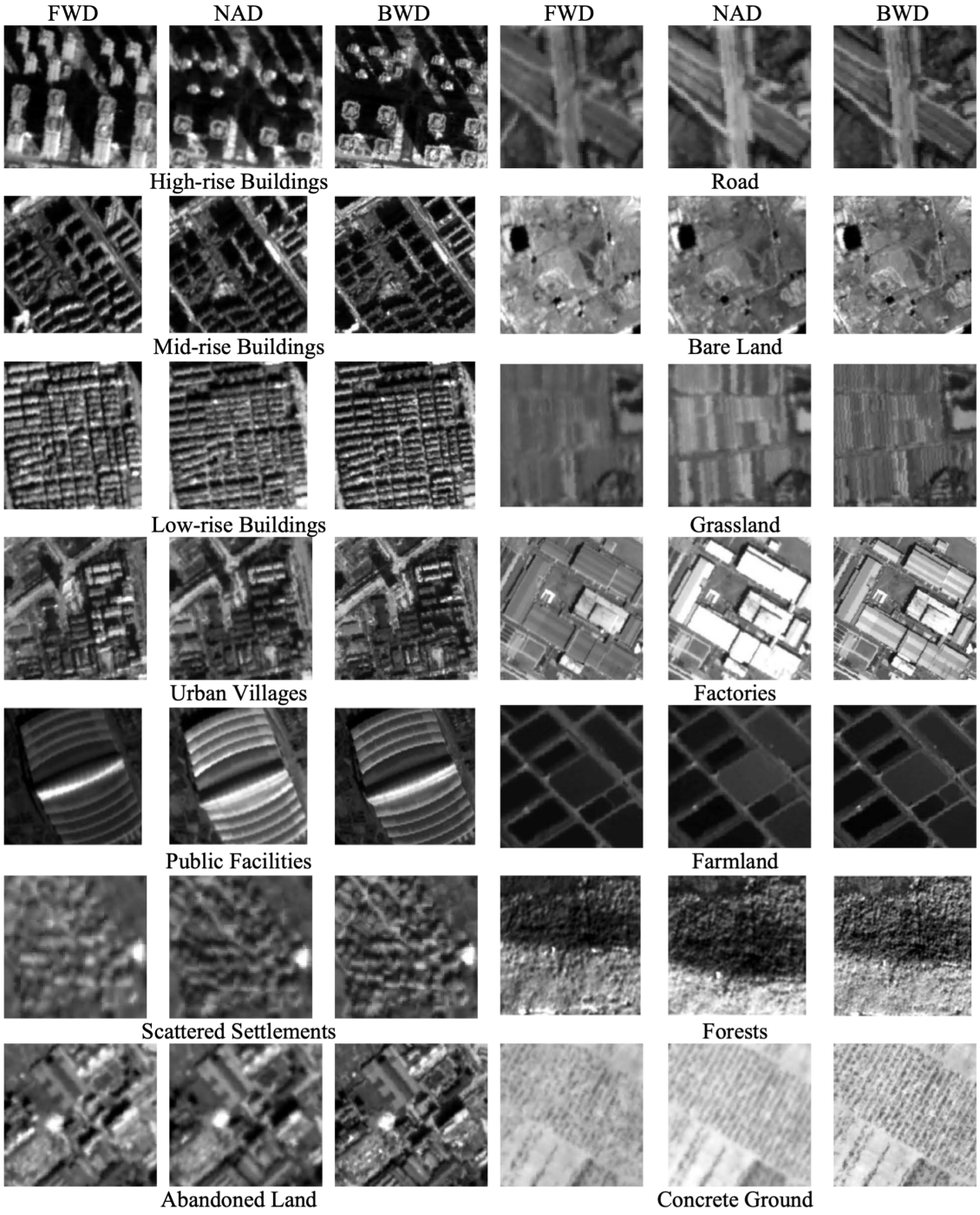

Fig. 2. Angular variations manifestations of typical urban features in MVHR imagery.

BUA by describing the angular differences between MVHR imagery while simultaneously suppressing the influence of other surface features.

BUA extraction from remote sensing imagery has been extensively studied, including manually designed features [21], [22], [23], [24] and deep learning (DL) networks [25]. The readers can refer to Section II-B for a detailed review. Nevertheless, in summary, there exist limitations and challenges that require attention: 1) most current research on BUA predominantly focuses on the 2-D features provided by single-view perspectives, failing to characterize the structure of objects and thus struggling to achieve accurate discrimination among them; 2) the potential of MVHR imagery to represent and characterize urban vertical structures has not been fully explored; deep fusion of spatial–spectral-angular information is lacking; and 3) existing datasets related to BUA suffer from inadequate coverage and sample diversity, and multiview BUA (MVBA) datasets and samples are scarce.

TABLE II
COMPREHENSIVE OVERVIEW OF THE RELATED DATASETS REGARDING BUA

Dataset name	Release Time	Spatial Resolution (m)	Pixels	Patches	Coverage (km^2)
ISPRS Vaihingen [1]	2012	0.09	2494×2064	33	11
ISPRS Potsdam [2]	2012	0.05	6000×6000	38	2.16
Massachusetts [30]	2013	1	1500×1500	151	2.25
Aerial Image Segmentation Dataset [31]	2013	0.3∼1.0	512×512	80	157.7
AIRS [32]	2015	0.075	10000×100004	1047	457
Inria Aerial Image Labeling Dataset [33]	2017	0.3	1500×1500	180	405
Digital Globe [34]	2018	0.5	2048×2048	803	1200
GID [35]	2018	0.8∼1.0	7200×6800	150	∼50000
SpaceNet Buildings Dataset [36]	2018	0.5∼1.0	200×200	12980	2544
DLRSD [37]	2018	1.0	256×256	2100	
WHU Building Dataset [38]	2018	0.075/2.7	512×512	8189/17388	450/550
EvLab-SS [39]	2018	0.1∼0.25	4500×4500	60	
DFC19-JAX [40]	2019	0.3	2048×2048	3083	$\geq$ 1160
LandCoverNet [41]	2020	10.0	256×256	1980	100
2020 CCF BDC [3]	2020	2.0	256×256	140000	
Hi-UCD [42]	2020	0.3	512×512	2500	30
SECOND [43]	2020	0.3	512×512	4662	
CrowdAI Mapping Dataset [33]	2020	0.03∼2	300×300	341058	
Examples of Typical Urban Buildings in China [45]	2021	0.29	500×500	120	11
LoveDa [46]	2022	0.3	512×512	5987	536.15
SARBud1.0 [1]	2022	10	256×256	140000	11
MVBA (Ours)	**2024**	**2.5**	**512 × 512**	**54000**	**55000**

To deal with these important issues, the main research content and contributions of this study are summarized as follows.

1) Considering the complementarity of 2-D spatial and angular information, we propose the CAFM based on MVHR imagery. This module can learn complementary spatial–spectral-angular information to comprehensively represent BUA.
2) In this study, we introduce the multiangular propagation decoder (MAPD), which progressively and sequentially propagates multiangle disparity information through bidirectional-adjacent fusion. This approach significantly enhances the capability to capture multiscale and multiangle variations from urban scenes.
3) Considering the lack of MVBA datasets, we utilize the ZY-3 multispectral and multiview imagery to construct the large-scale MVBA dataset. This dataset covers 41 representative cities in China with an area of 55 000 km^2, which can effectively represent various scenarios and distribution types of BUA, thereby advancing research in large-scale BUA extraction.

The remaining sections of this article are organized as follows. Section II provides an overview related to BUA extraction, including BUA extraction methodologies, MVHR feature representation, and existing datasets regarding BUA. Section III describes the production of the MVBA dataset. Section IV introduces the proposed cross-angle propagation network (CAPNet) for integrating spatial–spectral-angular information. Section V conducts a series of comparative experiments between CAPNet and state-of-the-art (SOTA) methods. Section VI aims to evaluate the benefits of CAFM and MAPD while discussing this study's limitations and potential future research directions. Section VII concludes the work.

[1]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx

[2]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[3]https://www.datafountain.cn/special/BDCI2020

II. RELATED WORK

A. BUA Datasets

The performance of DL-based BUA extraction is heavily reliant on the quality of samples. Therefore, constructing a high-quality dataset that accurately reflects the characteristics of BUA is indispensable [26], [27]. As shown in Table II, current studies have focused on local regions such as cities, countries, and continents to construct a series of datasets relevant to BUA extraction. According to the existing literature, the primary data sources for urban area mapping products include aerial imagery, high-resolution satellite remote sensing imagery, LiDAR point cloud data [28], and nighttime remote sensing imagery [29].

From Table II, it can be seen that the current datasets are primarily concentrated within relatively small geographical areas, which are insufficient to meet the requirements of large-scale BUA extraction. Particularly, currently, MVHR BUA datasets are lacking and are rarely investigated. To represent the distribution and morphology of BUA more comprehensively, this study integrates multispectral imagery with MVHR imagery to construct a large-scale MVBA dataset. It is evident that our constructed MVBA dataset exhibits advantages in extensive coverage, diverse distribution of BUA types, rich data diversity (including multiangle and multispectral information), and a large volume of samples. Readers can refer to III for the production of the dataset.

B. BUA Extraction

In complex urban environments, relying solely on spectral features often fails to yield satisfactory results [47]. Therefore, a substantial amount of research has been dedicated to designing spatial features to complement the limitations of spectral features, including texture [22], morphological [23], and shape features [24]. While handcrafted feature representations have succeeded in some classification tasks, they often fall short in comprehensively describing the complex characteristics of urban areas.

With the rapid advancement of DL, the application of convolutional neural networks (CNNs) has become increasingly widespread, showcasing enhanced feature representation capabilities across various computer vision applications. Compared to traditional methods, CNNs can automatically learn a substantial amount of high-level semantic features from raw images. For instance, the fully convolutional network (FCN) [48] advanced the image semantic segmentation by replacing fully connected layers with fully convolutional layers and employing deconvolutional structures to restore feature map dimensions, allowing the segmentation of images of arbitrary sizes. However, FCN overlooks global contextual information, which hinders the network from comprehending complex scenes. This results in the neglect of fine-grained details and issues such as unclear object boundaries. Subsequent to FCN, a series of segmentation models based on CNNs have emerged, e.g., SegNet [49], HRNet [50], U-Net [51], DeeplabV3 [52], UperNet [53], FPN [54], and ResUNet-a [55].

While CNNs offer significant advantages in feature extraction, they inherently suffer from translation invariance, information locality, and other inductive biases that constrain their capacity to capture features across a broad range. Introducing self-attention mechanisms [56], [57] can address these issues to some extent. The idea behind self-attention mechanisms is to select relevant information by capturing the relationships between any two positions in the feature maps. For instance, ISANet [58] extracts semantic information at different scales by introducing the decomposition of sparse affinity matrices and multistage attention modules. The transformer model [59] can utilize self-attention mechanisms to control the input–output interactions. For instance, PoolFormer [60] abstracts the transformer into a MetaFormer with a generalized architecture while employing a nonparametric pooling operation as a weak token mixer. SegFormer [61] introduces an MLP decoder that aggregates information from different layers, by combining local and global attention for powerful representation. However, the transformers-based models are subject to some limitations, such as high computational demands, slow processing time, and the need for abundant training data.

In recent years, a number of studies have attempted to utilize CNNs to extract BUA, but these efforts typically focus on specific regions. For instance, LMB-CNN [6] constructs patch-based graph models with the learned features, and then obtains BUA through postprocessing. BA-UNet [1] utilizes residual blocks in U-Net to address network degradation, and efficiently extracts BUA from high-resolution Gaofen-3 data. DSCNN [25] combines high-resolution panchromatic and multispectral imagery for automatic BUA extraction.

It is noted that many artificial urban features exhibit similar planar spatial characteristics (e.g., textures and shapes). Furthermore, urban objects with varying heights exhibit differences in brightness, structure, and spatial arrangement when observed from different viewing angles, owing to changing illumination, occlusion, and shadows.

Currently, most of the methods heavily rely on 2-D planar spatial features derived exclusively from single-view imagery, thus overlooking the benefits of angle-specific information. Therefore, there is a pressing need to explore more effective approaches for fully harnessing the combined spatial–spectral-angular information obtained from multispectral and MVHR imagery. This endeavor can help alleviate uncertainties in BUA extraction and offer a comprehensive representation of cities characterized by intricate vertical structures.

C. Feature Representation From Multiview Satellite Imagery

High-resolution remote sensing satellites equipped with stereoscopic observation modes offer the potential to extract 3-D structural information. A digital surface model (DSM) derived from MVHR imagery through stereo matching can be employed as additional data for further processing and analysis. For instance, Li et al. [62] utilized contour data from OpenStreetMap (OSM) as prior information and extracted normalized DSM (nDSM) from GeoEye stereo imagery to obtain height information. Subsequently, they applied the Bayesian theorem to fuse contour data with height information, to improve urban land-use classification accuracy. Huang et al. [63] utilized semi-global matching (SGM) to derive nDSM from ZY-3 stereo imagery, for the subsequent classifier training. However, the performance of such methods is primarily contingent on the quality of the DSM, which can be influenced by factors such as image variations, baseline-to-height ratio, and occlusion. Moreover, these approaches might overlook the inherent multiangle spatial information in stereo imagery.

An alternative approach involves converting the raw digital number (DN) value from MVHR into surface reflectance values that contain discriminative information about different objects. For instance, MV-OBIA [64] utilizes multiview information to classify multiview object instances corresponding to each orthophoto object and utilizes a voting procedure to assign a final label to the orthophoto object. Liu et al. [65] utilized the bidirectional reflectance distribution function (BRDF) model to assess the potential of utilizing multiview data in geographic object-based image analysis (GEOBIA). However, these studies treat spectral and angular variations as supplementary information, which does not fully exploit the rich multiangle spatial information.

Recent studies involve developing new spatial features to effectively leverage multiangle features. For instance, angular difference features (ADFs) [16] can comprehensively describe the differential characteristics of urban features in MVHR imagery from pixel, feature, and label levels. In addition, the ratio multiangular built-up index (RMABI) and normalized difference multiangular built-up index (NDMABI) [5] are

introduced to further enhance the characterization of building features using MVHR imagery.

The aforementioned feature representation from MVHR mainly operates at mid- or low-feature levels, lacking the capacity to capture multiangular semantic information. DL offers an opportunity to fully exploit angular information in MVHR imagery and depict the structure of urban areas. For instance, DFSN [66] employs a weight-shared ResNet-50 to extract deep semantic features from the left-view image, then processes stereo image pairs using a pretrained PSMNet [40] to obtain disparity features. These features are subsequently aggregated to generate final segmentation results.

However, it should be noted that the above-mentioned approaches simply concatenate multiangle features and subsequently input them into deep networks without fully exploiting the implicit complementary information among multiple views. To the best of our knowledge, there is currently limited research exploring spectral–spatial-angular information fusion. Therefore, it is necessary to investigate how to utilize multiangle features more effectively and unlock its potential in BUA extraction.

III. High-Resolution MVBA Dataset

This study leverages MVHR imagery to establish a large-scale, multiview, and high-resolution dataset (namely, MVBA). This dataset covers China's 41 major cities, encompassing an extensive area of 55 000 km^2 and incorporating a wide range of BUA samples. The MVBA dataset can effectively capture the diversity in geographical distribution, climatic conditions, and economic development patterns of BUA, and offer valuable insights into their unique characteristics.

In this study, we acquired BUA samples from the high-resolution land cover product Hi-ULCM [63], which is based on ZY-3 satellite imagery covering China's major cities. Hi-ULCM categorizes seven primary land cover classes: buildings, grass/shrubs, water, trees, soil, roads, and other impervious surfaces (OISAs). An accuracy assessment was conducted on over 40 000 test samples, yielding an impressive overall accuracy (OA) rate of 88.6%.

A. Generation of BUA Coverage

We propose a multiscale fusion algorithm to generate BUA samples considering their multiscale characteristics. Specifically, a series of windows with different sizes are employed, and the building density within each window is computed in a semi-overlapping manner. Subsequently, the results from different window sizes are fused to obtain the residential area intensity (RAI) value. Let R_{10}, R_{30}, R_{50}, R_{70}, and R_{100} denote the RAI value for window sizes of 10^2, 30^2, 50^2, 70^2, and 100^2, respectively, and RAI can be written as

$$\mathrm{RAI} = R_{10} + R_{30} + R_{50} + R_{70} + R_{100}. \quad (1)$$

Subsequently, the RAI value is normalized within the range [0, 1], and a threshold value is selected (in this study RAI is set to 0.1). Regions with RAI values exceeding this threshold were considered as BUA. Fig. 3(a) illustrates the relationship between BUA extraction accuracy and RAI thresholds,

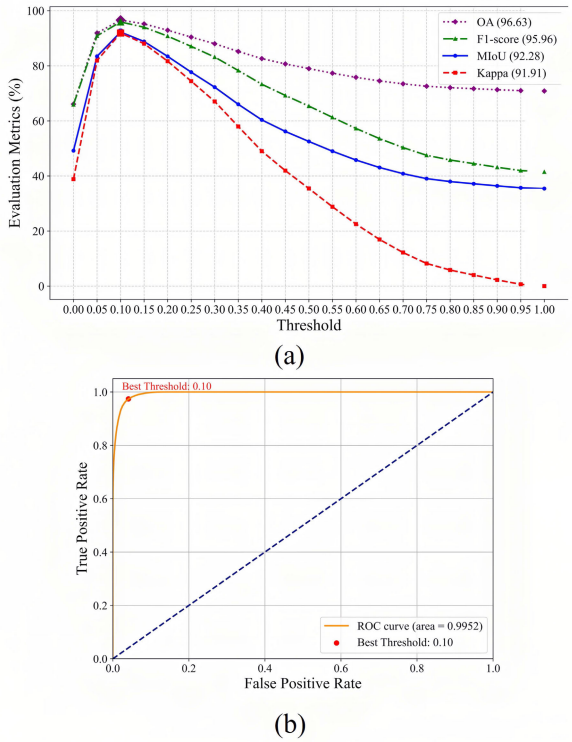

Fig. 3. (a) Relationship between the threshold and the accuracy of BUA extraction and (b) ROC curve.

with an interval of 0.05. Evaluation metrics include mean intersection over union (mIoU), macro-$F1$ (mF1), Kappa, and OA. Based on a comprehensive analysis of the accuracy trend curve [see Fig. 3(a)] and the receiver operating characteristic (ROC) curve [see Fig. 3(b)], a threshold value of 0.1 was selected. As depicted, the OA of the MVBA dataset achieves 96.63% when compared to manually annotated samples. It is worth noting that subsequent manual visual inspections were performed, which included the removal of roads, open spaces, and water, thereby reducing the influence of RAI thresholds on the samples. Finally, a visual inspection was conducted to manually rectify the errors in the initial BUA samples to ensure the accuracy of the dataset.

B. Construction of BUA Dataset

This study combined expert interpretation and a multiscale fusion algorithm to semi-automatically generate annotated binary labels corresponding to the MVHR imagery, using the NAD-view multispectral imagery as a reference. Subsequently, we cropped the MVHR imagery and the annotated binary labels to a set of patches with 512 × 512 pixels. To retain the continuity of the objects near the edges of the cropped regions and ensure that each BUA had a corresponding complete slice, we leave 25% overlapping portions between adjacent cropped areas. Fig. 4 illustrates the geographical distribution of sample cities and the extracted BUA in some cities. It is evident that the MVBA dataset can cover various scenes and geographical distributions, and provide a comprehensive assessment of the model's feature extraction performance and generalization capability. In addition, taking Wuhan as an

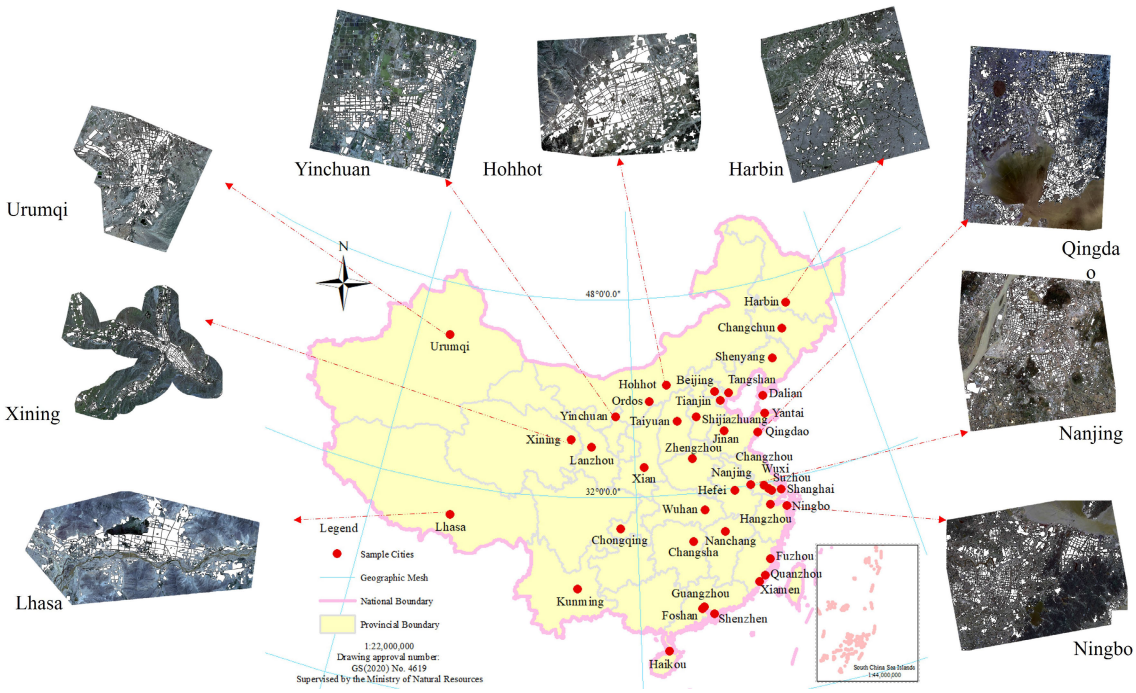

Fig. 4. Geographic distribution of sample cities and BUA examples in some sample cities.

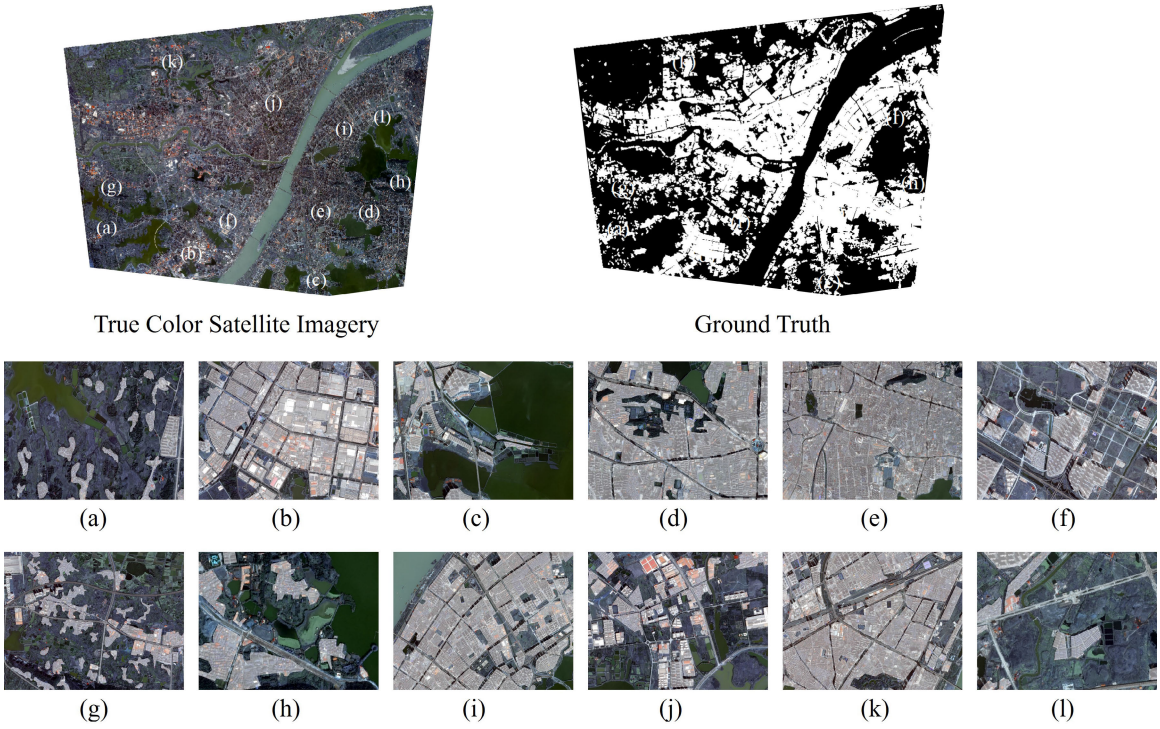

Fig. 5. Overall distribution of BUA. (a)–(l) True-color satellite imagery represents the different distribution types of BUA.

example (see Fig. 5), we observe that the overall distribution of BUA is accurate and diverse.

IV. Cross-Angle Propagation Network

The ZY-3 multiview satellite can simultaneously collect NAD-, FWD-, and BWD-view panchromatic and NAD-view multispectral imagery. Significant variations often exist between MVHR imagery of the same area due to multiple observational angles. Therefore, joint consideration of spatial–spectral-angular information is essential to enhance the discriminability of urban scenes. This study proposes the CAPNet that fuses multispectral and MVHR imagery for BUA extraction. As depicted in Fig. 6, CAPNet consists of four main components: the multiinput stream encoder (HorNet), CAFM, MAPD, and logical classification layer.

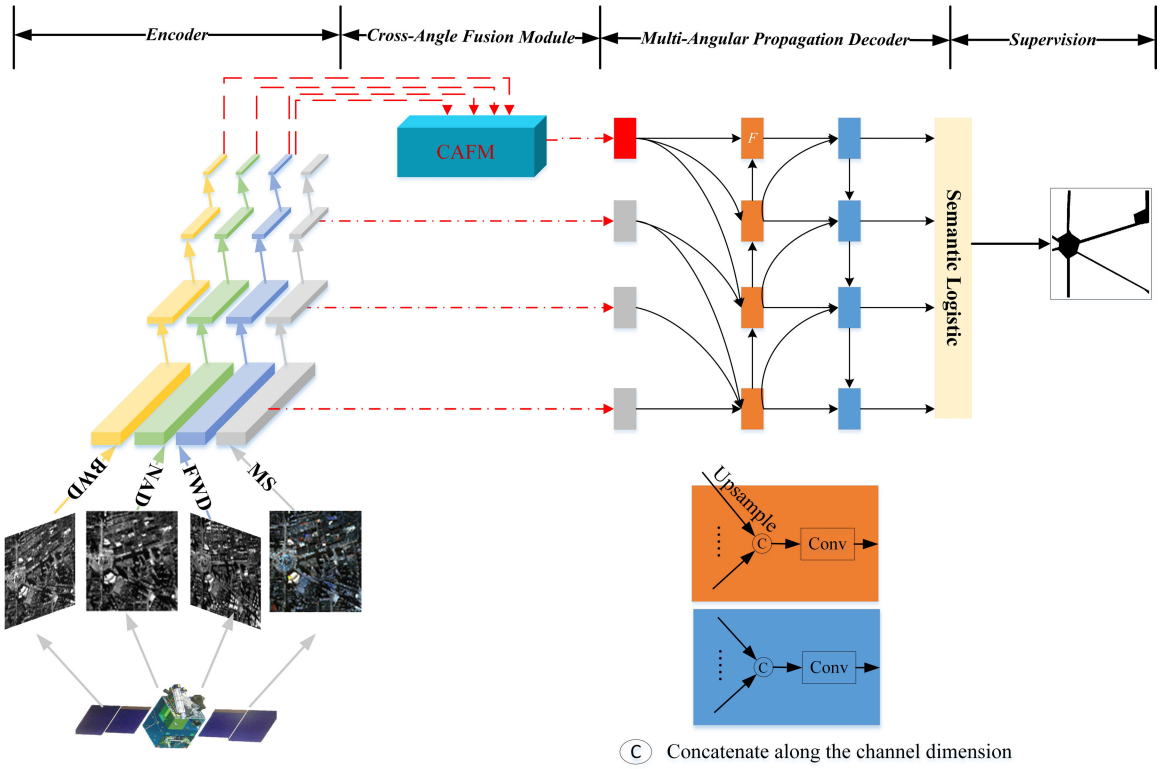

Fig. 6. Overall workflow of CAPNet for BUA extraction. Note that the "MS" represents the multispectral imagery.

Since the NAD-, FWD-, and BWD-view imagery are captured nearly simultaneously, considering their similarity and complementarity, in this study, we separately encoded the multiview features that reflect urban vertical structure to better preserve the original information from different viewpoints. This encoding method also facilitates more flexible adjustment of fusion methods to accommodate various scenarios and requirements. CAPNet can improve its ability to learn spatial–spectral-angular features from MVHR imagery by sharing all encoder parameters. This approach can significantly boost the efficiency of network training. Furthermore, in the final stage of the encoder, the CAFM is deployed to proficiently capture angle feature representations across MVHR imagery. Then, MAPD is designed to fuse multilevel and angle feature representations. Finally, the BUA extraction is accomplished through the logical classification layer.

A. Feature Extractor

The HorNet feature extractor [67] introduces recursive gated convolutions (gnConvs) for efficient and scalable input-adaptive, long-range, and high-order spatial interactions in a coarse-to-fine manner. The highly flexible and customizable operation gnConv is introduced as follows.

For the input feature $\boldsymbol{x} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W are the channel dimension, height, and width of the input feature, respectively, a $\text{Conv}_{1\times1}$ layer is initially employed to adjust the channel dimensions while partitioning it into $\boldsymbol{p}_o$ and $\boldsymbol{q}$ along the channel dimension

$$[\boldsymbol{p}_o, \boldsymbol{q}] = \text{Conv}_{1\times1}(\boldsymbol{x}) \in \mathbb{R}^{2C\times H\times W} \tag{2}$$

with $\boldsymbol{p}_o \in \mathbb{R}^{C/2^{n-1}\times H\times W}$, $\boldsymbol{q} \in \mathbb{R}^{2C-(C/2^{n-1})\times H\times W}$.

Subsequently, $\boldsymbol{q}$ undergoes a depth-wise convolution layer $\text{DWConv}_{7\times7}$, while being split into a series of projected n-order features $\{\boldsymbol{q}_k \in \mathbb{R}^{C/2^{n-k-1}\times H\times W}\}_{k=0}^{n-1}$. gnConv is then executed, where k-order features p_k are matched in channel dimension with the k-order features q_k using $\text{Conv}_{1\times1}$ layer and then element-wise multiplication is used to accomplish the interaction between adjacent features. This ensures that the higher order features are progressively preserved, resulting in the final recursion step p_n

$$[\boldsymbol{q}_0, \boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}] = DW\text{Conv}_{7\times7}(\boldsymbol{q}) \tag{3}$$

$$\boldsymbol{p}_{k+1} = \begin{cases} \boldsymbol{p}_o \odot \boldsymbol{q}_o, & k = 0 \\ \text{Conv}_{1\times1}(\boldsymbol{p}_k) \odot \boldsymbol{q}_k, & 1 \le k \le n-1 \end{cases} \tag{4}$$

where $\odot$ denotes elementwise multiplication. From (4), the k-order interactions in $\boldsymbol{q}_k$ can be interpreted as the attention weighting for $\boldsymbol{p}_k$. Hence, gnConv can extend self-attention from second-order to arbitrary-order interactions. Finally, the output of the gnConv is obtained through a projection layer $\text{Conv}_{1\times1}$

$$\boldsymbol{y} = \text{Conv}_{1\times1}(\boldsymbol{p}_n) \in \mathbb{R}^{C\times H\times W}. \tag{5}$$

We chose HorNet as the backbone network for our proposed CAPNet, as its recursive gated convolutions can facilitate high-order spatial interactions. This capability is beneficial for identifying complex structures within BUA, such as shapes, sizes, and relationships between adjacent buildings and appurtenances.

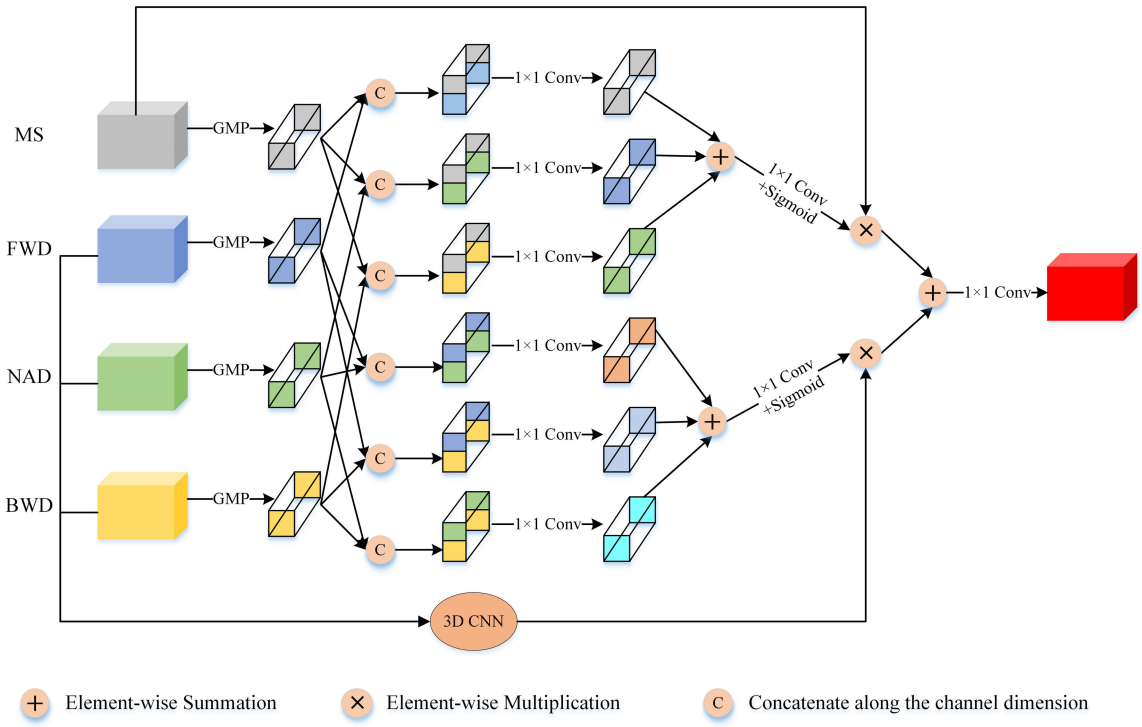

Fig. 7. Overview of the CAFM for multiangle features fusion.

B. Cross-Angle Fusion Module

Under various observation views, the spatial distribution of grayscale levels in MVHR imagery changes, yet the information provided by each view is highly correlated. As planar and stereoscopic features have distinct characterizing abilities for BUA, directly stacking multiangle features and feeding them into the network for training could introduce ambiguity in feature extraction. This is because there are significant differences in the semantics represented by different views.

To fully leverage the complementarity among multiangle features, we encode the multispectral, FWD-, NAD-, and BWD-view imagery as feature vectors $\boldsymbol{F}^{\mathrm{MS}}$, $\boldsymbol{F}^{\mathrm{FWD}}$, $\boldsymbol{F}^{\mathrm{NAD}}$, and $\boldsymbol{F}^{\mathrm{BWD}} \in \mathbb{R}^{C\times H\times W}$, respectively. This encoding process aims to extract multiangle features that reflect urban vertical structure. Here, C, H, and W denote these feature maps' channel dimensions, height, and width, respectively.

As depicted in Fig. 7, to facilitate the model's focus on differences among various views throughout the learning process, we employ the global max-pooling operation GMP$(\cdot)$ along the channel dimension to suppress spatial distribution information and condense the entire feature map into a single value. Consequently, it transforms spatial information into feature values denoted as $\boldsymbol{F}^{\mathrm{MS}}_{\mathrm{GMP}}$, $\boldsymbol{F}^{\mathrm{FWD}}_{\mathrm{GMP}}$, $\boldsymbol{F}^{\mathrm{NAD}}_{\mathrm{GMP}}$, and $\boldsymbol{F}^{\mathrm{BWD}}_{\mathrm{GMP}} \in \mathbb{R}^{C\times 1\times 1}$. This approach can effectively mitigate the offset introduced by angular disparities, and enable a more precise comparison of feature values at different positions under varying angles

$$\boldsymbol{F}^{\mathrm{MS}}_{\mathrm{GMP}} = \mathrm{GMP}(\boldsymbol{F}^{\mathrm{MS}}) \tag{6}$$
$$\boldsymbol{F}^{\mathrm{FWD}}_{\mathrm{GMP}} = \mathrm{GMP}(\boldsymbol{F}^{\mathrm{FWD}}) \tag{7}$$
$$\boldsymbol{F}^{\mathrm{NAD}}_{\mathrm{GMP}} = \mathrm{GMP}(\boldsymbol{F}^{\mathrm{NAD}}) \tag{8}$$
$$\boldsymbol{F}^{\mathrm{BWD}}_{\mathrm{GMP}} = \mathrm{GMP}(\boldsymbol{F}^{\mathrm{BWD}}). \tag{9}$$

Given that different angle combinations yield distinct ADFs, we map the acquired multiangle features onto channels and perform pairwise concatenation. Subsequently, we employ a 1×1 convolutional layer, accompanied by batch normalization and activation function, to capture nonlinear interactions among channels. This process is intended to characterize the angular difference properties of BUA observed from various views, which can be written as

$$\boldsymbol{F}^{\text{MS-FWD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{MS}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{FWD}}_{\mathrm{GMP}}\right)\right) \tag{10}$$
$$\boldsymbol{F}^{\text{MS-NAD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{MS}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{NAD}}_{\mathrm{GMP}}\right)\right) \tag{11}$$
$$\boldsymbol{F}^{\text{MS-BWD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{MS}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{BWD}}_{\mathrm{GMP}}\right)\right) \tag{12}$$
$$\boldsymbol{F}^{\text{FWD-NAD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{FWD}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{NAD}}_{\mathrm{GMP}}\right)\right) \tag{13}$$
$$\boldsymbol{F}^{\text{FWD-BWD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{FWD}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{BWD}}_{\mathrm{GMP}}\right)\right) \tag{14}$$
$$\boldsymbol{F}^{\text{NAD-BWD}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{NAD}}_{\mathrm{GMP}}, \boldsymbol{F}^{\mathrm{BWD}}_{\mathrm{GMP}}\right)\right) \tag{15}$$

where Cat$(:,:)$ represents the concatenation operation along the channel dimension and $\mathrm{Conv}_{1\times 1}$ represent a learnable 1×1 convolution layer with stride 1 used for lateral feature connections.

Subsequently, three adjacent connected angular features are summed. A 1×1 convolutional layer and the sigmoid function are then applied to learn nonexclusive relationships, emphasizing multiple channels. The process to obtain the weight maps for each view is formulated as

$$\boldsymbol{F}^{A1} = \sigma(\mathrm{Conv}_{1\times 1}(\boldsymbol{F}^{\text{MS-FWD}} + \boldsymbol{F}^{\text{MS-NAD}} + \boldsymbol{F}^{\text{MS-BWD}})) \tag{16}$$
$$\boldsymbol{F}^{A2} = \sigma(\mathrm{Conv}_{1\times 1}(\boldsymbol{F}^{\text{FWD-NAD}} + \boldsymbol{F}^{\text{FWD-BWD}} + \boldsymbol{F}^{\text{NAD-BWD}})). \tag{17}$$

As for the additive operations, F^{A1} focuses on integrating complementary information between spectral-angular combinations, providing the model with comprehensive spectral and morphological features. F^{A2} focuses on the multiview differences and their complementarity, aiding the model in understanding and utilizing the relationships between different viewing angles. These specific combinations enhance the model's ability to perceive and interpret complex BUA environments from various viewpoints.

Afterward, the FWD-, NAD-, and BWD-view features are stacked, and a 3-D operation, which facilitates cross-channel information interaction, is applied to obtain the cross-view features. The weights $\boldsymbol{F}^{A1}$ and $\boldsymbol{F}^{A2}$ are employed to conduct weighted operations on the features $\boldsymbol{F}^{\mathrm{MS}}$ and $\boldsymbol{F}^{\mathrm{MAF}}$, and dynamically adjust the fusion weights for different views to achieve spatial alignment of features

$$\boldsymbol{F}^{\text{3D-MAF}} = \mathrm{ST}\left(\boldsymbol{F}^{\mathrm{FWD}}, \boldsymbol{F}^{\mathrm{NAD}}, \boldsymbol{F}^{\mathrm{BWD}}\right) \tag{18}$$

$$\boldsymbol{F}^{\mathrm{MAF}} = 3D\left(\boldsymbol{F}^{\text{3D-MAF}}\right) \tag{19}$$

$$\boldsymbol{F}^{2} = \boldsymbol{F}^{A2} \cdot \boldsymbol{F}^{\mathrm{MAF}} \tag{20}$$

$$\boldsymbol{F}^{1} = \boldsymbol{F}^{A1} \cdot \boldsymbol{F}^{\mathrm{MS}} \tag{21}$$

where ST(:, :) represents the stacking operation and N represents the number of angles.

Concerning the 3-D operation, given an element $\boldsymbol{x}_{ija}$ within $\boldsymbol{F}^{\text{3D-MAF}} \in \mathbb{R}^{N\times C\times H\times W}$, where i and j denote coordinates in the spatial domain plane, and a denotes the angular position. The process to obtain $\boldsymbol{y}_{ija}$ within $\boldsymbol{F}^{\mathrm{MAF}}$ using a convolution kernel $\boldsymbol{w} \in \mathbb{R}^{N\times 1\times 1}$ can be formulated as

$$\boldsymbol{y}_{ija} = F\left(\boldsymbol{b} + \sum_{1}^{n}\sum_{w_i}^{K}\sum_{w_j}^{K}\sum_{w_a}^{N} w^{n}_{w_i w_j w_a} \boldsymbol{x}^{n}_{i+w_i, j+w_j, a+w_a}\right) \tag{22}$$

where $w^{n}_{w_i w_j w_a}$ represents the value located at (w_i, w_j, w_a) in the convolutional kernel of the nth layer feature map in $\boldsymbol{F}^{\text{3D-MAF}}$.

The dimensions (N, K, and K) denote the size of the 3-D convolution kernel along the angular, width, and height dimensions, respectively. $\boldsymbol{b}$ corresponds to the offset tensor and $F(\cdot)$ represents the activation function. This approach can preserve the local contextual relationship of spectral data blocks and multiangle tensor texture features in 3-D space, which reflects the correlation and variation of MVHR imagery. Finally, a 1×1 convolution layer is employed on the concatenation of the $\boldsymbol{F}^1, \boldsymbol{F}^2$ to extract an appropriate description $\boldsymbol{F}^{\mathrm{final}}$ of the implicit correlation among multiple angles from the fused feature vector

$$\boldsymbol{F}^{\mathrm{final}} = \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}\left(\boldsymbol{F}^{1}, \boldsymbol{F}^{2}\right)\right). \tag{23}$$

Traditional multiangle feature representation methods directly handle multiangle tensor features composed of rows, columns, and angular modes, which often results in the loss of local contextual relationships among multiangle textures [19]. To deal with this issue, this article proposes CAFM, to leverage the intrinsic structural distinctions among angular features. This module augments the discriminative capabilities when dealing with diverse, complex urban features by adopting an attention mechanism to capture and describe contextual relationships(i.e., the correlations among pixels from different views). It subsequently jointly learns deeper features in the spatial–spectral-angular domain with enhanced discriminative ability.

C. Multiangular Propagation Decoder

CAFM is designed to extract distinctions angular variations insights from multiviews, enabling the rational propagation and fusion of complementary multiscale, multiangle features characterized by weak interactions.

In order to further enrich the multilevel feature interactions and maximize the utilization of extracted ADFs, the proposed MAPD first built the angular-specific shallow-deep pathway, in which the low-level fine-scale information is propagated under the guide of the high-level semantic knowledge from shallow to deep layer, and the angular-specific deep-shallow pathway is used to propagate large-scale global semantic information from deep to shallow layer. This enables the decoder to perceive and leverage neighborhood and global information related to BUA, ultimately leading to a more discriminative representation across various scales.

The details of MAPD are illustrated in Fig. 8. Formally, the multilevel initial features from the encoder are denoted as: $\boldsymbol{S}_1 \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{S}_2 \in \mathbb{R}^{2C\times (H/2)\times (W/2)}$, $\boldsymbol{S}_3 \in \mathbb{R}^{4C\times (H/4)\times (W/4)}$, and $\boldsymbol{S}_4 \in \mathbb{R}^{8C\times (H/8)\times (W/8)}$. Note that $\boldsymbol{S}_4$ is the extracted high-level ADF by CAFM. Here, C, H, and W, respectively, represent the channel dimensions, height, and width of the feature maps extracted during the first stage of the encoder. For the shallow-deep pathway, the initial integration encompasses low-level features that lack multiangle disparities. In detail, we progressively up-sample the features $\boldsymbol{S}_2$, $\boldsymbol{S}_3$ to the same resolution as the feature $\boldsymbol{S}_1$, obtaining $\mathrm{UP}^2(\boldsymbol{S}_2), \mathrm{UP}^4(\boldsymbol{S}_3)$. Subsequently, these upsampled features are channel-wise concatenated with the feature $\boldsymbol{S}_1$ and then a 1×1 convolutional layer is added to extend it to a much richer space, obtaining $\boldsymbol{C}_1 \in \mathbb{R}^{8C\times H\times W}$. Following this, the features $\boldsymbol{S}_4$ is up-sampled to align with the resolution of $\boldsymbol{S}_1$ and is concatenated with $\boldsymbol{C}_1$ and $\mathrm{UP}^4(\boldsymbol{S}_3)$ to obtain $\boldsymbol{C}_2$. The process continues with the concatenation and dimensionality reduction of $\boldsymbol{C}_2$ with $\mathrm{UP}^8(\boldsymbol{S}_4)$, resulting in $\boldsymbol{C}_3$. Finally, $\boldsymbol{C}_3$ is concatenated and dimensionality reduction is performed alongside $\mathrm{UP}^8(\boldsymbol{S}_4)$, ultimately yielding $\boldsymbol{C}_4$. By aligning the channel dimensions of multilevel features $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_3$, and $\boldsymbol{C}_4$, we can effectively capture intricate semantic information and fine-grained details. This, in turn, bolsters the model's understanding of angular variations in BUA within complex scenes. This gradual approach enables the progressive capture of finer details, facilitating accurate extraction in scenarios involving small targets and complex scenes. The above-mentioned procedures can be expressed as

$$\boldsymbol{C}_i = \begin{cases} \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}(\boldsymbol{S}_1, \mathrm{UP}^2(\boldsymbol{S}_1), \mathrm{UP}^4(\boldsymbol{S}_3))\right), & i = 1 \\ \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}(\boldsymbol{C}_1, \mathrm{UP}^4(\boldsymbol{S}_3), \mathrm{UP}^8(\boldsymbol{S}_4))\right), & i = 2 \\ \mathrm{Conv}_{1\times 1}\left(\mathrm{Cat}(\boldsymbol{C}_{i-1}, \mathrm{UP}^8(\boldsymbol{S}_4))\right), & i = 3, 4 \end{cases} \tag{24}$$

where $\boldsymbol{C}_i \in \mathbb{R}^{8C\times H\times W}$, i is the index of different scale layers in the shallow-deep pathway. $\mathrm{UP}^j(\cdot)$ represents the bilinear

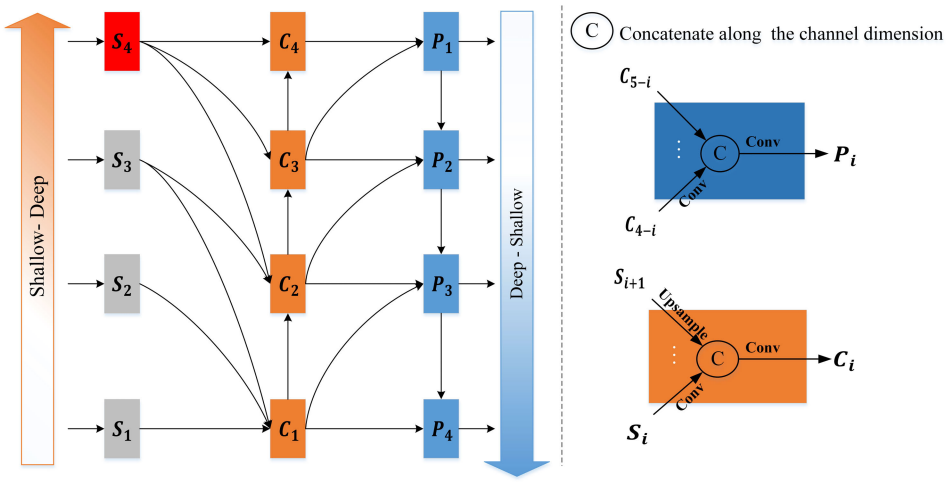

Fig. 8. (a) Feature information propagation way in the MAPD for enhancing the ability of hierarchical feature representation. Note that orange and blue rectangles represent the bottom–top and deep-shallow pathways, respectively. (b) Lightweight cross-scale connection node, as a crucial component of MAPD, functions as a bridge between features of varying scales in both preceding and current pathways. Its primary responsibility is the aggregation and reconstruction of multilevel feature dimensions, ensuring a seamless flow of information across scales.

interpolation algorithm with a scale factor of j. $\text{Conv}_{1\times1}$ represents the 1×1 convolution layer with a stride of 1, which is subsequently followed by batch normalization and an activation function. We employ 1×1 convolutional layer to achieve channel dimension transformation, incurring relatively low computational costs, and better handling of details and texture information. Upon observation, it becomes evident that the multiangle disparity information present in $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, $\boldsymbol{C}_3$, and $\boldsymbol{C}_4$ exhibits a gradual increase, where in $\boldsymbol{C}_1$ is devoid of such information. Consequently, we systematically propagate features rich in angle disparity information backward to those with lesser amounts, ensuring a comprehensive utilization of available multiscale, multiangle features. Specifically, the deep-shallow pathway connects adjacent features through lateral concatenation operation. In brief, the cross-scale connection node takes the corresponding adjacent feature of the shallow-deep pathway and the previous fused features as the input to update the current fused features. Afterward, the fused feature acts as the input for the next node. This process is repeated for all adjacent scale layers, resulting in fused multiscale semantic features $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_3$, and $\boldsymbol{P}_4$. This helps extract more discriminative representations by leveraging prior information. In addition, the connected features are carried with larger convolution kernel weights to facilitate channel compression, which can effectively expand the receptive field for feature perception and ensure the incorporation of regionally correlated feature details. The process of the deep-shallow pathway can be described as

$$\boldsymbol{P}_i = \begin{cases} \text{Conv}_{3\times3}(\text{Cat}(\boldsymbol{C}_3, \boldsymbol{C}_4)), & i = 1 \\ \text{Conv}_{3\times3}(\text{Cat}(\boldsymbol{P}_{i-1}, \boldsymbol{C}_{5-i}, \boldsymbol{C}_{4-i})), & i = 2, 3 \\ \text{Conv}_{3\times3}(\text{Cat}(\boldsymbol{P}_3, \boldsymbol{C}_1)), & i = 4 \end{cases} \tag{25}$$

where $\boldsymbol{P}_i \in \mathbb{R}^{8C\times H\times W}$, represents the feature layer used for BUA extraction after feature fusion. For both shallow-deep and deep-shallow pathways in this article, the convolutional kernels are used with a stride of 1. This is crucial for preserving as much spatial information as possible in an accurate BUA extraction.

In the context of BUA extraction tasks, our MAPD introduces an innovative fusion approach for multiangle disparity information and multiscale information. This approach, distinct from existing multiscale fusion decoder structures, incorporates shallow-deep, deep-shallow propagation, and lateral connections. Through this enhanced integration of multiangle, multiscale information, MAPD generates semantically rich features that maintain consistent dimensions.

D. Logical Classification Layer

After CAFM completes the interaction of multiangle disparity information, MAPD progressively facilitates semantic interaction between multiscale and multiangle information bidirectionally across different feature levels, ultimately generating a series of feature maps $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{P}_3$, and $\boldsymbol{P}_4$, encompassing varying degrees of textural details, semantic relevance, and multiangle disparity information. However, a notable issue arises from the lack of inherent connections among these features. Therefore, in this study, we directly aggregate these hierarchical features through channel concatenation and obtain the BUA extraction results in $\boldsymbol{S}^{\text{logist}}$. This enables the network to capture more comprehensive and discriminative multiangle and multiscale representations, significantly expanding the range for feature selection and combination. The process can be formulated as

$$\boldsymbol{S}^{\text{logist}} = \text{Conv}_{1\times1}(\text{Cat}(\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4)). \tag{26}$$

For ease of expression, we collectively refer to the 1×1 convolution layer, followed by batch normalization and an activation function as $\text{Conv}_{1\times1}$.

V. Experiment

A. Datasets

To comprehensively validate the effectiveness and feasibility of CAPNet for large-scale BUA extraction, in this study, experiments were conducted based on the MVBA and WV-3 multiview semantic stereo datasets. The results were then compared with SOTA models.

1) MVBA Dataset: This research conducts experiments in 41 representative cities across China, encompassing coastal, inland, mountainous, and plain regions, thereby representing a wide range of typical Chinese cities. Furthermore, the study areas were selected with deliberate consideration of the economic development levels, encompassing megacities, medium-sized urban centers, and smaller towns. The acquisition dates for all imagery are within the growing season, ranging from April to August. For our experiments, we have specifically chosen high-quality imagery with cloud cover below 10%. All imagery contains four multispectral bands: blue (450∼520 nm), green (520∼590 nm), red (630∼690 nm), and near-infrared (770∼890 nm). Furthermore, multiview images are available: FWD (500∼800 nm), NAD (500∼800 nm), and BWD (500∼800 nm).

2) WV-3 Multiview Semantic Stereo Dataset: The WV-3 multiview semantic stereo public dataset comprises 4188 pairwise two-view stereo semantic images sourced from the urban semantic 3-D (US3D) project [39]. These images cover an area of approximately 100 km^2, encompassing Jacksonville, Florida, and Omaha, Nebraska, USA. The size of each imagery is 1024 × 1024 pixels, and each imagery contains three spectral bands. Semantic classes included buildings, elevated roads and bridges, low vegetation, impermeable surfaces, and water.

B. Implementation Details

For fair comparisons, all experiments were implemented on two GeForce RTX 3090 graphics processing units (GPUs). The experiment configuration is detailed in Table III. The MVBA and WV-3 multiview semantic stereo datasets are randomly partitioned into training, validation, and test sets with a 7:2:1 ratio.

To evaluate the effectiveness of BUA extraction, this article conducted quantitative comparative experiments using the mIoU, mF1, and Kappa metrics

$$\text{mIoU} = \frac{1}{m+1}\sum_{i=0}^{m}\frac{\text{TP}}{\text{FN}+\text{FP}+\text{TP}} \tag{27}$$

$$\text{mF1} = \frac{1}{m+1}\sum_{i=0}^{m}\frac{2\times\text{TP}}{2\times\text{TP}+\text{FP}+\text{FN}} \tag{28}$$

$$\text{Kappa} = \frac{p_o - p_e}{1-p_e} \tag{29}$$

$$p_o = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{30}$$

$$p_e = \frac{(\text{TP}+\text{FP})\cdot(\text{TP}+\text{FN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2} + \frac{(\text{FN}+\text{TN})\cdot(\text{FP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})^2} \tag{31}$$

TABLE III
Details of Experiment Configuration

Experiments Platforms	
CPU	Intel(R) Xeon(R) Silver 4210
GPU	3090
Optimizer	
type	AdamW
lr	0.0001
betas	(0.9, 0.999)
weight_decay	0.05
decay_rate	0.9
decay_type	stage_wise
lr_config	
policy	poly
warmup	linear
warmup_iters	1500
warmup_ratio	1e-6
power	1.0
min_lr	0.0
Hyperparameters	
Batch size	2
Max iteration	40k
Others	
Loss function	Cross Entropy Loss

where m represents the total number of classes. True positive (TP) denotes the number of pixels where the predicted and true values are both BUA. True negative (TN) indicates the number of pixels where both predicted and true values correspond to non-BUA. False positive (FP) represents the number of pixels predicted as BUA but are, in fact, non-BUA. False negative (FN) corresponds to the number of pixels predicted as non-BUA but are actually BUA.

C. Results and Analysis

Table IV presents the comparison results of the proposed CAPNet with other SOTA models in the BUA extraction. For fairness, the NAD-, FWD-, and BWD-view panchromatic imagery and NAD-view multispectral imagery are directly concatenated and input into the comparison models along the channel dimension. The results demonstrate that, compared to the current SOTA models, the proposed CAPNet achieves more accurate BUA extraction, giving the highest mIoU, mF1, and Kappa values, which are 85.32%, 91.96%, and 83.93%, respectively. In terms of the mIoU evaluation metric, CAPNet, when compared to the baseline network (HorNet), exhibits an accuracy improvement of 3.09%. Furthermore, when compared to HRNet, ISANet, UNet, Poolformer, Segformer, and DeeplabV3+, CAPNet shows accuracy increments of 5.44%, 2.82%, 7.75%, 4.70%, 2.16%, and 1.92%, respectively. These results verify the better performance of CAPNet in multiview feature representation and BUA detection.

Moreover, it can be observed that due to the limitations of convolutional operations in capturing global feature representations, transformer-based models (such as Poolformer, Segformer, and ISANet) outperform CNNs-based models (HRNet and UNet) in the task of BUA extraction, particularly in exploring MVHR imagery. This also suggests that the MVBA dataset exhibits sufficient diversity, enabling

TABLE IV
QUANTITATIVE METRIC COMPARISONS OF DIFFERENT MODELS ON THE MVBA DATASET

Method	Encoder/Decoder	Per-Class IOU/F1 Score(%) BUA	Non-BUA	mIoU(%)	MeanF1(%)	Kappa(%)	Params (M)	Flops (GFLOPs)
HRNet [50]	HRNet48/FCNHead	86.99/93.04	72.77/84.24	79.88	88.64	77.30	65.80	93.65
ISANet [58]	ResNet50/ISAHead	89.00/94.18	75.99/86.36	82.50	90.27	80.54	56.80	227.49
UNet [51]	UNet/FCNHead	86.28/92.64	68.86/81.56	77.57	87.10	74.06	65.85	93.65
Poolformer [60]	PoolFormer/FPN	88.10/93.68	73.13/84.48	80.62	89.08	78.16	93.09	43.71
Segformer [61]	MixVision/Segformer	89.78/94.61	76.54/86.71	83.16	90.66	81.33	52.76	82.03
DeeplabV3+ [52]	ResNet101/ASPP	89.83/94.64	76.97/86.98	83.40	90.81	81.63	62.69	255.26
UperNet [53]	HorNet/UpeNret	89.01/94.18	75.46/86.01	82.23	90.10	80.20	52.21	231.5
CAPNet(Ours)	**4*HorNet/Propagation**	**91.40/95.51**	**79.23/88.42**	**85.32**	**91.96**	**83.93**	**79.56**	**387.69**

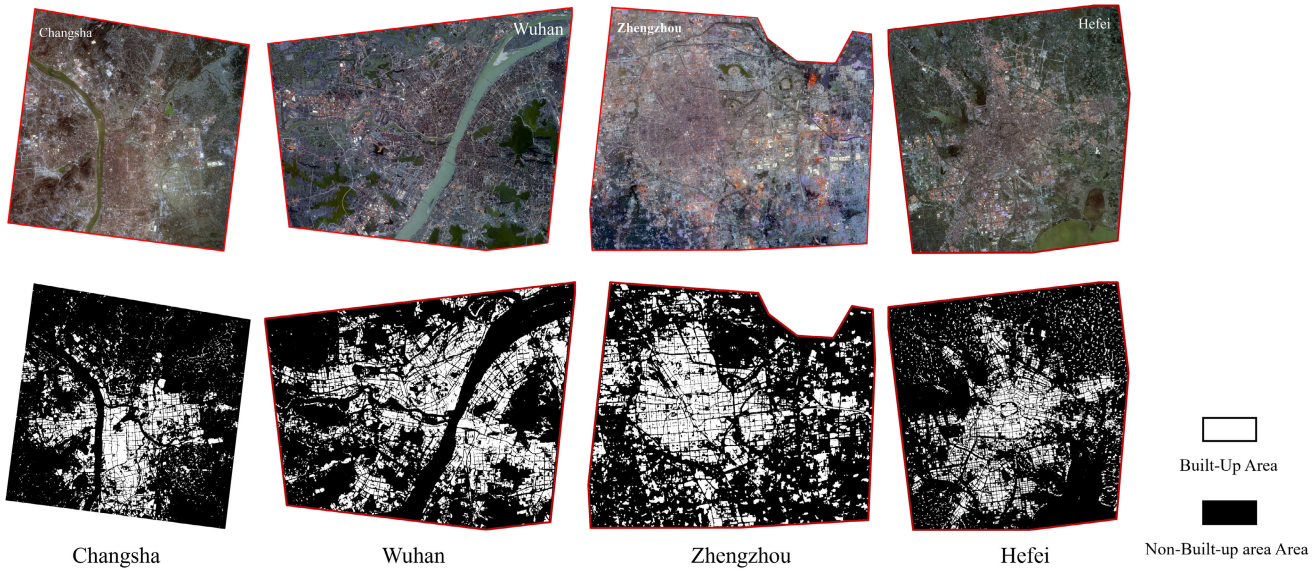

Fig. 9. Overall BUA extraction results for some test cities.

transformer-based models to better capture complex patterns and features from the data.

Fig. 9 shows the comprehensive performance of CAPNet in extracting a wide range of BUA. The first and second rows, respectively, depict the MVHR imagery and the corresponding extracted BUA for selected test cities. In these visualizations, white and black areas represent BUA and non-BUA, while a red dashed line outlines the boundary of the test city. It is worth noting that CAPNet consistently preserves detailed information regarding BUA across various urban landscapes. For instance, CAPNet accurately extracts both small residential areas and expansive commercial zones within the urban core, while effectively filtering out major roads, rivers, parks, or extensive bare soil/ground areas. Furthermore, it demonstrates its capability to identify scattered small villages in suburban regions.

Fig. 10 illustrates the randomly selected test patches of four representative Chinese provincial capital cities: Wuhan, Hefei, Changsha, and Zhengzhou. These cities have high population densities and are distributed across diverse regions, showcasing distinct landscapes. In 10, the red elliptical regions are marked as areas prone to misclassification or omission, located in the cities' core, suburban, or rural areas. Red and blue regions denote misclassified and missed BUA, respectively. White and black regions correspond to correctly extracted BUA and non-BUA. Specifically, in the test area located in the suburbs [P#(a)] of Changsha, BUA is dispersed and characterized by irregular structures, encompassing high buildings and small, low-density buildings of varying sizes, and shadows cast by the tall buildings further compound the scene. In the test area situated along the river [P#(b)] in Wuhan, the landscape primarily consists of bare soil, roads, and dense buildings, with limited vegetation cover. In the test area located in a town [P#(c)] in Zhengzhou, buildings are densely and systematically arranged and subdivided into several blocks by streets. In the test area located in a rural [P#(d)] in Hefei, BUA is sparsely scattered in areas of bare soil. These complex and challenging scenarios offer a valuable testing scenario to evaluate the model's capability in addressing missed detections and false positives.

Overall, it can be observed that the proposed CAPNet can extract clear outlines and comprehensive internal details of BUA in both urban and rural regions, where CAPNet effectively reduces omissions, minimizes false alarms, and exhibits fewer instances of adhesion and fragmentation. Results show that traditional U-Net exhibits noticeable misclassifications and omissions. HRNet captures abundant spatial information about BUA with clear boundaries, and it reasonably fills in

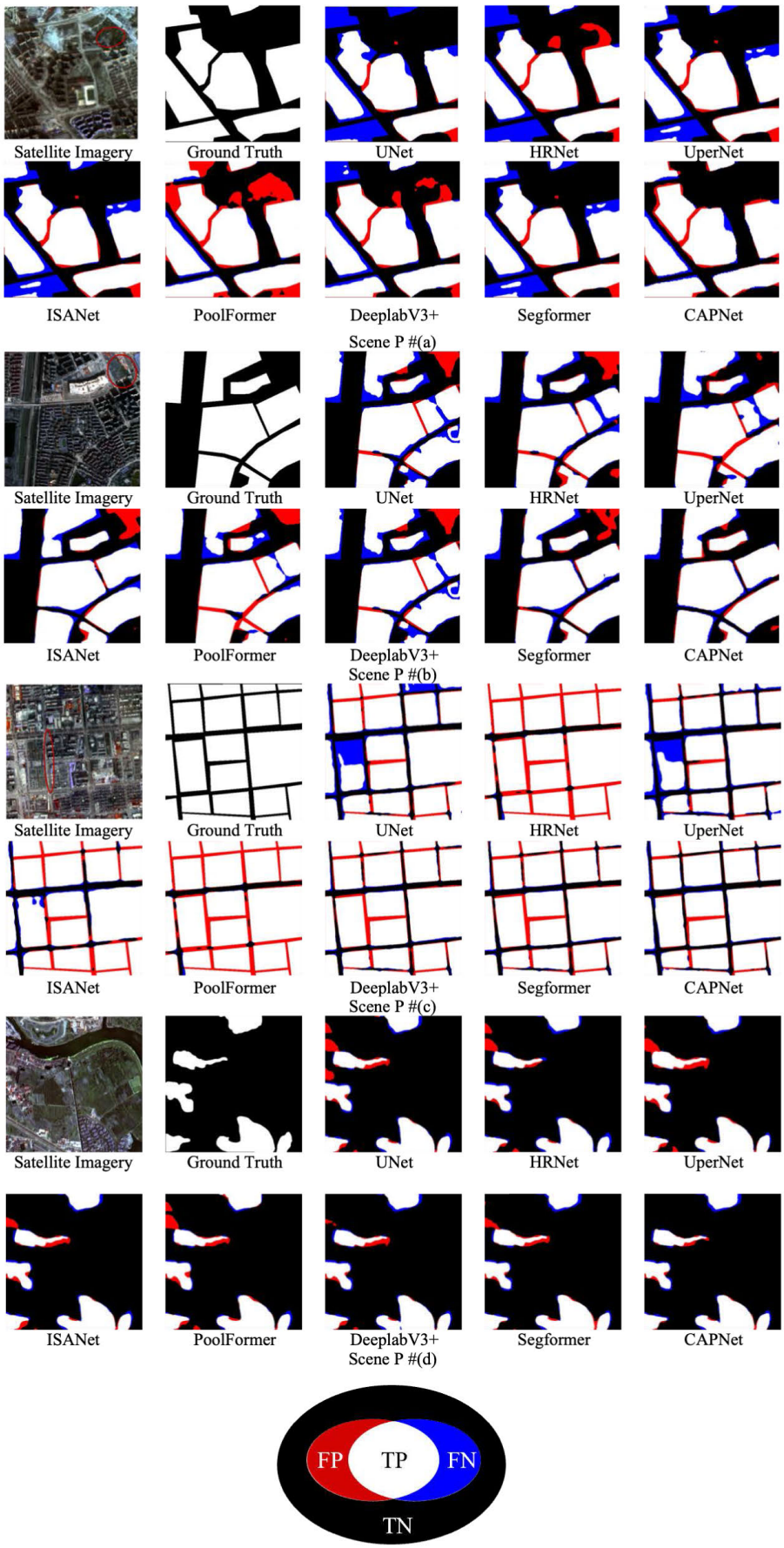

Fig. 10. Comparison of representative semantic segmentation results of different networks on the MVBA dataset. The scenes P#(a)–P#(d) illustrate the extraction results of BUA from four representative cities: Changsha, Wuhan, Zhengzhou, and Hefei, respectively.

TABLE V
QUANTITATIVE METRIC COMPARISONS OF DIFFERENT MODELS ON THE WV-3 MULTIVIEW SEMANTIC STEREO DATASET

Method	Encoder/Decoder	Per-Class IOU/F1 Score(%)					mIoU(%)	MeanF1(%)	Kappa(%)
		Groud	Tree	Building	Water	Clutter			
HRNet [50]	HRNet48/FCNHead	89.40/94.40	63.57/77.73	83.29/90.88	91.40/95.61	85.27/92.05	82.58	90.11	84.34
ISANet [58]	ResNet50/ISAHead	91.98/95.82	71.29/83.24	87.99/93.61	93.98/96.89	90.78/95.17	87.21	92.95	88.48
UNet [51]	UNet/FCNHead	88.76/94.05	62.80/77.15	81.79/89.98	88.39/93.84	82.03/90.13	80.76	89.03	83.21
Poolformer [60]	PoolFormer/FPNHead	81.31/89.69	47.13/64.07	68.55/81.34	78.36/87.87	55.52/71.40	66.17	78.87	70.91
Segformer [61]	MixVision/Segformer	87.14/93.13	57.82/73.27	78.86/88.19	87.60/93.39	72.24/83.88	76.73	86.37	80.36
DeeplabV3+ [52]	ResNet101/ASPP	85.06/95.20	68.07/81.00	86.00/92.47	92.70/97.11	87.69/93.44	85.06	91.67	86.70
UperNet [53]	HorNet/UperNet	91.05/94.87	68.43/80.83	86.06/92.30	92.47/95.86	88.39/93.62	85.12	91.50	86.63
Ours	**2*HorNet/Propagation**	**92.28/95.99**	**72.27/83.90**	**88.43/93.86**	**94.38/97.11**	**90.64/95.09**	**87.60**	**93.19**	**88.94**

gaps within the BUA to some extent. However, when dealing with areas with bare soil that have similar textures to BUA, HRNet may result in misclassifications. Poolformer provides relatively coarse extraction results of BUA, showing inaccurate boundary delineation. In contrast, in test areas P#(a) and P#(b), CAPNet can alleviate the adverse effects of bare soil areas near BUA that share similar visual textures, while also producing relatively complete boundaries. When dealing with the densely populated test area P#(c), only CAPNet can avoid the detection of roads between adjacent building clusters, and correctly identify them as non-BUA. In test area P#(d), CAPNet accurately identifies the small, isolated buildings in suburban areas as non-BUA but other models fail.

VI. DISCUSSION

A. Additional Experiments With WV-3 Multiview Dataset

We further utilized the publicly WV-3 multiview semantic stereo dataset to evaluate the effectiveness of CAPNet across various scene categories not just the BUA category. Since the WV-3 semantic stereo dataset encompasses only two viewing angles, we designated the left- and right-view imagery as the input for CAPNet, to capture the angular differential features.

The results presented in Table V clearly demonstrate that our proposed CAPNet achieves the highest mIoU, mF1, and Kappa scores, reaching 87.60%, 93.19%, and 88.94%, respectively, outperforming the current SOTA methods. Moreover, when considering the mIoU evaluation metric, CAPNet exhibits superior performance across various categories, including ground, tree, building, water, and clutter, with the highest scores of 92.28%, 72.27%, 88.43%, 94.38%, and 90.64%, respectively. Compared to the baseline network (HorNet), CAPNet exhibits gains of 1.23%, 3.84%, 2.37%, 1.91%, and 2.25% for these categories, resulting in an OA improvement of 2.48%.

Furthermore, when evaluated on individual categories, CAPNet is more effective in extracting buildings and impervious surfaces. Compared to HRNet, ISANet, UNet, Poolformer, Segformer, and DeeplabV3+, CAPNet enhances the accuracy of BUA extraction, by achieving improvements of 5.02%, 0.39%, 6.84%, 21.43%, 10.87%, and 2.54%, respectively.

As depicted in Fig. 11, we analyzed two typical test areas. Here, the white, cyan, blue, and yellow areas indicate ground, tree, building, and water regions, respectively. The red areas indicate misclassified and missed detection regions. Overall, CAPNet exhibits satisfactory performance and minimal artifacts. In the scenario P#(a), which exhibits a diverse range of multiscale features and distinct building morphologies, CAPNet outperforms other networks by accurately delineating small-scale trees and large-scale buildings. From the challenging high-density building scenario P#(b), it is evident that CAPNet achieves finer segmentation of individual buildings compared to other networks. Furthermore, in regions adjacent to buildings and trees, CAPNet produces finer results with fewer misclassifications. These results highlight that CAPNet can more effectively distinguish objects with similar characteristics within urban areas.

TABLE VI
PERFORMANCE OF DIFFERENT BACKBONE NETWORKS

Backbone	mIoU(%)
BiSeNetV2 [68]	81.52
MSCAN [69]	82.67
ConvNext [70]	81.86
MixVisionTransformer [61]	84.01
HorNet [67]	85.32

B. Ablation Experiments

We selected several commonly used architectures for comparison, including CNNs-based networks (BiSeNetV2 [68], SegNeXt [69], and ConvNext [70]) and transformer-based network (MixVisionTransformer [61]). As shown in Table VI, the experimental results indicate that HorNet demonstrates better accuracy performance. Specifically, in terms of MIoU, HorNet shows improvements of 3.80%, 2.65%, 3.46%, and 1.31% compared to BiSeNetV2, SegNeXt, ConvNext, and MixVisionTransformer, respectively. Therefore, HorNet is selected as the backbone network for CAPNet due to its superior performance.

Table VII indicates that utilizing all possible additive combinations may degrade the model accuracy slightly (−0.6%). It is noteworthy that in all possible additive combinations, we define $\boldsymbol{F}^{A1}$ as the sum of combinations that include $\boldsymbol{F}^{\text{FMS-NAD}}$, while $\boldsymbol{F}^{A2}$ represents the sum of combinations that do not include $\boldsymbol{F}^{\text{FMS-NAD}}$. In the single-view mode, only the MS-imagery was input into the CAPNet model (without CAFM). A possible explanation is that we have already performed all the pairwise combinations of multiview data

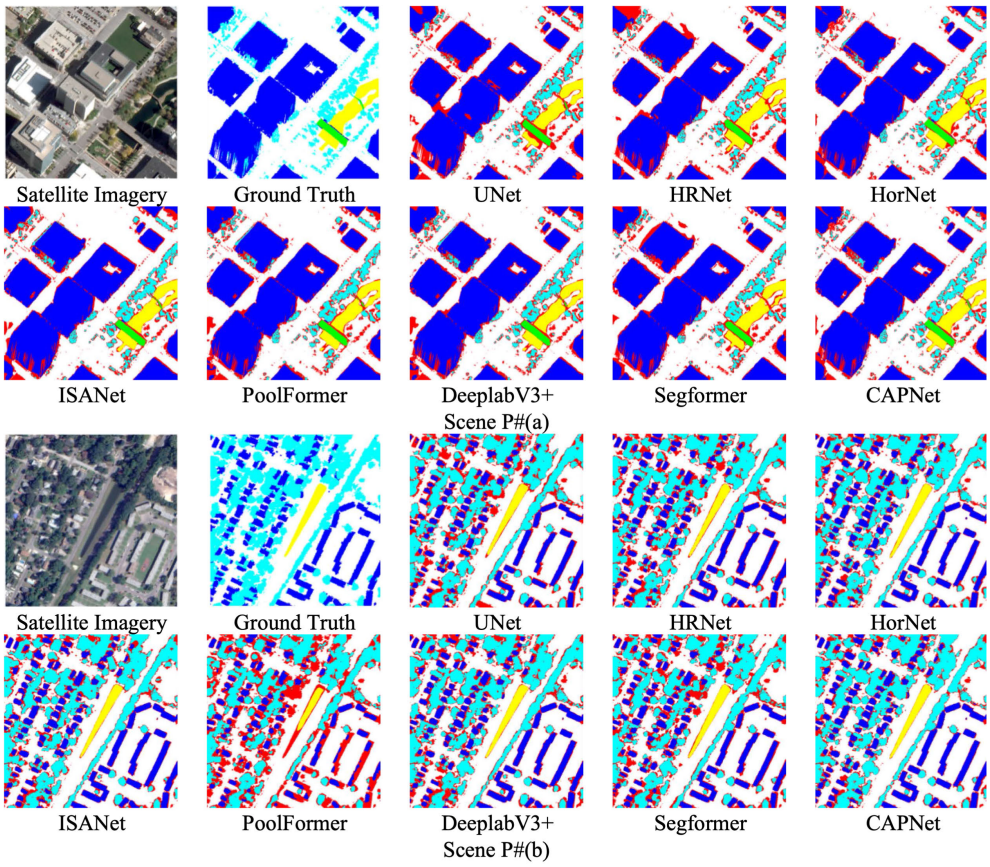

Fig. 11. Comparison of representative semantic segmentation results of different networks on the WV-3 multiview semantic stereo dataset. The scenes P#(a) and P#(b) illustrate the extraction results of two typical test areas: Scene P#(a) represents multiscale objects, and P#(b) represents a high-density building scenario.

TABLE VII
PERFORMANCE OF THE MODEL WITH DIFFERENT CONFIGURATIONS

Configurations	mIoU(%)
Singe-view Mode	82.85
All C_4^2 channel connections and all C_6^3 additive operations Mode	84.72
All C_4^2 channel connections and 2 additive operations Mode (CAPNet)	85.32

during the previous steps, and the ZY-3's MS and NAD-view imagery share identical viewing angles. Therefore, it becomes unnecessary to conduct additive operations for all possible combinations.

To further investigate the advantages of MAPD and CAFM in handling multispectral and multiangle features, this section conducts a series of ablation experiments based on MVBA.

As demonstrated in Table VIII, incorporating MAPD and CAFM into the baseline network has yielded gains of 1.12% and 2.23%, respectively, in terms of the mIoU. It can be seen that CAFM leads to a larger increase in accuracy compared to MAPD, and their combined utilization can further enhance the performance, resulting in a 3.09% improvement. The experiment results reveal that, compared to single-view mode, multiview features represented by the CAFM can more effectively capture the heterogeneity of objects' features, and thereby achieve better discrimination capabilities. In addition, the MAPD is beneficial for integrating contextual features across multiple scales and mitigating the information loss incurred by repeated up-and-down sampling.

TABLE VIII
ABLATION EXPERIMENTS OF CAFM AND MAPD

MAPD	CAFM	mIoU(%)
✘	✘	82.23
✔	✘	83.35
✘	✔	84.46
✔	✔	85.32

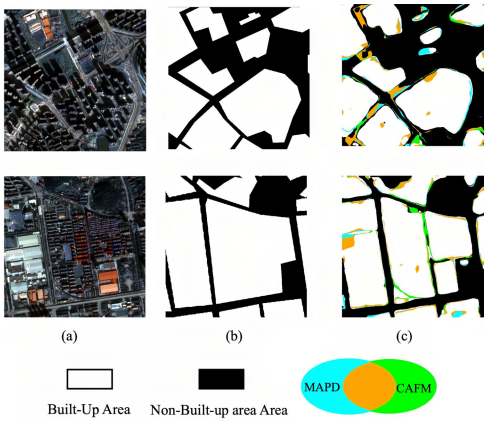

Fig. 12. Visualization of ablation experiments results: (a) true-color satellite imagery, (b) ground truth, and (c) BUA optimization map, displaying effects of MAPD and CAFM.

For a more comprehensive analysis of each module of the proposed CAPNet, Fig. 12 illustrates the performances when

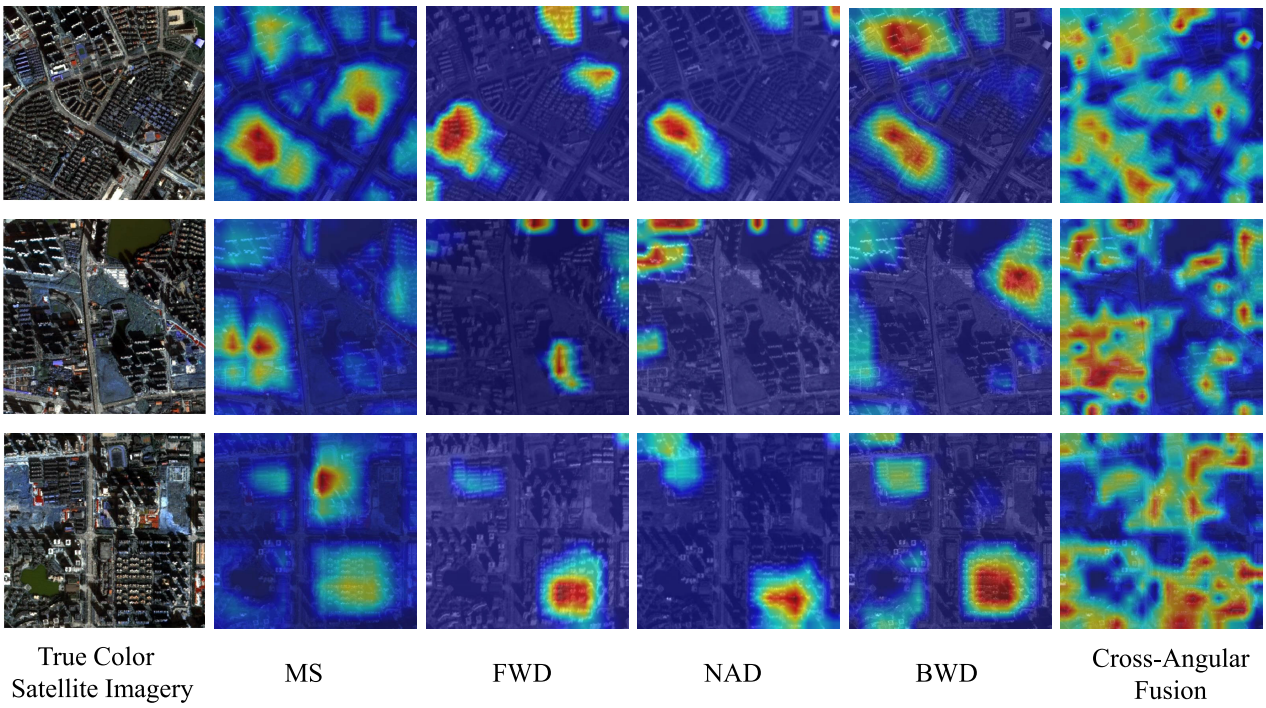

Fig. 13. Visual explanation to the category "BUA" generated at different views using Grad-CAM. Red regions indicate high scores for the "BUA" category.

MAPD and CAFM are adopted individually and in combination. Here, the white and black regions indicate BUA and non-BUA areas, respectively. Simultaneously, combinations of blue and orange, green and orange, and green, orange, and blue, respectively, denote correctly predicted as BUA when individually incorporating MAPD, CAFM, and the simultaneous introduction of MAPD and CAFM, in comparison to the baseline model. It can be observed that the initial results exhibit numerous artifacts and missed detections, resulting in fragmented patches within the BUA and difficulties in preserving boundary details. By incorporating MAPD, information loss is reduced. Specifically, the jaggedness at the boundaries is alleviated, with more refined and orderly segmentation contours while suppressing non-BUA noise. It can be also seen that using only single-angle information can lead to false alarms, imprecise boundaries, and misclassifications. Particularly, there exist omissions in the low-rise structures such as rural buildings. These issues primarily originate from objects like roads, farmlands, open spaces, and bare soil that exhibit similarities to BUA. To compensate for the limitations of single-view imagery in describing the structure of BUA, CAFM has the capability to learn spectral and structural differences of objects across MVHR imagery. It integrates angular and spectral information efficiently, and highlights pixels with significant angular differences, thus achieving more reliable BUA extraction results.

C. Visualization Analysis

This study introduces a CAFM based on MVHR imagery, which can effectively extract multiangular features and better represent the structural characteristics of BUA. As demonstrated in Fig. 13, using the Grad-CAM method [71], we highlight the regions where the CAFM significantly influences the prediction results, so as to understand how the CAFM assists the model's decision.

Fig. 13 illustrates that MS-, NAD-, FWD-, and BWD-view features can depict the characteristics of BUA from different viewing angles. However, each viewing mode has its own emphasis and often focuses on large areas with limited detail. In contrast, the angular-fusion features generated by the CAFM can highlight more comprehensive areas in densely built and high-rise regions, and at the same time, capture both detailed and global information of BUA.

D. Future Directions

For the MVBA extraction task, while this study has achieved satisfactory results through dataset construction and model design, there are still limitations as follows.

1) *Transferability Challenges:* When dealing with larger-scale BUA extraction, the model's generalization capacity is restricted when exposed to data with diverse characteristics and scenes. Therefore, future research could explore domain adaptation methods to mitigate the generalization errors and domain shift issues stemming from distribution inconsistencies.
2) *Data Source Diversity:* The data sources utilized in the proposed CAPNet are still limited (multispectral and multiview optical remote sensing imagery). Further effective fusion of diverse data sources (such as DEM, OSM, nighttime light imagery, and radar images)

can be employed to achieve a more accurate BUA extraction.

3) *Model Complexity:* The CAPNet employs a multibranch encoder structure, resulting in more network parameters. In future endeavors, model light-weighting could be considered to accelerate model computation speed while maintaining extraction effectiveness. This is particularly valuable for high-resolution BUA extraction tasks at the global scale.

VII. Conclusion

BUAs, which serve as the central zones for both residential living and industrial production, are intricately connected to the ecological conditions of the Earth's surface. This article focuses on the reliable extraction of high-resolution BUA, by utilizing China's MVHR imagery as the primary data source. Starting with constructing a sample dataset, the core of this study centers around the research of MVBA extraction algorithms.

We first utilize multispectral and MVHR imagery acquired from the ZY-3 satellites to construct a high-resolution and multiview dataset for BUA extraction (MVBA). Encompassing 41 representative cities in China, this dataset effectively captures the spatial distribution of BUA across various types, regions, and topographical scenes. MVBA can play a crucial role in promoting the utilization of DL techniques for large-scale, multiview, and high-resolution BUA extraction tasks.

Currently, there is a lack of research on utilizing angular features for urban land cover interpretation. Therefore, in consideration of the characteristics of BUA, this article proposes CAPNet for BUA extraction by leveraging spatial–spectral-angular features based on the MVBA dataset. CAPNet employs a shared encoder to simultaneously extract spectral and angular semantic features, and CAFM is designed for an effective multiangle feature fusion. In addition, it integrates adjacent layers on shallow-deep and deep-shallow pathways during the decoder stage, which narrows the semantic gap between multilevel-angle features and strengthens the information flow more efficiently. The experimental results confirm that CAPNet demonstrates superior sensitivity in detecting BUA and considerably improves the accuracy of BUA extraction from complex scenes.

Acknowledgment

The authors would like to thank the editors and anonymous reviewers for their insightful remarks, which significantly improved this article.

References

[1] F. Wu et al., "Built-up area mapping in China from GF-3 SAR imagery based on the framework of deep learning," *Remote Sens. Environ.*, vol. 262, Sep. 2021, Art. no. 112515.

[2] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A robust built-up area presence index by anisotropic rotation-invariant textural measure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 1, no. 3, pp. 180–192, Sep. 2008.

[3] T. Esch et al., "Breaking new ground in mapping human settlements from space—The global urban footprint," *ISPRS J. Photogramm. Remote Sens.*, vol. 134, pp. 30–42, Dec. 2017.

[4] Y. Cao, X. Huang, and Q. Weng, "A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas," *Remote Sens. Environ.*, vol. 297, Nov. 2023, Art. no. 113779.

[5] C. Liu, X. Huang, Z. Zhu, H. Chen, X. Tang, and J. Gong, "Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities," *Remote Sens. Environ.*, vol. 226, pp. 51–73, Jun. 2019.

[6] Y. Tan, S. Xiong, and P. Yan, "Multi-branch convolutional neural network for built-up area extraction from remote sensing image," *Neurocomputing*, vol. 396, pp. 358–374, Jul. 2020.

[7] S. Li, S. Fu, and D. Zheng, "Rural built-up area extraction from remote sensing images using spectral residual methods with embedded deep neural network," *Sustainability*, vol. 14, no. 3, p. 1272, Jan. 2022.

[8] A. Verma, A. Bhattacharya, S. Dey, C. López-Martínez, and P. Gamba, "Built-up area mapping using Sentinel-1 SAR data," *ISPRS J. Photogramm. Remote Sens.*, vol. 203, pp. 55–70, Sep. 2023.

[9] Y. Chen, S. Yao, Z. Hu, B. Huang, L. Miao, and J. Zhang, "Built-up area extraction combing densely connected dual-attention network and multiscale context," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5128–5143, 2023.

[10] A. Schneider, M. A. Friedl, D. K. McIver, and C. E. Woodcock, "Mapping urban areas by fusing multiple sources of coarse resolution remotely sensed data," *Photogrammetric Eng. Remote Sens.*, vol. 69, no. 12, pp. 1377–1386, Dec. 2003.

[11] M. Pesaresi et al., *Operating Procedure for the Production of the Global Human Settlement Layer From Landsat Data of the Epochs 1975, 1990, 2000, and 2014*, document JRC97705, 1975, pp. 1–62.

[12] C. Corbane et al., "Enhanced automatic detection of human settlements using Sentinel-1 interferometric coherence," *Int. J. Remote Sens.*, vol. 39, no. 3, pp. 842–853, Feb. 2018.

[13] X. Huang, J. Li, J. Yang, Z. Zhang, D. Li, and X. Liu, "30 m global impervious surface area dynamics and urban expansion pattern observed by Landsat satellites: From 1972 to 2019," *Sci. China Earth Sci.*, vol. 64, no. 11, pp. 1922–1933, Nov. 2021.

[14] X. Huang, J. Yang, W. Wang, and Z. Liu, "Mapping 10 m global impervious surface area (GISA-10m) using multi-source geospatial data," *Earth Syst. Sci. Data*, vol. 14, no. 8, pp. 3649–3672, Aug. 2022.

[15] R. Khatami, G. Mountrakis, and S. V. Stehman, "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research," *Remote Sens. Environ.*, vol. 177, pp. 89–100, May 2016.

[16] X. Huang, H. Chen, and J. Gong, "Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 127–141, Jan. 2018.

[17] F. Peng, J. Gong, L. Wang, H. Wu, and P. Liu, "A new stereo pair disparity index (SPDI) for detecting built-up areas from high-resolution stereo imagery," *Remote Sens.*, vol. 9, no. 6, p. 633, Jun. 2017.

[18] R. Qin, J. Tian, and P. Reinartz, "Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images," *Int. J. Remote Sens.*, vol. 37, no. 15, pp. 3455–3476, Aug. 2016.

[19] Y. Cao and X. Huang, "A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112590.

[20] X. Huang et al., "A multispectral and multiangle 3-D convolutional neural network for the classification of ZY-3 satellite images over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10266–10285, Dec. 2021.

[21] W. Guo, D. Lu, Y. Wu, and J. Zhang, "Mapping impervious surface distribution with integration of SNNP VIIRS-DNB and MODIS NDVI data," *Remote Sens.*, vol. 7, no. 9, pp. 12459–12477, Sep. 2015.

[22] L. Moya, H. Zakeri, F. Yamazaki, W. Liu, E. Mas, and S. Koshimura, "3D gray level co-occurrence matrix and its application to identifying collapsed buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 14–28, Mar. 2019.

[23] X. Huang, W. Yuan, J. Li, and L. Zhang, "A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 654–668, Feb. 2017.

[24] Q. Li, X. Huang, D. Wen, and H. Liu, "Integrating multiple textural features for remote sensing image change detection," *Photogramm. Eng. Remote Sens.*, vol. 83, no. 2, pp. 109–121, Feb. 2017.

[25] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 3988–4004, Nov. 2018.

[26] T. F. R. Ribeiro, F. Silva, J. Moreira, and R. L. D. C. Costa, "Burned area semantic segmentation: A novel dataset and evaluation using convolutional networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 202, pp. 565–580, Aug. 2023.

[27] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.

[28] E. Maltezos, A. Doulamis, N. Doulamis, and C. Ioannidis, "Building extraction from LiDAR data applying deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 155–159, Jan. 2019.

[29] D. Chang, Q. Wang, J. Xie, J. Yang, and W. Xu, "Research on the extraction method of urban built-up areas with an improved night light index," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2505305.

[30] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, 2013.

[31] J. Yuan, S. S. Gleason, and A. M. Cheriyadat, "Systematic benchmarking of aerial image segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 6, pp. 1527–1531, Nov. 2013.

[32] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," 2018, *arXiv:1807.09532*.

[33] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3226–3229, doi: 10.1109/IGARSS.2017.8127684.

[34] I. Demir et al., "DeepGlobe 2018: A challenge to parse the Earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[35] X.-Y. Tong et al., "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[36] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[37] Z. Shao, K. Yang, and W. Zhou, "Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset," *Remote Sens.*, vol. 10, no. 6, p. 964, Jun. 2018.

[38] 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. 2018.

[39] M. Zhang, X. Hu, L. Zhao, Y. Lv, M. Luo, and S. Pang, "Learning dual multi-scale manifold ranking for semantic segmentation of high-resolution images," *Remote Sens.*, vol. 9, no. 5, p. 500, May 2017.

[40] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1524–1532.

[41] H. Alemohammad and K. Booth, "LandCoverNet: A global benchmark land cover classification training dataset," 2020, *arXiv:2012.03111*.

[42] S. Tian, A. Ma, Z. Zheng, and Y. Zhong, "Hi-UCD: A large-scale dataset for urban semantic change detection in remote sensing imagery," 2020, *arXiv:2011.03247*.

[43] K. Yang et al., "Semantic change detection with asymmetric Siamese networks," 2020, *arXiv:2010.05687*.

[44] S. P. Mohanty et al., "Deep learning for understanding satellite imagery: An experimental survey," *Frontiers Artif. Intell.*, vol. 3, Nov. 2020, Art. no. 534696.

[45] F. Fang, K. Wu, and D. Zheng, 2021, "A dataset of building instances of typical cities in China," *Science Data Bank*, doi: 10.11922/SCIENCEDB.00620.

[46] J. Wang, Z. Zheng, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021, pp. 1–16.

[47] R. Avudaiammal, P. Elaveni, S. Selvan, and V. Rajangam, "Extraction of buildings in urban area for surface area assessment from satellite imagery based on morphological building index using SVM classifier," *J. Indian Soc. Remote Sens.*, vol. 48, no. 9, pp. 1325–1344, Sep. 2020.

[48] 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.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[50] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[51] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[52] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[53] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

[54] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7036–7045.

[55] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 94–114, Apr. 2020.

[56] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[57] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3141–3149.

[58] L. Huang, Y. Yuan, J. Guo, C. Zhang, X. Chen, and J. Wang, "Interlaced sparse self-attention for semantic segmentation," 2019, *arXiv:1907.12273*.

[59] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.

[60] W. Yu et al., "MetaFormer is actually what you need for vision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10809–10819.

[61] 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. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–14.

[62] M. Li, K. M. de Beurs, A. Stein, and W. Bijker, "Incorporating open source data for Bayesian classification of urban land use from VHR stereo images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 11, pp. 4930–4943, Nov. 2017.

[63] X. Huang et al., "High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images," *Sci. Bull.*, vol. 65, no. 12, pp. 1039–1048, Jun. 2020.

[64] T. Liu and A. Abd-Elrahman, "Multi-view object-based classification of wetland land covers using unmanned aircraft system images," *Remote Sens. Environ.*, vol. 216, pp. 122–138, Oct. 2018.

[65] T. Liu, A. Abd-Elrahman, B. Dewitt, S. Smith, J. Morton, and V. L. Wilhelm, "Evaluating the potential of multi-view data extraction from small unmanned aerial systems (UASs) for object-based classification for wetland land covers," *GIScience Remote Sens.*, vol. 56, no. 1, pp. 130–159, Jan. 2019.

[66] H. Chen et al., "Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4967–4970.

[67] Y. Rao, "HorNet: Efficient high-order spatial interactions with recursive gated convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10353–10366.

[68] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "BiSeNet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, Nov. 2021.

[69] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "SegNeXt: Rethinking convolutional attention design for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 1140–1156.

[70] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11966–11976.

[71] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Renxiang Zuo received the B.Eng. degree in surveying and mapping engineering and the M.E. degree in architectural and civil engineering from the School of Geomatics Science and Technology, Nanjing Tech University, Nanjing, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in science and technology of remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include deep learning and remote sensing image processing.

Xin Huang (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2009.

He is currently a Full Professor with Wuhan University, Wuhan, China, where he teaches remote sensing and image interpretation. He is the Head of the Institute of Remote Sensing Information Processing (IRSIP), School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 200 peer-reviewed articles (SCI papers) in international journals. He was supported by the National Program for Support of Top-Notch Young Professionals in 2017, China National Science Fund for Excellent Young Scholars in 2015, and the New Century Excellent Talents in University from the Ministry of Education of China in 2011. His research interests include remote sensing image processing methods and applications.

Prof. Huang was a recipient of the Boeing Award for the Best Paper in Image Analysis and Interpretation from American Society for Photogrammetry and Remote Sensing (ASPRS) in 2010, the John I. Davidson President's Award from ASPRS in 2018, and the National Excellent Doctoral Dissertation Award of China in 2012. In 2011, he was recognized by the IEEE Geoscience and Remote Sensing Society (GRSS) as the Best Reviewer of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the Winner of the IEEE GRSS Data Fusion Contest in 2014 and 2021. He was an Associate Editor of *Photogrammetric Engineering and Remote Sensing* (2016–2019), IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (2014–2020), and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2018–2022). He has been serving as an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (since 2022). He has been an Editorial Board Member of *Remote Sensing of Environment* (since 2019).

Jiayi Li (Senior Member, IEEE) received the B.S. degree from Central South University, Changsha, China, in 2011, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She is currently an Associate Professor with the School of Remote Sensing and Information Engineering, Wuhan University. She has authored more than 60 peer-reviewed articles (Science Citation Index (SCI) articles) in international journals. Her research interests include hyperspectral imagery, sparse representation, computation vision and pattern recognition, and remote sensing images.

Dr. Li is a reviewer for more than 30 international journals, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, RSE, and ISPRS-J. She is the Young Editorial Board Member of *Geospatial-Information Science* (GSIS) and a Guest Editor of *Remote Sensing* (an open access journal from MDPI) and *Sustainability* (an open access journal from MDPI).

Xiaofeng Pan received the Ph.D. degree in environmental science from Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China, in 2010.

He now works with Shenzhen Ecological and Environmental Monitoring Center Station of Guangdong Province, Shenzhen, China, mainly engaged in monitoring the quality of the ecological environment and comprehensive analysis of data.