A Multisource Dynamic Fusion Network for Urban Functional Zone Identification on Remote Sensing, POI, and Building Footprint

Hangfeng Qiao, Huiping Jiang, Gang Yang, Faming Jing, Weiwei Sun, *Senior Member, IEEE*, Chenyang Lu, and Xiangchao Meng, *Senior Member, IEEE*

***Abstract*—Urban functional zones (UFZ) identification with remote sensing imagery (RSI) is attracting increasing attention in urban planning and resource allocation in urban areas, etc. The UFZ is a comprehensive unit comprising geographical, how to effectively integrate the RSI and points of interest (POI) with different physical and socioeconomic characteristics is important and promising. However, there are two challenges for the UFZ identification. On one hand, the UFZ is closely related to buildings, and most current methods lack an in-depth understanding of building semantics. Therefore, an efficient integration of building footprint (FT) data deserves further investigation. On the other hand, these RSI, POI, and FT data are heterogeneous; how to effectively leverage complementary information among these highly heterogeneous modalities to enhance the comprehensive understanding of urban. To solve the above challenges, this article introduces an end-to-end deep learning-based multisource dynamic fusion network for UFZ identification on RSI, POI, and FT. In the proposed method, an adaptive weight interactive fusion module is designed to comprehensively integrate the complementary information among the heterogeneous RSI, POI, and FT data sources. In addition, a multiscale feature focus module is proposed to extract multiscale image features and emphasize critical characteristics. This method was applied to UFZ classification in Ningbo, Zhejiang Province, China, and the experimental results demonstrate the competitive performance.**

***Index Terms*—Deep learning (DL), multimodal data fusion, remote sensing imagery (RSI), social sensing data, urban functional zone (UFZ).**

Manuscript received 4 March 2024; revised 28 April 2024; accepted 7 May 2024. Date of publication 31 May 2024; date of current version 5 June 2024. This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23D010001 and Grant LY22F010014, in part by National Natural Science Foundation of China under Grant 42171326 and Grant 42271340, in part by Ningbo Natural Science Foundation under Grant 2022J076, in part by 2025 Science and Technology Major Project of Ningbo City under Grant 2021Z107 and Grant 2022Z032, and in part by Zhejiang Province "Pioneering Soldier" and "Leading Goose" Research Project under Grant 2023C01027. *(Corresponding author: Xiangchao Meng.)*

Hangfeng Qiao, Chenyang Lu, and Xiangchao Meng are with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China (e-mail: 2211100145@nbu.edu.cn; luchenyang1@nbu.edu.cn; mengxiangchao@nbu.edu.cn).

Huiping Jiang is with the Key Laboratory of Regional Sustainable Development Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China (e-mail: jianghp@igsnrr.ac.cn).

Gang Yang and Weiwei Sun are with the Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo 315211, China (e-mail: sunweiwei@nbu.edu.cn).

Faming Jing is with the Ningbo Institute of Surveying Mapping and Remote Sensing, Ningbo 315211, China (e-mail: 284639904@qq.com).

Digital Object Identifier 10.1109/JSTARS.2024.3404094

I. Introduction

URBAN functional zone (UFZ) is the basic unit of urban planning, serving as geographical space for performing urban functions and aggregating social resources. A fine-scale and accurate classification of the UFZ using remote sensing plays a crucial role in sustainable development, effective planning, and resource allocation in urban areas [1], [2]. The UFZ is a comprehensive concept comprising geographical, social, and economic attributes. Therefore, an effective integration of multisource data, such as remote sensing and points of interest (POI), with different physical and socioeconomic characteristics is crucial but challenging for the classification of the UFZ [3].

In existing studies, most of UFZ classification methods can be concluded into two categories: single-modal-based methods and multimodal-based methods.

A. Single-Modal-Based UFZ Classifying Methods

The high-resolution remote sensing imagery (RSI) is generally favored in the single-modal-based UFZ classification methods due to the large scene geographical imaging [4], [5], [6] and the precise identification of different land cover types in dividing urban areas into distinct UFZ [7], [8], [9]. Currently, existing methods for identifying UFZ using RSI can be broadly categorized into two types: 1) scene-based methods and 2) object-based methods. The scene-based methods primarily rely on overall features and landscape information in RSI, making them suitable for regions with significant variations in land cover. Zhang and Du [10] utilized object features and categories to generate low-level semantic information, employing the latent Dirichlet allocation method to represent image scenes as distinct UFZ. Huang et al. [11] introduced the bag of visual words model, treating UFZ as a series of "visual words" and using the visual features represented by these words to encode UFZ characteristics. However, scene-based methods face challenges in recognizing small-scale features in complex urban environments. In contrast, object-based methods can more precisely capture specific land cover information, enabling accurate delineation of UFZ. Zhou et al. [12] initially divided very high-resolution remote sensing images into super objects corresponding to UFZ. They employed a random point generation algorithm to generate voting points and then applied a majority voting strategy to

assign functional attributes to UFZ. Building upon Zhou's work, Du et al. [13] proposed a pixelwise prediction multiscale semantic segmentation network by identifying and merging objects with similar functions. However, these methods solely relying on RSI generally generate suboptimal identification performance [14]. This is attributed to several factors: 1) RSI can provide spatial and spectral information but often struggles to offer direct information about socioeconomic attributes; and 2) different functional zones are often intertwined, making it challenging for RSI to accurately differentiate these mixed land cover types without social sensing data.

With the rapid development of multimedia information communication technology, an increasing number of researchers are utilizing social sensing big data for UFZ classification. Unlike remote sensing data, social sensing data is generated in real time, has a wide range of sources, and can more accurately reflect urban socioeconomic functions and human spatial activities [15], [16]. Common social sensing data include mobile phone signal data [17], [18], POI [19], [20], [21], taxi trajectory data [22], and geotagged leveraging traffic interaction information extracted from taxi GPS trajectory data, and proposed a geographically convolutional neural network model with geographic semantic embedding representation for UFZ classification at the road segment level. Guilin et al. [25] proposed a UFZ classification method based on mobile signaling data, utilizing the feature relationships associated with the characteristics of UFZ corresponding to mobile user calling and movement behaviors. These social sensing data are easily accessible, among which POI aligns particularly well with urban land-use types [26], [27]. Therefore, how to effectively utilize POI data for identifying UFZ has become one of the research hotspots [28]. POI-based UFZ classification methods can be mainly categorized into three types:

1) Methods based on clustering analysis [29], this approach aggregates POI of similar types into clusters, forming distinct UFZ for each cluster.
2) Rule-based and weighted methods [30], this approach devises a set of rules and weights to score different types of POI, subsequently delineating UFZ based on these scores. However, this method is constrained by the subjectivity of expert scoring.
3) Methods based on deep learning (DL). With the continuous improvement of DL theory, particularly in their outstanding feature extraction capabilities [31], [32], DL techniques have found extensive applications in UFZ classification tasks and they often outperform the first two approaches. For instance, Lu et al. [33] transformed discrete POI into hierarchical distance heatmaps, enabling the application of convolutional neural network (CNNs) to POI. Huang et al. [34] trained their model by maximizing mutual information between POI–region–urban hierarchy, facilitating unsupervised learning of urban region representations.

Several studies have indicated that each type of geospatial data possesses unique advantages [21], [35], [36], and buildings stand out as one of the most crucial elements influencing the urban structure and city planning [37]. Building footprints (FT) serve as a significant data source characterizing population aggregation, energy consumption, and regional development [38], [39], [40]. Accurate and timely information on FT is essential data support for sustainable development and urban–rural planning [41]. Historically, the geographical boundaries of buildings in UFZ have been vague and discrete. FT data not only provide precise location and shape information for clusters of buildings but also assist in identifying structural features of different buildings (such as height, purpose, and type). This aids in associating UFZ with specific buildings or groups of buildings, enhancing spatial information. However, previous article has primarily focused on scene recognition and classification, paying comparatively less attention to the semantic aspects of buildings in UFZ analysis [42]. In-depth investigations into the three-dimensional (3-D) structure of urban areas remain limited. Currently, there is few dedicated methods for UFZ identification using FT exclusively. Instead, it is employed as supplementary data combined with RSI to achieve favorable recognition results. In mixed-functional zones, especially where buildings exhibit significant physical differences, FT may play a crucial role in distinguishing between structures [43].

B. *Multimodal-Based UFZ Classifying Methods*

The types of UFZ are closely related to human socioeconomic activities, and these pieces of information are not provided by remote sensing images solely. Therefore, more and more researchers are considering incorporating social sensing data to assist in identification. Integration methods of RSI and social sensing data can be categorized into two types: 1) feature-level fusion [44], this approach extracts features from different data sources and integrates them into a common feature space; and 2) decision-level fusion [45], [46], each mode independently generates a preliminary UFZ classification result, and the results from different modes are then combined to form an final results. Feature-level fusion methods are widely applied because they consider the correlations and interaction effects between modalities. For instance, Zhang et al. [14] extracted spatial relationship features between different targets, synthesized very high spatial resolution (VHSR) images, multilevel road networks, and POI data, proposing a distance-weighted graph attention model. Guo et al. [47] combined high spatial resolution (HSR) images with POI data, embedded hierarchical group convolution modules, and attention mechanisms in the network. They utilized a mask layer for deep feature filtering of irregular UFZ, automatically classifying UFZ blocks of different sizes and irregular shapes. However, feature-level fusion methods tend to introduce redundant information, while decision-level fusion, as each modality independently produces decisions, is more suitable for handling heterogeneous data sources. Yet, this method overlooks the potential correlations between modalities. For example, Bai et al. [19], based on VHSR images and POI data, separately learned regional vector embeddings for geographical and socioeconomic attributes. They proposed an unsupervised multimodal geographic representation learning framework. Bao et al. [48], by combining remote sensing and POI data, introduced the deep edge feature map into segmentation classification, enhancing the recognition of building boundaries.

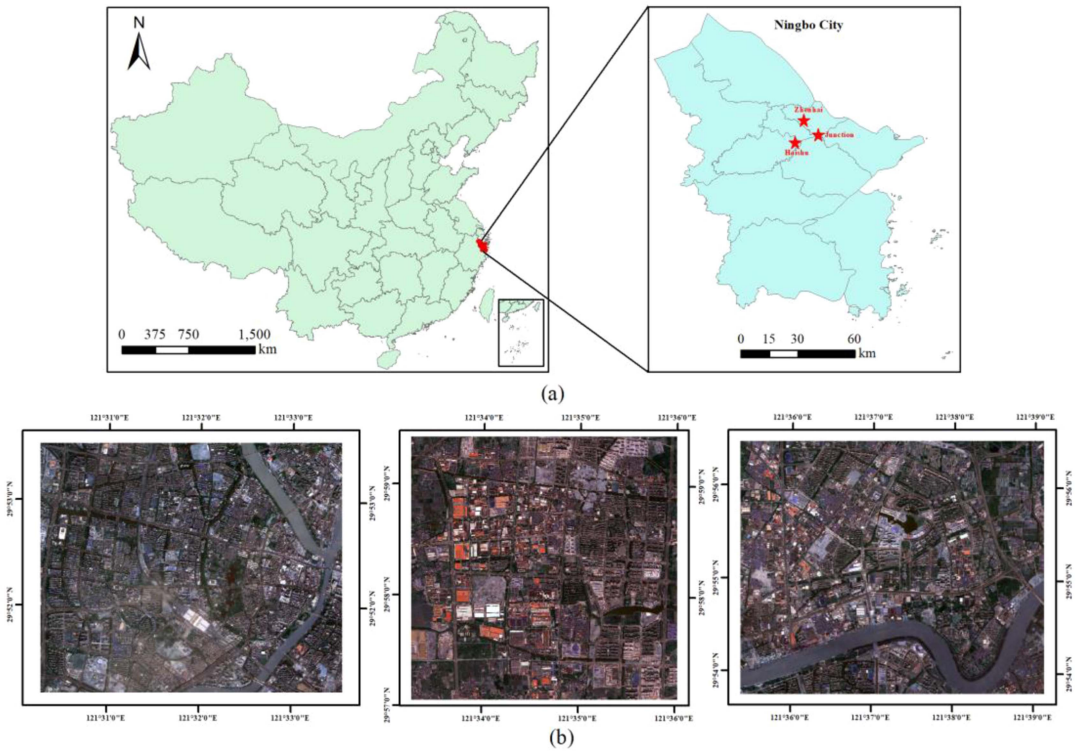


Fig. 1. Illustration of the study areas. (a) Geographic locations of study area. (b) Three representative study areas, the first is Haishu district, the second is Zhenhai district, and the third is Jiangbei–Zhenhai–Yinzhou intersection area.

C. Challenges and Contributions

Based on the above analyses of UFZ classification, we put forward the following two challenges to be solved:

Challenge 1. How to compensate for the lack of elevation and building semantic information in UFZ classification: While remote sensing images provide rich information, such as texture and spectral data, and POI offers insights into economic activities, there is still a gap in exploring the 3-D structure and boundary information of buildings. Previous articles have not thoroughly explored the interrelation between UFZ and specific buildings, and the representation of building semantics has been somewhat limited. Consequently, the efficient integration of building FT data deserves further investigation.

Challenge 2. How to effectively leverage complementary information among heterogeneous modalities: Due to the inherent modality differences between RSI and social sensing data, where remote sensing images provide spatial and spectral surface data, building FTs and POI are based on location-based raster vector data, and existing methods not only overlook modality differences by directly integrating them but also fail to consider the semantic distribution imbalance between remote sensing data and social sensing data. This makes it challenging to fully exploit complementary information among different modalities by considering modality differences.

In this article, considering the aforementioned challenges, we propose a novel end-to-end multimodal fusion network. The main contributions of this article can be summarized as follows.

1) To address the *Challenge 1*, we introduce a multisource dynamic fusion network (MSDFN) suitable for fine-grained UFZ identifying. This network leverages object features from remote sensing images, socioeconomic information from POI, and elevation information from building FTs, enabling a clearer identification of regional boundaries and demonstrating effective recognition in complex terrains, such as fragmented landscapes.
2) To overcome the *Challenge 2*, we develop an adaptive weight interactive fusion module (AW-IFM) that dynamically allocates weights based on the inherent characteristics of each respective dataset. This module robustly and efficiently interacts with feature information among different modalities, thereby narrowing modality differences.

The rest of this article is organized as follows. Section II describes the study area and dataset. Section III introduces the general workflow and key components of the proposed method. Section IV provides the experimental results and analysis. Sections V presents the discussions. Finally, Section VI concludes this article.

II. Study Areas and Datasets

A. Study Areas

In this article, we selected three typical areas within Ningbo city in Zhejiang province, China, as the research area, with geographical coordinates ranging from 28°51' to 30°33'N and 120°55' to 122°16'E [Fig. 1(a)]. Located on the southeast coast

TABLE I
SPECIFICS OF RSI, POI, AND FT

Data type	Study area	Amount of data (RSI is dimension size, POI and FT are record counts)	Sampling source
RSI	Haishu	5274 × 4368 × 4	Gaofen-2 satellite (https://data.cresda.cn/#/2dMap)
	Zhenhai	4426 × 4384 × 4	
	Jiangbei–Zhenhai–Yinzhou intersection area	6048 × 5028 × 4	
POI	Haishu	30710	Amap API (https://lbs.amap.com/tools/picker)
	Zhenhai	5097	
	Jiangbei–Zhenhai–Yinzhou intersection area	6161	
FT	Haishu	8752	Bigemap (http://www.bigemap.com)
	Zhenhai	2063	
	Jiangbei–Zhenhai–Yinzhou intersection area	4873	

of China, Ningbo serves as a crucial port city and one of the economic centers of Zhejiang province. As of the end of 2022, Ningbo covers a total area of 9816 square kilometers, with a permanent population of 9.618 million and an urbanization rate of 78.9%. The region exhibits a relatively diversified economic structure, including well-developed sectors in agriculture, industry, and commerce.

For this article, we focused on three specific areas within Ningbo [Fig. 1(b)]. The first study area is predominantly situated in the Haishu district, representing the most urbanized region in Ningbo with high-rise buildings, shopping centers, dining establishments, and cultural and entertainment facilities, making it a crucial hub for urban life. The second study area is positioned in the core zone of Zhenhai district, undergoing active urbanization with a predominant industrial focus, featuring relatively dense building types, and exhibiting significant differences among various UFZ. The third study area is located at the intersection of Jiangbei district, Zhenhai district, and Yinzhou district, characterized by a diverse and scattered distribution of urban functions.

B. Datasets

This article utilized following three types of data: high-resolution RSI, POI, and building FT data, and their specific information is shown in Table I.

1) The RSI was captured by the Gaofen-2 satellite, consisting of 1-m resolution panchromatic images and 4-m resolution multispectral images with four bands. Through radiometric calibration, atmospheric correction, and orthorectification, the panchromatic and multispectral images were fused to generate a final 1-m resolution multispectral image.
2) POI data were obtained from the Amap API and contained information about specific geographical locations or places, such as names, categories, geographic coordinates, addresses, descriptions, etc. These data reveal the types of activities conducted by people in specific places and can complement socioeconomic information. Due to the absence of clear classification standards, we reclassified POI into 10 categories based on the "code for classification of urban and rural land use and planning standards of development land GB50137" issued by the Ministry of Housing and Urban–Rural Development of the People's Republic of China. These categories include public leisure, medical services, transportation facilities, commerce, education and culture, residential accommodation, companies and enterprises, government land, scenic spots, and parking lots.
3) FT data were sourced from Bigemap, including information about the height, length, and area of the buildings. These physical properties can reflect the functional attributes of an area and highlight differences between various land-use types. All the RSI, POI, and FT data in the three study areas are acquired in 2019.

C. UFZ Categories

Since there is no uniform rule of UFZ classification, similar to the POI classification, we also refer to the Chinese national standard "code for classification of urban and rural land use and planning standards of development land GB50137" and define 13 types of UFZ, including area to be developed, commercial zone, educational facility, residential neighborhood, mixed-use residential and commercial zone, industrial zone, tourist attraction, public leisure area, government facility, medical facility, green space, water body, and transportation infrastructure. Their specific definitions are provided in Table II. It is worth noting that Jiangbei–Zhenhai–Yinzhou intersection dataset lacks public leisure area and government facility, resulting in only 11 types of UFZ.

III. METHODOLOGY

In this section, we propose a multimodal fusion framework tailored for the UFZ classification by integration of RSI, POI, and FT data. The framework is intricately divided into four key components: preprocessing, feature extraction, feature fusion, and feature refinement and regression, as illustrated in Fig. 2. In the proposed method, the RSI, POI, and FT are encoded into a shared domain after preprocessing. Subsequently, we introduce an AW-IFM to dynamically assign weights to the data based on their inherent characteristics. The integration process is gradual, involving FT and POI, with RSI serving as the primary focal point. To enhance the feature representation capability across multiple scales, we primarily design the channel–dilated–spatial joint module (CDS-JM) and multiscale feature focus module (MS-FFM). Finally, UFZ classification is executed through a combination of convolutional and fully connected layers. It is worth mentioning that, to address the issue of category imbalance in UFZ classifying, our framework supplements the conventional cross-entropy loss with the dice loss.

A. Data Preprocessing

This article employs a traditional patch-based approach, uniformly dividing each dataset into fixed-size blocks. Leveraging

TABLE II
SPECIFIC DEFINITIONS OF EACH UFZ TYPE ON CHINESE NATIONAL STANDARD "CODE FOR CLASSIFICATION OF URBAN AND RURAL LAND USE AND PLANNING STANDARDS OF DEVELOPMENT LAND GB50137"

Types	Definitions
Area to be developed	Undeveloped area, primarily consisting of buildings under construction but may also include open spaces and standalone parking lots.
Commercial zone	Commercial zone, encompassing retail, dining, and entertainment activities.
Educational facility	Educational institutions and research facilities.
Residential neighborhood	General residential housing area, characterized by predominantly low-rise buildings, forming the primary residential communities.
Mixed-use residential and commercial zone	Mixed-use residential and commercial zone, allowing for a blend of commercial and residential activities, often featuring high-rise buildings.
Industrial zone	Industrial zone, comprising clusters of companies and factories.
Tourist attraction	Tourist attractions, drawing visitors and sightseers.
Public leisure area	Recreational and entertainment facilities, including libraries, sports arenas, and public squares.
Government facility	Government agencies and government-owned facilities.
Medical facility	Medical institutions, primarily hospitals.
Green space	Urban green spaces and farmlands.
Water body	Water bodies, encompassing rivers, streams, and lakes.
Transportation infrastructure	Transportation infrastructure, such as roads, bridges, and transportation hubs.

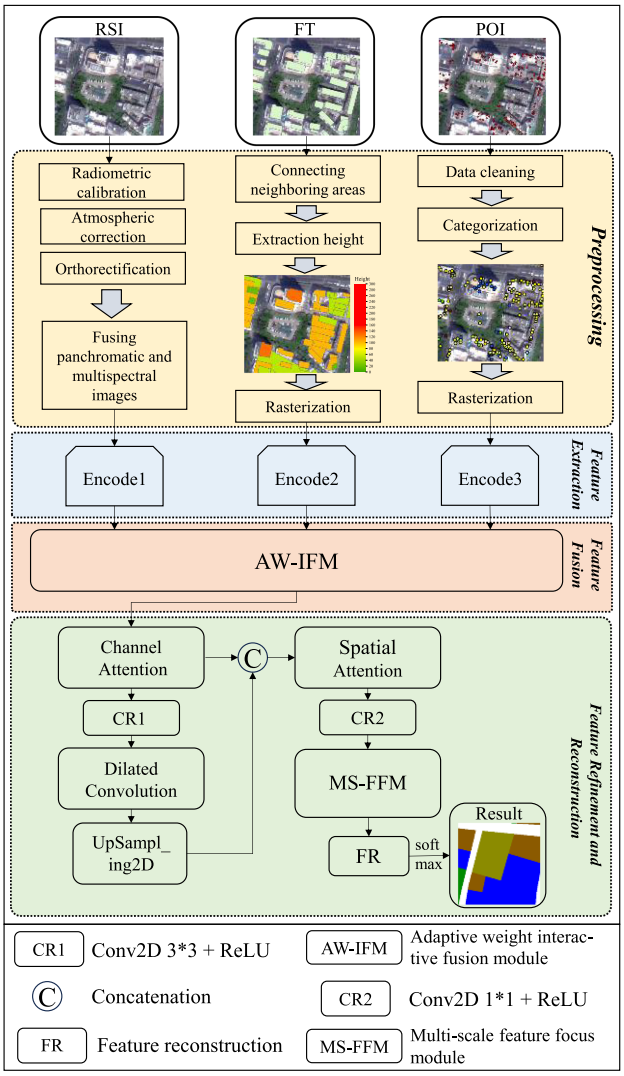


Fig. 2. Overview of the proposed MSDFN.

an end-to-end network, the model requires only three preprocessed data inputs to generate a classification result map, making it convenient and easily implementable. In this article, we consider each uniformly sized image as a basic unit for UFZ, and the preprocessing procedures for the three types of data, namely, RSI, building FT, and POI, are illustrated in Fig. 2.

For RSI data, after preprocessing, we directly extract both their low-level and high-level features without the need for additional processing. Regarding the FT data, we utilize ArcMap 10.8 to connect adjacent building FT blocks into connected regions, followed by rasterization to generate output images. We specifically select the physical attribute of height from FT data, which can compensate for the limitations of extracting features in 2-D and analyzing the floor-level patterns of building structures in UFZ, given the significant variation in building heights among different types. Concerning the POI data, we first filter out invalid data, such as public toilets and kiosks, categorize them as per Section II, and then rasterize the output images. Finally, the three types of data are collectively input into our MSDFN for UFZ mapping.

B. Feature Extraction

This article employs CNN to extract feature information from multimodal data. CNN is capable of extracting various features when processing RSI, including low-level features, such as edges, textures, colors, and shapes, as well as high-level semantic features, such as the overall structure of buildings, relative positions, and hierarchical relationships, even capturing local information in complex scenes [49], [50]. However, POI and FT are both vector data, with a structure different from that of RSI, making direct fusion challenging. To address this issue, we perform previous data preprocessing and transform POI and FT data into image format using ArcMap. Unlike other methods that straightforwardly concatenate nonremote sensing data, we consider the uniqueness of each data source, allowing CNN to independently extract their feature information. Specifically, we

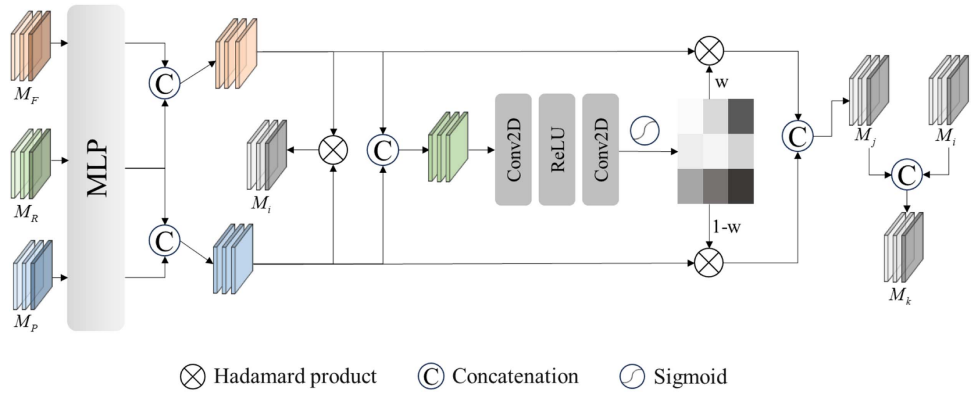


Fig. 3. Architectural schematic diagram of AW-IFM.

extract the height information from the FT data. For POI data, we employ a majority voting strategy, selecting the category with the highest frequency within a region as the POI category for that region. The calculation method is as follows:

$$T_i = \frac{\frac{n_i}{N_i}}{\sum_1^j \frac{n_i}{N_i}} (i = 1, 2, 3, \ldots, j) \tag{1}$$

where i represents the type of POI, j represents the total number of POI categories, n_i represents the number of the ith type of POI in a specific region, N_i represents the total number of the ith type of POI in the entire image, and T_i represents the ratio of the frequency density of the ith type of POI to the frequency density of all types of POI. When converting the processed POI data features to raster format, the selected pixel size for each study area is uniformly set to 80. It is worth noting that, due to the varying resolutions of each RSI and the corresponding differences in UFZ sizes, the output POI image is a raster image with the same resolution as the RSI. The same principle applies to FT data.

In the feature extraction module of Fig. 2, we employ an encoder architecture similar to VGG16 [35], [51]. This encoder structure is utilized for extracting features from the three multisource and multimodal data types: RSI, POI, and FT. The architecture consists of four stages denoted as "Encode," each composed of a 3×3 convolutional layer and an ReLU layer. Additionally, between the convolutional and ReLU layers in the second and fourth stages, we include an additional max-pooling layer with a pooling window size of 2×2 to facilitate the extraction of spatial features. After extracting feature information from the three data types, we will proceed to interactively fuse them in the subsequent steps.

C. Feature Fusion

When utilizing FT data for the recognition of simple-structured functional areas, such as scenic spots, urban green spaces, open spaces, and transportation, there may be some data redundancy introduced. However, in identifying areas composed of significantly different building types, such as residential, commercial, and industrial, enhancing the corresponding physical features can improve recognition effectiveness [43]. Similarly, although POI data can provide more detailed geographic entity information in scenarios with high demands for urban activity diversity, in situations requiring consideration of building structures and micro features, it is necessary to combine other data sources to enhance accuracy. RSI serves as the bridge between these two types of data. Therefore, inspired by nonlocal attention [52], we introduce corresponding weights to these data to facilitate the positive impact of effective features on fusion and suppress the influence of irrelevant features on fusion results. Our proposed adaptive weight interaction fusion module is illustrated in Fig. 3.

Specifically, AW-IFM first combines the feature maps $M_R \in \mathbb{R}^{H\times W\times C}$ of RSI with the feature maps $M_F \in \mathbb{R}^{H\times W\times C}$ of FT and $M_P \in \mathbb{R}^{H\times W\times C}$ of POI in pairs, where H, W, and C denote the height, width, and channel of the feature maps. Then, each combination undergoes an interaction through a shared multilayer perceptron (MLP) layer, represented as follows:

$$F_F, F_R, F_P = \text{MLP}(M_F, M_R, M_P) \tag{2}$$

where F_F, F_R, and F_P represent the output of their respective feature maps. Taking RSI data as the main component and FT and POI data as supplementary, we progressively fuse them to obtain fusion results $M_{RF} \in \mathbb{R}^{H\times W\times 2C}$ and $M_{RP} \in \mathbb{R}^{H\times W\times 2C}$. This approach ensures that each fusion step fully utilizes the information from each data source, allowing the model to better learn the multimodal features of the data. Subsequently, adaptive weights are assigned to the two fusion results, relying on the respective characteristics of the data for weight generation. Similarly, the three types of multisource and multimodal data undergo an MLP layer, generating a single fusion feature through a step-by-step fusion strategy. Unlike directly fusing the three data sources, our approach considers the correlation between the data, preserving not only the characteristics of the source data but also reducing fusion complexity.

Next, for the fused feature map $M_{RFP} \in \mathbb{R}^{H\times W\times 4C}$, we employ a 1×1 convolutional layer to learn the correlation between the previously obtained M_{RF} and M_{RP} features. The obtained correlation coefficients are then normalized using the sigmoid function to generate weight matrices $w \in [0, 1]$ and $w' = 1 - w \in [0, 1]$. The values of w are applied to M_{RF}, and those of w' are applied to M_{RP}. Subsequently, the final output feature map is calculated through a weighted summation,

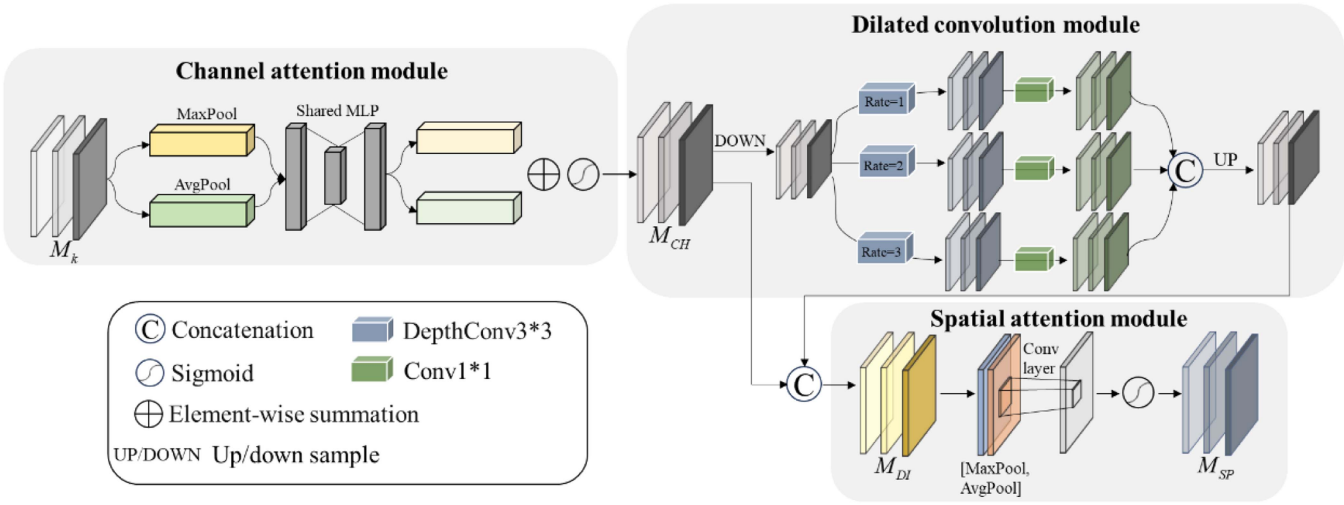


Fig. 4. Channel–dilated–spatial joint module.

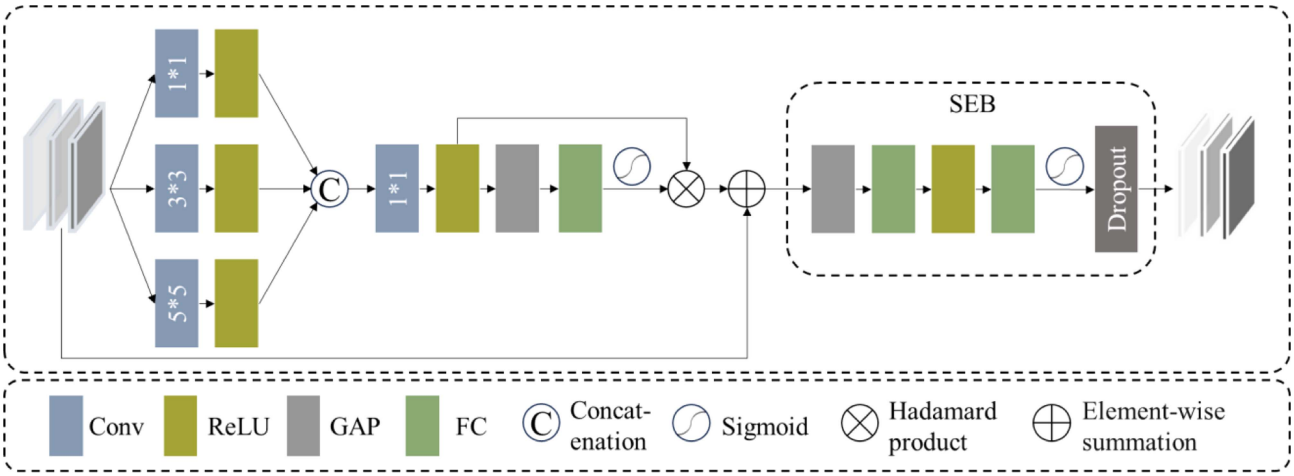


Fig. 5. Multiscale feature focus module (MS-FFM).

expressed as follows:

$$\begin{cases} M_i = M_{RF} \otimes M_{RP} \\ M_j = concat(w \otimes M_{RF}, w' \otimes M_{RP}) \\ M_k = concat(M_i, M_j) \end{cases} \tag{3}$$

where $\otimes$ denotes the Hadamard product, $M_i \in \mathbb{R}^{H\times W\times 2C}$ and $M_j \in \mathbb{R}^{H\times W\times 4C}$ represent the intermediate fusion features respectively, and $M_k \in \mathbb{R}^{H\times W\times 6C}$ signifies the final output of the fusion feature result. Overlaying features from different fusion strategies allows adaptation to diverse data and scenarios, enhancing robustness. Subsequently, the obtained fusion feature map is propagated to the next stage for further processing.

D. Feature Refinement and Feature Reconstruction

For the fused intermediate feature map, we introduce dilated convolutions and attention mechanisms to enhance feature representation capabilities in both spatial and channel dimensions, to optimize network performance. In the fusion of map channels, different channels contribute differently to UFZ classification. To capture channel features accurately, we introduce a channel attention mechanism to allocate weights to each channel. Furthermore, considering that UFZ is composed of land use in multiple regions, we introduce a spatial attention mechanism in its spatial dimension to adaptively adjust features, enhancing sensitivity to the internal details of UFZ. Moreover, building upon the traditional convolutional block attention module (CBAM) module [53], we introduce dilated convolutions and an MS-FFM, enabling the model to better handle different scales of geographic structures in UFZ, expanding the receptive field, facilitating the understanding of the overall structure and patterns of land use, and retaining finer local features. This is crucial for UFZ classification, as UFZ encompasses rich geographical details typically, such as building outlines and road networks.

We categorize all the modules mentioned above into the CDS-JM (Fig. 4) and the MS-FFM (Fig. 5). The explanation of the CDS-JM is as follows. As shown in Fig. 4, for the result feature map M_k obtained from feature fusion, we initially

pass it through a channel attention module (CAM). The CAM effectively weights different channels and highlights channel information that is meaningful for the UFZ classification task.

After passing through the CAM, we employ an improved dilated convolutional layer to enhance computational efficiency and effectiveness. It is achieved by utilizing depthwise separable convolution, which decomposes the standard convolution into depthwise convolution and pointwise convolution. Initially, the input feature M_{CH} undergoes a 3×3 convolutional layer, generating output that encompasses information from different receptive fields of the input features. By employing depthwise separable convolution, the convolutional layer performs downsampling, reducing the spatial dimensions of the feature map while increasing the number of channels. Subsequently, three depthwise separable convolutional layers are utilized, each consisting of two sublayers. The first sublayer is a depthwise separable convolutional layer with a 3×3 convolutional kernel, and dilation rates of 1, 2, and 3, implementing dilated convolutions on the input feature. The second sublayer is a 1×1 convolutional layer with eight convolutional kernels, responsible for channel information fusion. This approach decomposes each depthwise separable convolution into depthwise and pointwise convolutions, effectively enhancing the computational efficiency of the model. Following this, the output of the three depthwise separable convolutional layers is merged using a bilinear interpolation upsampling layer, restoring the spatial dimensions of the feature map to the original size. Finally, by concatenating the original input feature M_{CH} with the upsampled feature, we obtain the module's output M_{DI}.

Following the dilated convolution, we subject M_{DI} to the spatial attention module. It makes the shape of feature map transform to $\mathbb{R}^{H \times W \times 8C}$, and this transformed feature map is subsequently fed into the MS-FFM (Fig. 5).

Even within the same type of UFZ, there exists significant variation in scale, and this difference is even more pronounced among different types of UFZ. Considering this aspect, we propose a novel MS-FFM designed to extract critical feature information across various scales in an image, significantly enhancing the network's feature representation capabilities and optimizing its perceptual and discriminative performance toward UFZ. The module initiates by employing convolutional kernels of sizes 1×1, 3×3, and 5×5 to extract features at multiple scales. By synthesizing scale-specific features, we obtain an extended representation. Global average pooling and attention weighting are applied to capture global information and fuse multiscale features, forming the preliminary output. Since feature maps from the same level share some commonalities and feature block weights need to be allocated, we introduce a spatial excitation block (SEB) at the end of the module [54]. We capture global information through global average pooling, then introduce an activation function and attention weights through a fully connected layer for spatial excitation. Moreover, we incorporate dropout regularization to enhance robustness, selectively emphasizing crucial features of the module.

Eventually, the specific operations for feature reconstruction of FR in Fig. 2 are illustrated as follows:

$$M_{\text{out}} = \text{Conv}(\text{ReLU}(\text{GAP}(\text{Conv}(\text{ReLU}(\text{Conv}(M_{\text{in}})))))) \quad (4)$$

where GAP denotes global average pooling, and M_{in} and M_{out} represent the input and output of FR, respectively. After that, all feature maps are transformed into 1-D feature vectors. These vectors are then fed into a softmax function for UFZ recognition and classification.

E. Loss Functions

Due to significant variations in regions at different urbanization stages, UFZ categories exhibit extreme imbalance. In addition to utilizing the traditional cross-entropy loss function $\mathcal{L}_{\text{ce}}$ to constrain network training, we introduce a dice loss function $\mathcal{L}_{\text{dice}}$ to enhance pixel-level classification accuracy. To the best of our knowledge, there has been no prior research applying the $\mathcal{L}_{\text{dice}}$ to UFZ classification in the current field. Furthermore, to mitigate the risk of overfitting, we employ L2-norm regularization on the network's weight parameters. The overall loss function for the network is given by

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \beta \mathcal{L}_2 + \lambda \mathcal{L}_{\text{dice}} \quad (5)$$

where β is a constant with values of 0.001 and λ is the balancing factor for the three losses, its value will be determined in the hyperparameter analysis in Section V. By jointly utilizing these losses, the entire segmentation network can learn complementary knowledge. The specific calculation formulas for $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{dice}}$ are as follows:

$$\mathcal{L}_{\text{ce}} = -\sum_{i} \sum_{k=1}^{C} y_{i,k} \log(p_{i,k}) \quad (6)$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\,|Y_{\text{pr}} \cap Y_{\text{re}}|}{|Y_{\text{pr}}| + |Y_{\text{re}}|}. \quad (7)$$

In $\mathcal{L}_{ce}$, $y_{i,k}$ and $p_{i,k}$ represent the ground truth label and predicted probability value for class k of the ith sample, respectively, and C is the total number of categories in UFZ. In $\mathcal{L}_{\text{dice}}$, the predicted result Y_{pr} and the true label Y_{re} are both probability distributions after the softmax operation, where each value at a position indicates the model's predicted probability for the corresponding class. In the calculation of $\mathcal{L}_{\text{dice}}$, these probability distributions are treated as binary masks, considering positions with probabilities greater than a certain threshold as positive class and others as negative class. Then, elementwise comparisons are performed on the binary masks, obtaining the intersection by elementwise multiplication and the union by elementwise addition.

IV. Experimental Results and Analysis

A. Experimental Settings

All the network frameworks in the experiments are implemented using TensorFlow 2.5.0 and are run on an Intel(R) Core(TM) i9-10920X CPU with an NVIDIA GeForce RTX 3090 GPU. Since the proposed method is based on CNN patches, each image in the dataset is cropped into blocks of the same size for input into the network. Considering the varied sizes and shapes of UFZ in different datasets, we employ a multicropping strategy. Specifically, for Haishu, Zhenhai, and Jiangbei–Zhenhai–Yinzhou intersection area, we crop the images into 54

patches of size 586 × 728, 81 patches of size 491 × 487, and 54 patches of size 672 × 838, respectively. Subsequently, we randomly split the data from each patch image into a 4:1 ratio for training and testing sets. Since this article uses an end-to-end network, pretraining of model parameters is unnecessary. The testing results are then restored to their original positions, and the metrics are calculated by taking the average over the total number. The number of epochs for all datasets is set to 100, and training continues until the training loss converged. The learning rate for this article is set to 0.001, utilizing the Adam optimizer, with a combination of cross-entropy loss, regularization loss for weight parameters, and class imbalance loss.

B. Evaluation Metrics and Compared Methods

To evaluate the classification results of UFZ, we utilize four metrics: overall accuracy (OA), kappa coefficient (KAPPA), average $F1$ score (Ave_F1), and mean intersection over union (MIoU). These metrics quantify the model's classification performance comprehensively, considering its performance across different categories and providing a comprehensive and objective evaluation. OA measures the OA of the model across all categories, KAPPA assesses the consistency between the classification results and random classification, Ave_F1 averages precision and recall for each category, and MIoU measures the relationship between the intersection and union of predicted results and true labels, reflecting the model's performance at the pixel level. Their specific formulas are as follows.

(1) OA:
$$p_0 = \sum_{i=1}^{n} \frac{x_{ii}}{N} \tag{8}$$

where x_{ii} denotes the element of the ith row and the jth column in the confusion matrix, n is the number of classes, and N is the total number of all the samples.

(2) KAPPA:
$$K = \frac{p_0 - p_e}{1 - p_e} \tag{9}$$

where $p_e = \sum_{i=1}^{n} (\sum_{j=1}^{n} x_{i,j} \sum_{j=1}^{n} x_{j,i})/N^2$, and $x_{i,j}$ denotes the element of ith row and jth column in the confusion matrix.

(3) F1 score:
$$F1_i = \frac{2p_i r_i}{p_i + r_i} \tag{10}$$

where p_i and r_i are the precision and recall score of class i, respectively, and $p_i = x_{ii}/\sum_{j=1}^{n} x_{ij}, r_i = x_{ii}/\sum_{j=1}^{n} x_{ji}$. $F1_i$ metric evaluates the classification performance for a specific class i. While the Ave_$F1$ is the mean of all the $F1$ scores across different categories, providing a comprehensive measure of the overall classification results for all n classes

Ave_F1:
$$\overline{F1} = \frac{1}{n} \sum_{i=1}^{n} F1_i. \tag{11}$$

(4) MIoU:
$$\text{MIoU} = \frac{1}{n} \sum_{i=1}^{n} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{12}$$

where n represents the total number of categories, TP_i denotes the true positive count for class i, FP_i represents the false positive count for class i, and FN_i signifies the false negative count for class i. MIoU calculates the MIoU for each category, where IoU is utilized to assess the degree of overlap between the pixel predictions of the model and the ground truth for each class.

In order to assess the performance of our method, we select following six representative segmentation methods based on DL for comparative experiments.

1) *Random forest (RF):* It is an ensemble learning method that constructs multiple decision trees and makes predictions based on their collective voting, enabling classification and regression analysis of data.
2) *ResNet:* It employs ResNet18 to extract features, establishing "skip connections" between preceding and subsequent layers, and accomplishing segmentation through stacked convolution and pooling operations.
3) *Fully convolutional network (FCN):* It conducts pixel-level classification on images, transforming traditional fully connected layers in CNNs into convolutional layers.
4) *Multimodal fusion network (MFN) [43]:* This is a dual-branch convolutional network that integrates features from two modalities, introducing both channel attention and spatial attention.
5) *LANet [55]:* It introduces a patch attention module based on local attention to enhance the embedding of contextual information and enriches the semantic information of low-level features by embedding local focuses of high-level features.
6) *MDL–RS–CNNs [56]:* This is a multimodal fusion network performing type recognition through cross-modal fusion.
7) *UisNet [57]:* It utilizes transformer-based modules to receive multimodal data, making it a fine-grained semantic segmentation approach.

Noteworthy, all comparison methods mentioned above were provided with data inputs as described in the original text.

C. Experimental Results

1) Experimental Results in Haishu Area: In this section, we conduct a comprehensive comparison of six baseline methods for UFZ classification on Haishu dataset (Fig. 6 and Table III). This region is predominantly composed of commercial residences, encompassing a diverse range of UFZ categories. The functional attribute categories of UFZ correspond to those in Table II, comprising 13 distinct categories: area to be developed (Are.), commercial zone (Com.), educational facility (Edu.), residential neighborhood (Res.), mixed-use residential and commercial zone (Mix.), industrial zone (Ind.), tourist attraction (Tou.), public leisure area (Pub.), government facility (Gov.), medical facility (Med.), green space (Gre.), water body (Wat.), and transportation infrastructure (Tra.). As mentioned in Section II, Haishu is positioned in the city center, characterized by thriving commercial activities and the convergence of various types of UFZ. Our analysis is as follows.

Firs, methods based on residual blocks or dense blocks (RF and ResNet) exhibited the poorest performance due to their

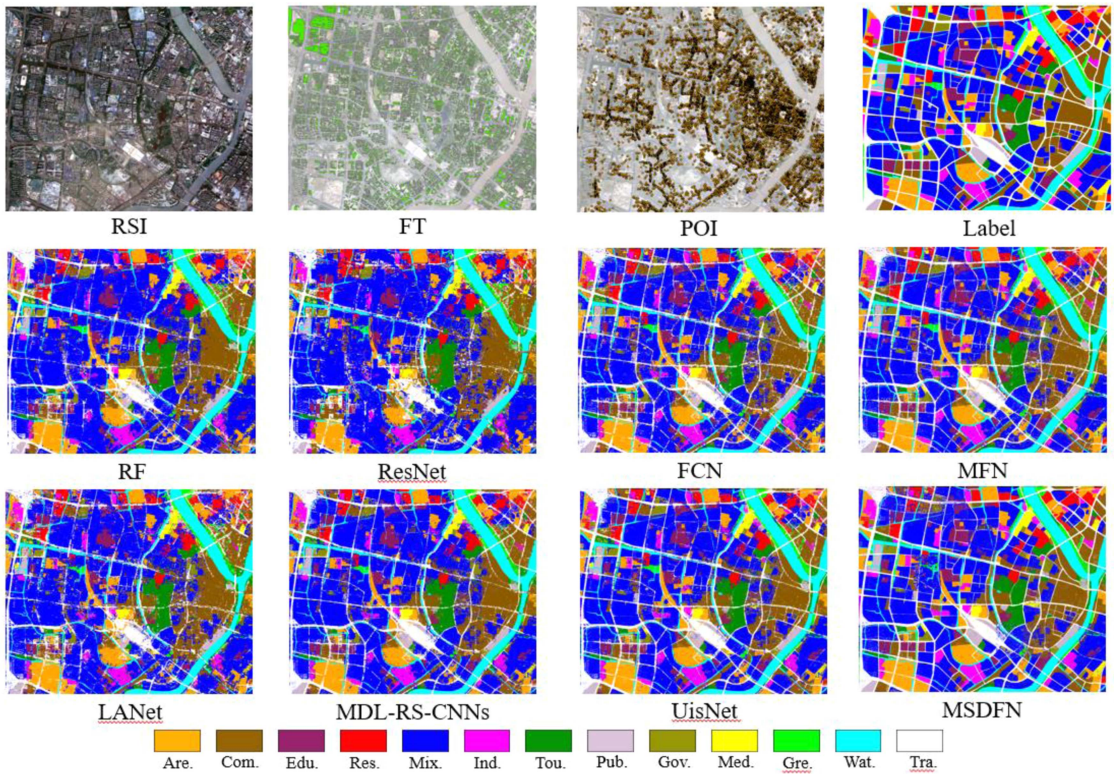


Fig. 6. Experimental results for different methods in Haishu area.

TABLE III
QUANTITATIVE EVALUATION RESULTS IN HAISHU AREA (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD)

Methods	F1													OA	KAPPA	Ave_F1	MIoU
	Are.	Com.	Edu.	Res.	Mix.	Ind.	Tou.	Pub.	Gov.	Med.	Gre.	Wat.	Tra.				
RF	0.7455	0.7151	0.5085	0.5901	0.8705	0.5097	0.7303	0.6196	0.4694	0.5608	0.5401	0.8457	0.5465	0.7213	0.6557	0.7158	0.4592
ResNet	0.6988	0.6343	0.3862	0.5427	0.8302	0.4396	0.6834	0.5032	0.3191	0.4299	0.4705	0.8179	0.4497	0.6570	0.5757	0.6478	0.3970
FCN	0.8231	0.7525	0.6440	0.6966	0.8655	0.6673	0.7732	0.7316	0.6042	0.6693	0.6459	0.9039	0.7073	0.7857	0.7392	0.7842	0.5680
MFN	0.8683	0.8098	0.7445	0.7700	0.8822	0.7659	0.8269	0.8017	0.7177	0.7572	0.7276	0.9256	0.7728	0.8330	0.7978	0.8326	0.6369
LANet	0.7625	0.6981	0.4943	0.5968	0.8417	0.5371	0.7245	0.6076	0.4590	0.5637	0.5440	0.8732	0.5875	0.7197	0.6564	0.7154	0.4736
MDL–RS–CNNs	0.8164	0.7906	0.6327	0.7030	0.9176	0.6366	0.7752	0.7210	0.5616	0.6453	0.5806	0.8932	0.6899	0.7998	0.7536	0.7969	0.5814
UisNet	0.8459	0.7309	0.5241	0.7580	0.8983	0.5965	0.6743	0.8683	0.8663	0.8441	0.6345	0.8667	0.8404	0.8139	0.7733	0.8115	0.6201
MSDFN	**0.9234**	**0.8799**	**0.8169**	**0.8513**	**0.9314**	**0.9006**	**0.8914**	**0.8843**	**0.7957**	**0.8635**	**0.7847**	**0.9538**	**0.8638**	**0.8980**	**0.8766**	**0.8978**	**0.7543**

simple and shallow feature extraction capabilities. FCN demonstrated similar performance but achieved reasonably satisfactory classification results in slightly more complex research scenarios. The context exploration based on attention (LANet) obtained better performance by adaptively allocating weights to feature maps in both spatial and channel dimensions. However, this method has certain limitations in modeling the global relationships of multimodal data. To address this issue, MDL–RS–CNNs proposed a cross-modal fusion module conducive to intermodality information transfer, demonstrating competitive improvement. UisNet effectively improves performance by utilizing transformer modules. Among all baseline methods, MFN exhibited the best performance, as it simultaneously addressed attention context exploration and multimodal fusion relationships, utilizing mean square error as supervision to enhance intraclass consistency and interclass discriminability. The proposed MSDFN method leverages adaptive weights for more stable and robust feature fusion. Through the visualization of results, our method achieves more accurate visual effects compared to other methods, with many misclassified areas in other methods.

2) Experimental Results in Zhenhai Area: In the early stages of urbanization, Zhenhai is characterized by UFZ-dominated enterprises and factories, with surrounding farmland. Although the distribution of UFZ in Zhenhai differs significantly from Haishu, our experiments yielded similar results, with an overall classification accuracy slightly higher than that of Haishu. This is especially evident in types related to POI, such as commercial areas, residential areas, educational zones, medical facility areas, etc. It is evident that our method makes more comprehensive use

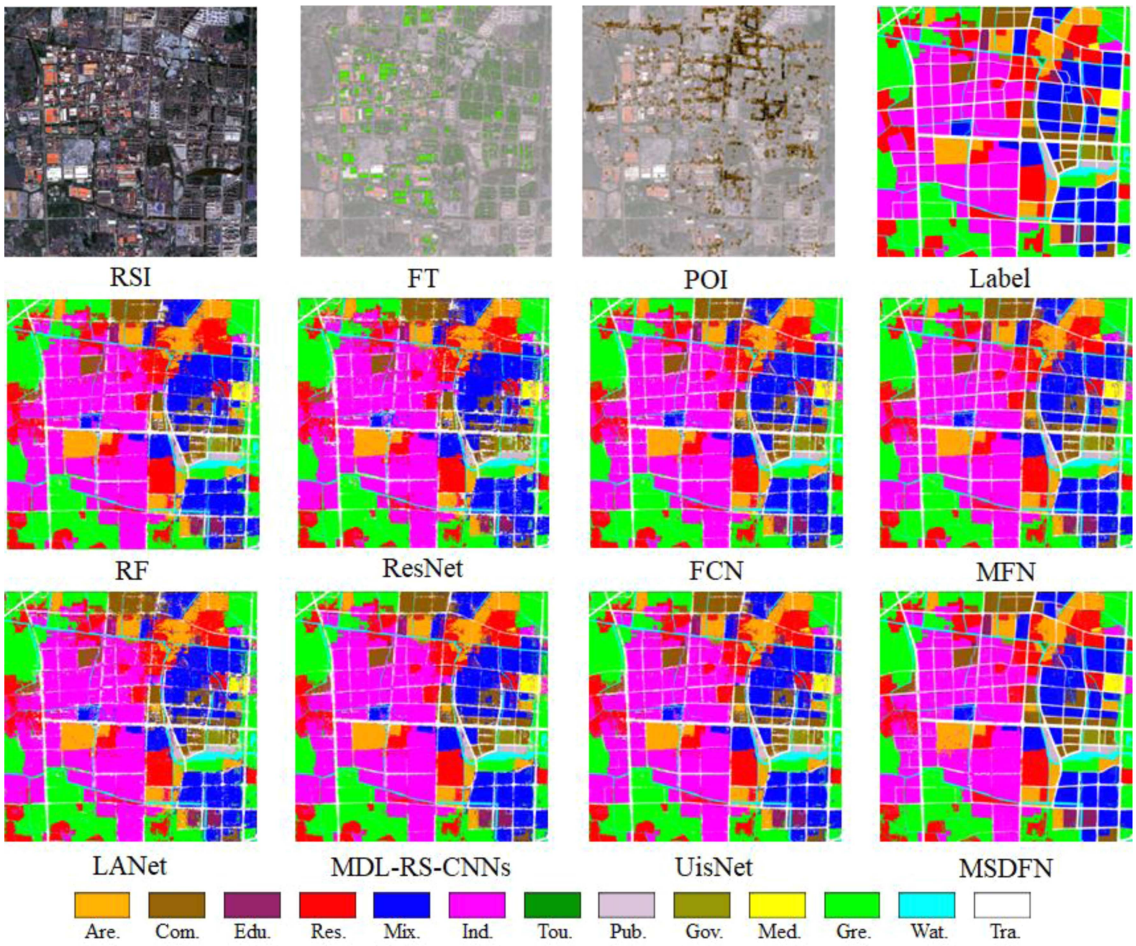


Fig. 7. Experimental results for different methods in Zhenhai area.

TABLE IV
QUANTITATIVE EVALUATION RESULTS IN ZHENHAI AREA (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD)

Methods	F1													OA	KAPPA	Ave_F1	MIoU
	Are.	Com.	Edu.	Res.	Mix.	Ind.	Tou.	Pub.	Gov.	Med.	Gre.	Wat.	Tra.				
RF	0.8646	0.7686	0.6991	0.7885	0.8619	0.8985	0.2041	0.7428	0.7973	0.7231	0.9402	0.7627	0.5630	0.8239	0.7926	0.8204	0.5870
ResNet	0.8428	0.7118	0.6437	0.7437	0.8275	0.8703	0.0037	0.6724	0.7146	0.7014	0.9275	0.7342	0.5253	0.7920	0.7551	0.7884	0.5179
FCN	0.8740	0.8050	0.7582	0.8121	0.8762	0.9006	0.2510	0.7755	0.8074	0.7419	0.9423	0.8228	0.7035	0.8535	0.8278	0.8525	0.6346
MFN	0.9004	0.8416	0.8200	0.8553	0.8978	0.9202	0.3956	0.8270	0.8568	0.8192	0.9526	0.8654	0.7731	0.8849	0.8647	0.8844	0.6956
LANet	0.8471	0.7442	0.6702	0.7690	0.8345	0.8816	0.0204	0.7281	0.8003	0.6347	0.9332	0.7776	0.5898	0.8128	0.7796	0.8101	0.5532
MDL-RS-CNNs	0.8930	0.8250	0.7555	0.8191	0.9061	0.9298	0.1816	0.7535	0.8429	0.7951	0.9498	0.7839	0.6537	0.8617	0.8371	0.8592	0.6311
UisNet	0.9063	0.8420	0.8060	0.8531	0.9051	0.9253	0.3596	0.8085	0.8597	0.7984	0.9612	0.8460	0.7277	0.8818	0.8609	0.8807	0.6974
MSDFN	**0.9450**	**0.9138**	**0.9238**	**0.8906**	**0.9451**	**0.9343**	**0.7090**	**0.9228**	**0.9481**	**0.8866**	**0.9666**	**0.9182**	**0.8415**	**0.9226**	**0.9090**	**0.9223**	**0.8063**

of POI. As Fig. 7 shows, our method exhibits the clearest boundaries for UFZ; moreover, the amount of confusing noise points is the least. Compared to the suboptimal MFN method, MSDFN outperforms it in terms of OA, KAPPA, Ave_F1, and MIoU by 3.55%, 4.17%, 3.47%, and 10.77%, respectively. The significant improvement in MIoU indicates that MSDFN achieves clearer boundaries for the buildings in this region, demonstrating a better understanding of semantics and segmentation effectiveness. Simultaneously, for challenging small targets like tourists, our method can also accurately identify, as shown in Table IV , while other methods perform poorly.

3) Experimental Results in Jiangbei–Zhenhai–Yinzhou Intersection Area: The Jiangbei–Zhenhai–Yinzhou intersection area is situated at the meeting point of multiple urban districts, surrounded by water bodies and green spaces. Various types of UFZ are scattered, reflecting a stage of urban development. In contrast to other methods that struggle with identifying blurred and numerous misrecognition areas in this dataset, our approach demonstrates outstanding performance in maintaining intraclass consistency while significantly reducing fragmented areas, its quantitative evaluation results are shown in Table V. This is particularly notable in the case of detailed regions and small

TABLE V
QUANTITATIVE EVALUATION RESULTS IN JIANGBEI–ZHENHAI–YINZHOU INTERSECTION AREA (THE BEST RESULTS ARE HIGHLIGHTED IN BOLD)

Methods	F1													OA	KAPPA	Ave_F1	MIoU
	Are.	Com.	Edu.	Res.	Mix.	Ind.	Tou.	Pub.	Gov.	Med.	Gre.	Wat.	Tra.				
RF	0.5290	0.8225	0.8929	0.6723	0.8433	0.8355	0.8166	/	/	0.6873	0.8699	0.8818	0.4740	0.7552	0.7229	0.7510	0.4708
ResNet	0.4944	0.8034	0.8663	0.6345	0.8146	0.8056	0.7906	/	/	0.6383	0.8497	0.8661	0.4477	0.7276	0.6917	0.7229	0.4610
FCN	0.5644	0.8348	0.9049	0.6794	0.8520	0.8394	0.8301	/	/	0.7190	0.8731	0.9079	0.6418	0.7859	0.7577	0.7844	0.5207
MFN	0.8272	0.8638	0.9218	0.8216	0.8770	0.8689	0.8611	/	/	0.7688	0.9008	0.9293	0.7015	0.8579	0.8388	0.8572	0.5662
LANet	0.7465	0.8057	0.8736	0.7555	0.8334	0.8286	0.8018	/	/	0.6301	0.8775	0.8944	0.5808	0.8027	0.7760	0.8011	0.5012
MDL-RS-CNNs	0.7672	0.8335	0.9249	0.8067	0.8915	0.8635	0.8235	/	/	0.6692	0.8857	0.9068	0.5890	0.8331	0.8105	0.8304	0.5576
UisNet	0.7732	0.7250	0.8152	0.8209	0.9052	0.7880	**0.9344**	/	/	0.5854	0.8605	0.8974	0.7706	0.8274	0.8187	0.8251	0.5442
MSDFN	**0.8902**	**0.9296**	**0.9488**	**0.8981**	**0.9274**	**0.9187**	0.9189	/	/	**0.8288**	**0.9204**	**0.9584**	**0.8295**	**0.9125**	**0.9008**	**0.9123**	**0.6319**

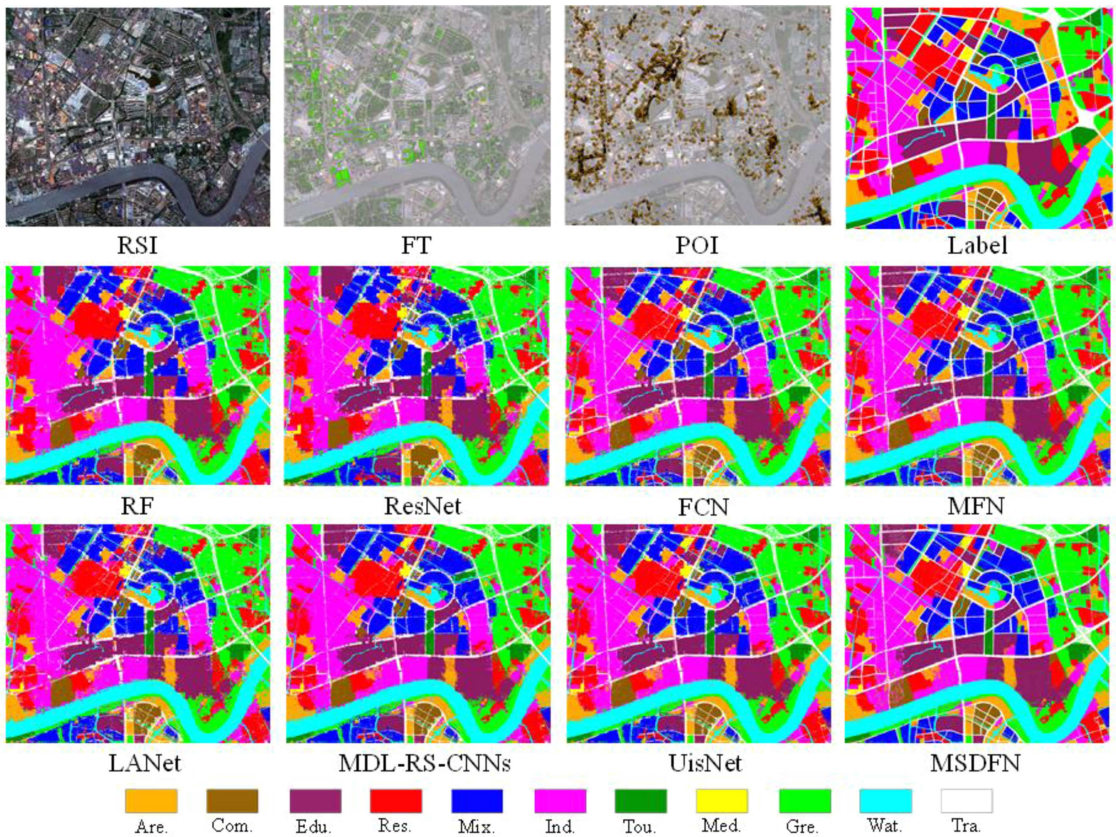


Fig. 8. Experimental results for different methods in Jiangbei–Zhenhai–Yinzhou intersection area.

targets in Fig. 8. Furthermore, the enlarged results of partial regions of the Jiangbei–Zhenhai–Yinzhou intersection dataset in Fig. 8 are shown in Fig. 9. In these enlarged results, it is evident that our method performs best in filling fragmented segmentation areas while preserving the original road features.

V. DISCUSSION

A. Performance of RSI, POI, and FT

To demonstrate the effectiveness of supplementing RSI with POI and FT data, we conduct overall classification experiments for each data source in Haishu, Zhenhai, and Jiangbei–Zhenhai–Yinzhou intersection area. The quantitative results are presented in Table VI. Given that the fusion module requires multiple data sources for integration, to ensure the fairness of the experiments, we do not employ any fusion or attention modules on any datasets. The remaining aspects were consistent with the proposed network framework.

From Table VI, it can be observed that RSI plays a decisive role in UFZ recognition. However, regardless of the study area, the UFZ classification results obtained by jointly using RSI, FT, and POI data sources far exceed the results obtained by using any single or combination of two data sources alone. This strongly indicates the helpfulness of POI and FT data in improving classification outcomes. Regarding FT data, we initially focus on the top 10 building categories (since areas, such as transportation, water bodies, and green spaces, do not have building

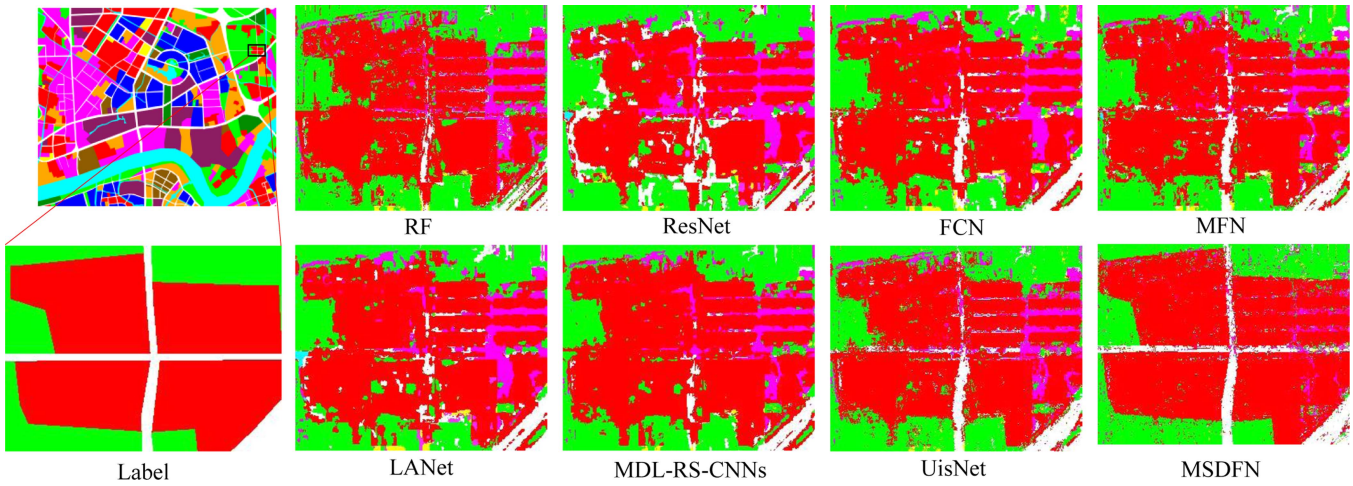


Fig. 9. Comparison of enlarged results of partial regions in the Jiangbei–Zhenhai–Yinzhou intersection dataset among different comparison methods.

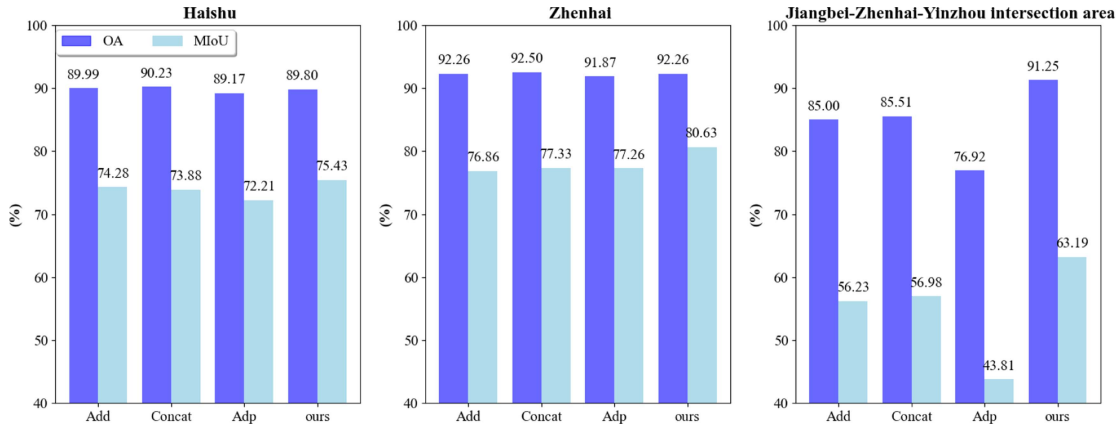


Fig. 10. Comparison experiments of different fusion methods in each study area.

FTs and consequently cannot be identified as corresponding UFZ). Without the assistance of RSI, overall FT data exhibit better recognition performance for residential and industrial areas but are less sensitive to commercial areas, schools, and other regions. The recognition accuracy in some areas is even below 20%, possibly due to the high fragmentation level of these regions. However, when RSI data are combined with FT data, the recognition performance significantly improves. The OA for Haishu, Zhenhai, and Jiangbei–Zhenhai–Yinzhou intersection area increases by 4.57%, 6.68%, and 5.69%, respectively. This suggests that FT data actively enhances UFZ recognition capabilities in the presence of buildings.

Similarly, POI data demonstrate favorable recognition outcomes in POI-dense areas, such as commercial districts, industrial zones, and hospitals, but exhibits poorer recognition performance in other regions. With the assistance of POI, the recognition accuracy of RSI in Haishu, Zhenhai, and Jiangbei–Zhenhai–Yinzhou intersection area improves by 9.56%, 9.55%, and 4.19%, respectively. The inclusion of POI facilitates the classification of UFZ associated with human economic and social activities.

Upon combining the three modal data sources, the recognition performance in each study area sees further improvement. In comparison to using only RSI data, the OA in Haishu, Zhenhai, and Jiangbei–Zhenhai–Yinzhou intersection area increases by 11.61%, 14.71%, and 8.54%, respectively. Kappa improves by 13.16%, 18.39%, and 10.04%, Ave_F1 increases by 11.73%, 15.16%, and 8.58%, and MIoU elevates by 18.07%, 25.32%, and 16.27%. This indicates that the introduction of multimodal data actively promotes UFZ recognition.

B. Ablation Study

1) Hyperparameters: We evaluate the hyperparameters of our method, primarily the parameter λ in the loss function (5). λ serves as a factor to balance the cross-entropy loss, weight decay loss, and dice loss. The results for λ in Jiangbei–Zhenhai–Yinzhou intersection area are presented in Table VII. From the table results, we ultimately choose $\lambda = 1$ as the parameter value for MSDFN, as it yields the highest values for OA and MIoU in the experiments, with respective values of 0.9075 and 0.6185.

2) Performance of AW-IFM, CDS-JM, and MS-FFM: MSDFN consists of three key components: AW-IFM, CDS-JM, and MS-FFM. To assess the necessity and effectiveness of these modules, we conduct ablation experiments on various

TABLE VI
QUANTITATIVE COMPARATIVE EXPERIMENTS WITH DIFFERENT SOURCE DATA OF RSI WITH POI AND FT

Datasets	Data	F1													OA	KAPPA	Ave_F1	MIoU
		Are.	Com.	Edu.	Res.	Mix.	Ind.	Tou.	Pub.	Gov.	Med.	Gre.	Wat.	Tra.				
Haishu	RSI	0.7544	0.6598	0.4778	0.5331	0.8311	0.5432	0.6743	0.5962	0.3731	0.3809	0.5475	0.9006	0.6515	0.7141	0.6481	0.7093	0.4176
	FT	0.3038	0.4936	0.2283	0.2985	0.8040	0.1462	0.5061	0.2214	0.1230	0.0497	0.0124	0.3775	0.1782	0.4641	0.3153	0.4253	0.1742
	POI	0.3907	0.5963	0.2927	0.2207	0.8356	0.1925	0.4545	0.2443	0.3295	0.4378	0.0434	0.4443	0.2109	0.5154	0.3802	0.4825	0.2462
	FT+POI	0.4876	0.6219	0.3830	0.5152	0.8354	0.2916	0.4815	0.4411	0.3890	0.4914	0.0741	0.4220	0.2351	0.5538	0.4385	0.5288	0.3014
	RSI+FT	0.8152	0.7461	0.6265	0.6959	0.8572	0.6750	0.7576	0.6853	0.5561	0.5278	0.6584	0.9149	0.7268	0.7809	0.7335	0.7792	0.5154
	RSI+POI	0.8450	0.7776	0.6900	0.6916	0.8766	0.7156	0.8062	0.7906	0.6386	0.7132	0.7017	0.9220	0.7426	0.8096	0.7680	0.8084	0.5870
	RSI+FT+POI	**0.8874**	**0.8402**	**0.7940**	**0.8218**	**0.9034**	**0.8049**	**0.8588**	**0.8433**	**0.7639**	**0.7985**	**0.7695**	**0.9342**	**0.8079**	**0.8612**	**0.8320**	**0.8609**	**0.6708**
Zhenhai	RSI	0.8569	0.7427	0.6958	0.7632	0.8082	0.8773	0.3699	0.6732	0.6963	0.7195	0.9203	0.8333	0.6951	0.8213	0.7899	0.8206	0.5701
	FT	0.7097	0.5783	0.3406	0.5451	0.7560	0.8044	0.0000	0.1835	0.2107	0.7381	0.8951	0.0732	0.1330	0.6318	0.5636	0.5981	0.3304
	POI	0.6221	0.5458	0.1687	0.5171	0.6947	0.8311	0.0900	0.1244	0.7375	0.7792	0.7404	0.0601	0.0343	0.5799	0.4988	0.5318	0.3192
	FT+POI	0.7767	0.6715	0.5597	0.6003	0.7701	0.8484	0.0900	0.4093	0.7828	0.5383	0.8817	0.0672	0.1414	0.6674	0.6061	0.6328	0.4144
	RSI+FT	0.9016	0.8243	0.8179	0.8440	0.8896	0.9167	0.4772	0.7721	0.7762	0.7803	0.9522	0.8607	0.7648	0.8782	0.8569	0.8777	0.6665
	RSI+POI	0.8855	0.8231	0.7728	0.8230	0.8558	0.9057	0.4433	0.7970	0.8424	0.7934	0.9350	0.8597	0.7559	0.8632	0.8392	0.8627	0.6575
	RSI+FT+POI	**0.9181**	**0.8814**	**0.8660**	**0.8828**	**0.9166**	**0.9354**	**0.4942**	**0.8541**	**0.8907**	**0.8563**	**0.9592**	**0.8829**	**0.8172**	**0.9067**	**0.8903**	**0.9064**	**0.7328**
Jiangbei-Zhenhai-Yinzhou intersection area	RSI	0.5169	0.5653	0.8142	0.6241	0.7299	0.7625	0.5739	/	/	0.2617	0.8253	0.9153	0.6131	0.7156	0.6780	0.7134	0.3603
	FT	0.0949	0.2274	0.8136	0.5553	0.6381	0.6297	0.0165	/	/	0.0849	0.4647	0.7168	0.0767	0.4870	0.4169	0.4383	0.1960
	POI	0.2834	0.7649	0.8184	0.5733	0.7668	0.7054	0.7846	/	/	0.6824	0.6021	0.7165	0.1039	0.5874	0.5336	0.5686	0.3736
	FT+POI	0.2884	0.8202	0.8416	0.6106	0.8199	0.7296	0.7990	/	/	0.6844	0.5957	0.7172	0.1353	0.6096	0.5586	0.5906	0.3902
	RSI+FT	0.5486	0.6970	0.8469	0.6660	0.8094	0.8072	0.6295	/	/	0.3848	0.8516	0.9255	0.6868	0.7613	0.7298	0.7595	0.4139
	RSI+POI	0.5924	0.8624	0.9227	0.7024	0.8642	0.8543	0.8633	/	/	0.7779	0.8866	0.9318	0.7093	0.8112	0.7864	0.8102	0.5181
	RSI+FT+POI	**0.6067**	**0.8825**	**0.9333**	**0.7293**	**0.8920**	**0.8779**	**0.8790**	/	/	**0.7974**	**0.9001**	**0.9371**	**0.7491**	**0.8317**	**0.8096**	**0.8307**	**0.5410**

The best results are highlighted in bold.

TABLE VII
EVALUATION OF THE EFFECTIVENESS OF THE HYPERPARAMETERS IN (5)

λ	OA	MIoU
0	0.8619	0.6128
0.5	0.8539	0.5956
0.7	0.8559	0.6002
0.9	0.8569	0.5657
1	**0.9075**	**0.6185**
1.1	0.8507	0.5571

The best results are highlighted in bold.

datasets to evaluate their impact on classification accuracy. The experimental results are presented in Table VIII.

1) *AW-IFM:* As shown in Table VIII, the removal of AW-IFM results in a decrease in various metrics, demonstrating the effectiveness of the proposed AW-IFM in fusing remote sensing and social perception data. Furthermore, building upon the results in Table VIII, we investigate the impact of different fusion methods on recognition results. We conduct experiments using elementwise addition (Add), concatenation (Concat), adaptive fusion (Adp), and our dynamic fusion method which is based on data characteristics. The results are illustrated in Fig. 10. For adaptive fusion, each data source is assigned a randomly initialized weight parameter, and the sum of these three weight parameters equals 1. The network is trained through feedback to automatically adjust the weights until achieving the optimal output results.
From Fig. 9, it can be observed that, regardless of the dataset, the elementwise addition method performs poorly due to its simple and rough fusion strategy. The concatenation method is similar to addition but with slight improvements in various metrics. Unexpectedly, the adaptive fusion yields the poorest predictive results. We suspect that the adaptive method requires more computational resources, and our comparative experiment is only conducted for 100 epochs, which might not have allowed this network model to fully converge. This is further supported by the fact that the optimal test results mostly emerged close to the 100th epoch in the experiment. Conversely, our data-driven fusion method demonstrates the best overall performance, with the MIoU metric reaching the highest values for each dataset. In the third dataset, the OA also surpasses the other three fusion methods, reaching 91.25%. This indicates that our fusion method exhibits optimal recognition performance in scenarios with diverse UFZ types. In the other two study areas, our method shows OA results similar to the elementwise addition and concatenation methods. Therefore, we chose the data-driven method to implement feature fusion.
2) *CDS-JM:* CDS-JM serves as a crucial backbone module in the network. As evident from Table VIII, the removal of CDS-JM led to a significant decrease in the network's performance.

TABLE VIII
ABLATION STUDY TO ANALYZE THE PERFORMANCE OF AW-IFM, CDS-JM, AND MS-FFM

Datasets	AW-IFM	CDS-JM	MS-FFM	OA	MIoU
Haishu		√	√	**0.8999**	0.7428
	√		√	0.8279	0.6248
	√	√		0.8832	0.7113
	√	√	√	0.8980	**0.7543**
Zhenhai		√	√	**0.9226**	0.7686
	√		√	0.9089	0.7838
	√	√		0.9221	0.7672
	√	√	√	**0.9226**	**0.8063**
Jiangbei–Zhenhai–Yinzhou intersection area		√	√	0.8500	0.5623
	√		√	0.8354	0.5385
	√	√		0.8487	0.5643
	√	√	√	**0.9125**	**0.6319**

The best results are highlighted in bold.

3) *MS-FFM:* MS-FFM integrates multiscale features, global information, attention mechanisms, and enhances module performance through residual connections and SEB. This enables the model to have greater expressiveness and better adaptability to complex tasks. The ablation experiments in Table VIII also validate this point. Additionally, MS-FFM demonstrates outstanding performance in complex urban areas. In the diverse Haishu dataset, the removal of the MS-FFM module resulted in a decrease of 1.48% and 4.30%. The impact is more pronounced in Jiangbei–Zhenhai–Yinzhou intersection area, with reductions of 6.38% and 6.76%, respectively. This indicates the importance of considering multiscale modeling, particularly in areas with significant urbanization.

C. Issues and Prospects

While our multimodal fusion network has achieved fine-grained delineation of UFZ across the entire map, certain limitations persist. For instance, the model relies on a substantial amount of accurately annotated training samples, primarily generates through manual labeling. Future efforts should focus on devising innovative algorithms to reduce the workload associates with data preprocessing. Additionally, the study areas cover in this article are relatively small in scale, and as of now, there is no UFZ recognition modeling for large-scale scenes, such as in Zhejiang Province. Addressing this gap will be a key focus of future article.

VI. CONCLUSION

The effective utilization of RSI and social sensing data is crucial for accurately identifying UFZ. This article proposes an adaptive weighted integration fusion module designed to condense complementary information and eliminate redundant information based on the inherent characteristics of multisource data. Additionally, the inclusion of building FT data supplements missing semantic information related to buildings.

Experimental results on datasets from three regions in Ningbo demonstrate that the proposed method exhibits excellent identification performance across UFZ with different distribution patterns. Under the dominance of RSI, the inclusion of FT and POI significantly supplements missing feature information, particularly in socioeconomic and building-related areas. Comparative experiments show that the proposed MSDFN outperforms other UFZ identification methods. Ablation study further confirms the effectiveness of the proposed modules.

REFERENCES

[1] W. Gao, L. Nan, B. Boom, and H. Ledoux, "PSSNet: Planarity-sensible semantic segmentation of large-scale urban meshes," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 32–44, Feb. 2023.
[2] D. He, Q. Shi, J. Xue, P. M. Atkinson, and X. Liu, "Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation," *Remote Sens. Environ.*, vol. 299, Dec. 2023, Art. no. 113884.
[3] A. Kavvada et al., "Towards delivering on the sustainable development goals using Earth observations," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111930.
[4] Q. Liu, X. Chen, X. Meng, H. Chen, F. Shao, and W. Sun, "Dual-task interactive learning for unsupervised spatio-temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Apr. 2023, Art. no. 5402015.
[5] Q. Liu, X. Meng, F. Shao, and S. Li, "PSTAF-GAN: Progressive spatio-temporal attention fusion method based on generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5408513.
[6] S. Zhang, X. Meng, Q. Liu, G. Yang, and W. Sun, "Feature-decision level collaborative fusion network for hyperspectral and LiDAR classification," *Remote Sens.*, vol. 15, no. 17, Jan. 2023, Art. no. 4148.
[7] Q. Zhu, X. Sun, Y. Zhong, and L. Zhang, "High-resolution remote sensing image scene understanding: A review," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3061–3064.
[8] Y. Zhang, P. Liu, and F. Biljecki, "Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 153–168, Apr. 2023.
[9] Z. Pan, J. Xu, Y. Guo, Y. Hu, and G. Wang, "Deep learning segmentation and classification for urban village using a worldview satellite image based on U-Net," *Remote Sens.*, vol. 12, no. 10, Jan. 2020, Art. no. 1574.

[10] X. Zhang and S. Du, "A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings," *Remote Sens. Environ.*, vol. 169, pp. 37–49, Nov. 2015.

[11] X. Huang, J. Yang, J. Li, and D. Wen, "Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 403–415, May 2021.

[12] W. Zhou, D. Ming, X. Lv, K. Zhou, H. Bao, and Z. Hong, "SO–CNN based urban functional zone fine division with VHR remote sensing image," *Remote Sens. Environ.*, vol. 236, Jan. 2020, Art. no. 111458.

[13] S. Du, S. Du, B. Liu, and X. Zhang, "Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach," *Remote Sens. Environ.*, vol. 261, Aug. 2021, Art. no. 112480.

[14] K. Zhang et al., "Distance weight-graph attention model-based high-resolution remote sensing urban functional zone identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jan. 2022, Art. no. 5608518.

[15] M. Schultz, J. Voss, M. Auer, S. Carter, and A. Zipf, "Open land cover from OpenStreetMap and remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 63, pp. 206–213, Dec. 2017.

[16] L. Mu, L. Wang, Y. Wang, X. Chen, and W. Han, "Urban land use and land cover change prediction via self-adaptive cellular based deep learning with multisourced data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5233–5247, Dec. 2019.

[17] X. Zhao and Y. Guo, "Planning bikeway network for urban commute based on mobile phone data: A case study of Beijing," *Travel Behav. Soc.*, vol. 34, Jan. 2024, Art. no. 100672.

[18] B. Zhang, C. Zhong, Q. Gao, Z. Shabrina, and W. Tu, "Delineating urban functional zones using mobile phone data: A case study of cross-boundary integration in Shenzhen–Dongguan–Huizhou area," *Comput., Environ. Urban Syst.*, vol. 98, Dec. 2022, Art. no. 101872.

[19] L. Bai et al., "Geographic mapping with unsupervised multi-modal representation learning from VHR images and POIs," *ISPRS J. Photogrammetry Remote Sens.*, vol. 201, pp. 193–208, Jul. 2023.

[20] R. Wu, J. Wang, D. Zhang, and S. Wang, "Identifying different types of urban land use dynamics using point-of-interest (POI) and random forest algorithm: The case of Huizhou, China," *Cities*, vol. 114, Jul. 2021, Art. no. 103202.

[21] G. Lou, Q. Chen, K. He, Y. Zhou, and Z. Shi, "Using nighttime light data and POI big data to detect the urban centers of Hangzhou," *Remote Sens.*, vol. 11, no. 15, Aug. 2019, Art. no. 1821.

[22] S. Hu et al., "Urban function classification at road segment level using taxi trajectory data: A graph convolutional neural network approach," *Comput., Environ. Urban Syst.*, vol. 87, May 2021, Art. no. 101619.

[23] R. Cao et al., "Integrating aerial and street view images for urban land use classification," *Remote Sens.*, vol. 10, no. 10, Sep. 2018, Art. no. 1553.

[24] Y. Zhu and S. Newsam, "Land use classification using convolutional neural networks applied to ground-level images," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, 2015, pp. 1–4.

[25] J. Guilin, H. U. Fangyu, and S. H. I. Lixing, "Urban functional area identification based on call detail record data," *J. Comput. Appl.*, vol. 36, no. 7, Jul. 2016, Art. no. 2046.

[26] K. Liu, L. Yin, F. Lu, and N. Mou, "Visualizing and exploring POI configurations of urban regions on POI-type semantic space," *Cities*, vol. 99, Apr. 2020, Art. no. 102610.

[27] Y. Xu et al., "A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method," *Comput., Environ. Urban Syst.*, vol. 95, Jul. 2022, Art. no. 101807.

[28] X. Zhang, S. Du, and Q. Wang, "Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 132, pp. 170–184, Oct. 2017.

[29] F. Mo, X. Fan, C. Chen, C. Bai, and H. Yamana, "Sampling-based epoch differentiation calibrated graph convolution network for point-of-interest recommendation," *Neurocomputing*, vol. 571, Feb. 2024, Art. no. 127140.

[30] P. Li, J. Liu, A. Luo, Y. Wang, J. Zhu, and S. Xu, "Deep learning method for Chinese multisource point of interest matching," *Comput., Environ. Urban Syst.*, vol. 96, Sep. 2022, Art. no. 101821.

[31] J. Wang, A. Ma, Y. Zhong, Z. Zheng, and L. Zhang, "Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery," *Remote Sens. Environ.*, vol. 277, Aug. 2022, Art. no. 113058.

[32] C. Zheng, J. Nie, Z. Wang, N. Song, J. Wang, and Z. Wei, "High-order semantic decoupling network for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 2023, Art. no. 5401415.

[33] W. Lu, C. Tao, H. Li, J. Qi, and Y. Li, "A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data," *Remote Sens. Environ.*, vol. 270, Mar. 2022, Art. no. 112830.

[34] W. Huang, D. Zhang, G. Mai, X. Guo, and L. Cui, "Learning urban region representations with POIs and hierarchical graph infomax," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 134–145, Feb. 2023.

[35] Y. Feng et al., "An SOE-based learning framework using multisource big data for identifying urban functional zones," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7336–7348, Jun. 2021.

[36] Y. Wang, Y. Li, X. Song, X. Zou, and J. Xiao, "Correlation analysis between NPP-VIIRS nighttime light data and POIs data—A comparison study in different districts and counties of Nanchang," *Inst. Phys. Publishing Conf. Ser.: Earth Environ. Sci.*, vol. 693, no. 1, Mar. 2021, Art. no. 012103.

[37] X. Kong, "China must protect high-quality arable land," *Nature*, vol. 506, no. 7486, Feb. 2014, Art. no. 7.

[38] X. Li, Y. Zhou, P. Gong, K. C. Seto, and N. Clinton, "Developing a method to estimate building height from Sentinel-1 data," *Remote Sens. Environ.*, vol. 240, Apr. 2020, Art. no. 111705.

[39] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, "Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images," *Remote Sens. Environ.*, vol. 264, Oct. 2021, Art. no. 112589.

[40] X. Huang, D. Wen, J. Li, and R. Qin, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sens. Environ.*, vol. 196, pp. 56–75, Jul. 2017.

[41] G. Wu et al., "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, Mar. 2018, Art. no. 407.

[42] K. Nogueira, O. A. B. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, Jan. 2017.

[43] L. Xie, X. Feng, C. Zhang, Y. Dong, J. Huang, and K. Liu, "Identification of urban functional areas based on the multimodal deep learning fusion of high-resolution remote sensing images and social perception data," *Buildings*, vol. 12, no. 5, Apr. 2022, Art. no. 556.

[44] Y. Cheng et al., "Multi-scale feature fusion and transformer network for urban green space segmentation from high-resolution remote sensing images," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 124, Nov. 2023, Art. no. 103514.

[45] W. Zhou, C. Persello, M. Li, and A. Stein, "Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information," *Remote Sens. Environ.*, vol. 297, Nov. 2023, Art. no. 113767.

[46] R. Cao et al., "Deep learning-based remote and social sensing data fusion for urban region function recognition," *ISPRS J. Photogrammetry Remote Sens.*, vol. 163, pp. 82–97, May 2020.

[47] Z. Guo, J. Wen, and R. Xu, "A shape and size free-CNN for urban functional zone mapping with high-resolution satellite images and POI data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Sep. 2023, Art. no. 5622117.

[48] H. Bao, D. Ming, Y. Guo, K. Zhang, K. Zhou, and S. Du, "DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data," *Remote Sens.*, vol. 12, no. 7, Mar. 2020, Art. no. 1088.

[49] M. Wu, X. Zhao, Z. Sun, and H. Guo, "A hierarchical multiscale superpixel-based classification method for extracting urban impervious surface using deep residual network from worldview-2 and LiDAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 210–222, Jan. 2019.

[50] Q. Liu, X. Meng, F. Shao, and S. Li, "Supervised-unsupervised combined deep convolutional neural networks for high-fidelity pansharpening," *Inf. Fusion*, vol. 89, pp. 292–304, Jan. 2023.

[51] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf.*, 2018, pp. 169–175.

[52] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[53] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.

[54] X. Wang, Z. Hu, S. Shi, M. Hou, L. Xu, and X. Zhang, "A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved UNet," *Sci. Rep.*, vol. 13, no. 1, May 2023, Art. no. 7600.

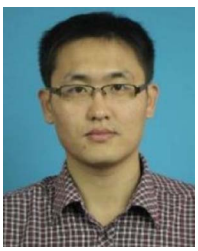
[55] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 426–435, Jan. 2021.
[56] D. Hong et al., "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[57] R. Fan, F. Li, W. Han, J. Yan, J. Li, and L. Wang, "Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Sep. 2022, Art. no. 5630316.

Hangfeng Qiao received the B.S. degree in telecommunication engineering from the Ningbo University, Ningbo, China, in 2022, where he is currently working toward the M.S. degree in telecommunication engineering.

His research interests include deep learning for image processing and multisource image fusion.

Huiping Jiang received the Ph.D. degree in cartography and geographic information system from the University of Chinese Academy of Sciences, Beijing, China, in 2019.

Following his 3-year postdoctoral study, he has been with the Key Laboratory of Regional Sustainable Development Modeling, Chinese Academy of Sciences, Beijing, China, as an Assistant Research Fellow, since 2023. His research interests include urban remote sensing, big data for SDG 11, sustainable urbanization, and human dimensions of global environmental change.

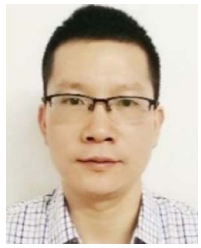


Gang Yang received the M.S. degree in geographical information system from the Hunan University of Science and Technology, Xiangtan, China, in 2012, and the Ph.D. degree in cartography and geographical information engineering from the School of Resource and Environmental Sciences, Wuhan University, Wuhan, China, in 2016.

He is currently an Associate Professor with the Ningbo University, Ningbo, China. His research interests include missing information reconstruction of remote sensing image, cloud removal of remote sensing image, and remote sensing time-series products temporal reconstruction.

Faming Jing received the B.S. degree in surveying and mapping engineering from the Wuhan University, Wuhan, China, in 2006, and the master's degree in geodesy and survey engineering from the Wuhan University, Wuhan, China, in 2008.

He is currently a Senior Engineer with the Ningbo Institute of Surveying, Mapping and Remote Sensing, Ningbo, China. His research interests include investigation and analysis of land resources and spatial data mining.

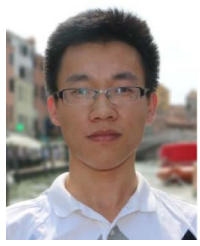


Weiwei Sun (Senior Member, IEEE) received the B.S. degree in surveying and mapping from the Tongji University, Shanghai, China, in 2007, and the Ph.D. degree in cartography and geographic information engineering from the Tongji University, Shanghai, China, in 2013.

From 2011 to 2012, he was with the Department of Applied Mathematics, University of Maryland, College Park, MD, USA, and is a Visiting Scholar with the famous Prof. J. Benedetto to study on the dimensionality reduction of hyperspectral image. From 2014 to 2016, he was with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, working as a Postdoc to study intelligent processing in hyperspectral imagery. From 2017 to 2018, he was with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS, USA, and is a Visiting Scholar in hyperspectral image processing. He is currently a Full Professor with the Ningbo University, Ningbo, China. He has authored and coauthored more than 80 journal papers. His research interest includes hyperspectral image processing with machine learning.

Chenyang Lu received the B.Sc. degree in mechatronics engineering from the Zhejiang University, Hangzhou, China, in 2017, the M.Sc. degree in control systems technology from the Eindhoven University of Technology (TU/e), Eindhoven, The Netherlands, in 2018, and the Ph.D. degree in artificial intelligence from the TU/e, in 2023.

He is currently a Lecturer with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include computer vision, autonomous driving, and artificial intelligence.



Xiangchao Meng (Senior Member, IEEE) received the B.S. degree in geographic information system from the Shandong University of Science and Technology, Qingdao, China, in 2012, and the Ph.D. degree in cartography and geography information system from the Wuhan University, Wuhan, China, in 2017.

He is currently a Professor with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include multisource data fusion and machine learning.