International Journal of Geo-Information

MDPI

Article

A Deep-Learning-Based Multimodal Data Fusion Framework for Urban Region Function Recognition

Mingyang Yu [1], Haiqing Xu [1,*], Fangliang Zhou [1], Shuai Xu [1] and Hongling Yin [2]

[1] School of Surveying and Geo-Informatics, Shandong Jianzhu University, Ji'nan 250101, China; ymy@sdjzu.edu.cn (M.Y.); 2021165123@stu.sdjzu.edu.cn (F.Z.); 2021165102@stu.sdjzu.edu.cn (S.X.)
[2] School of Architecture and Urban Planning, Shandong Jianzhu University, Ji'nan 250101, China; 12949@sdjzu.edu.cn
* Correspondence: 2021160105@stu.sdjzu.edu.cn

Abstract: Accurate and efficient classification maps of urban functional zones (UFZs) are crucial to urban planning, management, and decision making. Due to the complex socioeconomic UFZ properties, it is increasingly challenging to identify urban functional zones by using remote-sensing images (RSIs) alone. Point-of-interest (POI) data and remote-sensing image data play important roles in UFZ extraction. However, many existing methods only use a single type of data or simply combine the two, failing to take full advantage of the complementary advantages between them. Therefore, we designed a deep-learning framework that integrates the above two types of data to identify urban functional areas. In the first part of the complementary feature-learning and fusion module, we use a convolutional neural network (CNN) to extract visual features and social features. Specifically, we extract visual features from RSI data, while POI data are converted into a distance heatmap tensor that is input into the CNN with gated attention mechanisms to extract social features. Then, we use a feature fusion module (FFM) with adaptive weights to fuse the two types of features. The second part is the spatial-relationship-modeling module. We designed a new spatial-relationship-learning network based on a vision transformer model with long- and short-distance attention, which can simultaneously learn the global and local spatial relationships of the urban functional zones. Finally, a feature aggregation module (FGM) utilizes the two spatial relationships efficiently. The experimental results show that the proposed model can fully extract visual features, social features, and spatial relationship features from RSIs and POIs for more accurate UFZ recognition.

Keywords: multimodal data fusion; UFZ map; spatial relationship modeling; vision transformer

Citation: Yu, M.; Xu, H.; Zhou, F.; Xu, S.; Yin, H. A Deep-Learning-Based Multimodal Data Fusion Framework for Urban Region Function Recognition. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 468. https://doi.org/10.3390/ijgi12120468

Academic Editor: Wolfgang Kainz

Received: 7 September 2023
Revised: 5 November 2023
Accepted: 19 November 2023
Published: 21 November 2023

1. Introduction

Today, more than half of the world's population lives in cities, yet cities cover only a tiny fraction of the Earth's surface. Asian and African countries are continuously urbanizing, and the urban population continues to grow; the world's urban population is expected to increase by 500 million by 2030 [1]. Therefore, with a growing urban population, it is critical to manage and monitor limited urban areas. An urban functional zone is a concept that describes people's different activities in a certain area, such as industrial areas, commercial areas, or residential areas [2,3]. As a basic urban unit, an accurate UFZ map is very important for urban planning, management, and decision making [4,5]. However, with the rapid development of urbanization in the world, urban functional zone maps managed by the government cannot be updated in a timely manner [6,7]. Therefore, it is crucial to make accurate and timely UFZ maps.

With the rapid development of related disciplines and technologies, high-resolution remote-sensing image data have gradually shown potential in the task of UFZ recognition. Research based on remote-sensing images continues to develop, and the use of RSIs is recognized as one of the most effective and efficient methods [8,9]. In the past, some

traditional methods, such as the scale-invariant feature transform (SIFT) [10] and the histogram of oriented gradients (HOG) [11,12], were used to identify UFZs from remote-sensing images. These methods perform well in traditional land use and land cover (LULC) classification tasks but do not perform well in urban area recognition tasks with complex structures and fine semantics. Therefore, some scholars have proposed topic models to further identify UFZs from remote-sensing imagery. Additionally, probabilistic topic models [13,14] can further exploit the potential semantic relationships of urban areas by using object features. However, such methods are still only based on low-level manual features and cannot effectively represent the complex high-level semantic feature relationships of UFZs. In recent years, computer science has developed rapidly, and the advantages of many deep-learning methods [15–17] in identifying urban functional zones and scenario-level LULC classification have become increasingly obvious. For example, benefiting from the development of the transformer [18], Wang [19] designed a U-shaped transformer network based on the UNet [20] model architecture to interpret high-resolution urban scene images and obtained satisfactory results on four challenging datasets. Zhou [21] introduced the concept of the super object (SO) and classified the urban functional area based on the methods of frequency statistics and convolutional neural networks. Du [22] mapped large-scale and fine-grained urban functional zones from remote-sensing images using a multi-scale semantic segmentation network and an object-based approach. In general, the use of high-resolution RSIs can achieve relatively good results for producing fine-grained urban area classification maps [13,23,24]. However, RSI data perform well in extracting the physical characteristics of ground objects, such as the distribution of buildings and the spatial structure of cities, but they fail to reflect the dynamically changing properties and human social activity information [25,26]. In addition, some UFZ components are visually very similar, and the method based on remote-sensing image data can extract the visual semantic features efficiently, but it cannot analyze the social semantic features between and within different functional areas. The urban functional zone contains various functional properties of regions that are directly related to human social activities [27,28]. This relationship is complex to some extent. Therefore, it is difficult to obtain high-precision UFZ classification maps using only RSI data.

It has been shown that social perception and human activities are better methods for dynamically identifying urban areas [26,29]. Socioperceptual big data that record human activity in real time are becoming increasingly available, such as points of interest (POIs) [2,30], mobile phone positioning data [31,32], social media check-in data [33], geotagged photos [34,35], and vehicle GPS trajectory data [36,37]. Unlike remote-sensing imagery, these data are byproducts of human social activities, and many have temporal features; therefore, they contain rich socioeconomic attributes. For example, by fusing remote-sensing image data and taxi trajectory data, Qian [38] used the road network as the basic segmentation unit to identify urban functional districts based on the residual network framework. Cao [39] proposed two strategies: enforcing cross-modal feature consistency (CMFC) and cross-modal triplet (CMT) constraints. Time-dependent social sensing signature features were extracted based on a long short-term memory network (LSTM) and a one-dimensional convolutional neural network and fused with the remote-sensing image data for more accurate urban functional area classification. POIs are the main static data in social sensing data; they are not only easy to obtain but can also provide comprehensive land use information based on human activities and geographical locations [40]. Xu [41] calculated the statistical features of POIs and fused them with remote-sensing image data to identify UFZs. Lu [42] proposed a unified deep-learning framework that can simultaneously extract visual features, social features, and spatial relationship features from RSI data and POI data. Bao [43] proposed the deeper-feature convolutional neural network (DFCNN) and integrated remote-sensing data and POI data to identify urban functional zones. There have been many efforts by scholars to recognize urban functional zones, but many of the previous studies could not effectively identify the implicit relationship between UFZs and POIs because the relationship between them is not one-to-one or one-to-many,

and the synergy mechanism between POIs and RSIs has rarely been studied. Simple fusion strategies (stacking or adding) [41,43] have difficulty taking full advantage of the complementary advantages of multimodal data. Additionally, fine-grained spatial relationships are important for identifying UFZs for the following reasons [44]. The first is that the layout of urban functional zones itself helps to reduce visual ambiguity. For example, determining whether a retail area is a commercial or residential area requires information about the surrounding area. Second, aggregating spatial information between and within different urban functional areas is conducive to determining the types of UFZs. For example, there are often recreational parks next to residential areas, while there are few large green-space areas around industrial areas. This co-occurrence relationship is very clear to some extent.

To address the above problems, we built a unified deep-learning framework, Lufz-CrossFormer (Lu-CF), which can efficiently exploit the complementary advantages between RSIs and POIs and simultaneously capture the potential spatial relationships of urban functional zones. The framework is divided into two parts: a complementary feature fusion module and a spatial-relationship-modeling module. In the first module, we use a common CNN to extract visual features from remote-sensing imagery data and a CNN with a gated attention mechanism to extract social features from point-of-interest data. To utilize the POIs conveniently, we first convert them into the corresponding hierarchical distance heatmap tensor according to the number of POI categories. Then, we use a layer-weighted model (LWM) to capture the possible co-occurrence relationship between POIs and UFZs. Finally, an adaptive feature fusion module (FFM) is used to efficiently fuse the two types of features. In the second part, the fused features are recoded according to the location relationship. We designed a new network structure based on the long-short-distance attention (LSDA) network CrossFormer [45], which can simultaneously capture the local spatial relationships and global spatial relationships of the fused features. Finally, a feature aggregation module (FGM) is used to efficiently utilize the two spatial relationships for more accurate urban functional area recognition. The main contributions of this paper are summarized as follows:

(1) We designed a unified deep-learning framework and integrated remote-sensing images and POIs to recognize urban functional zones. Our method can extract visual features, social features, and spatial relationship features from different data for more accurate urban functional zone recognition, while existing relevant studies rarely take into account all three features simultaneously.
(2) We investigated which POI categories have a greater impact on the final urban functional zone recognition accuracy, as well as the advantages of using POI data compared to using RSI data, through a series of experiments, which contributes to a further understanding of the role of multimodal data in the urban functional zone recognition task.
(3) The synergy mechanism of remote-sensing image data and point-of-interest data in the urban functional zone recognition task has rarely been studied. In this study, we used a feature fusion module to adaptively fuse the visual and social features and further analyzed the specific effects of this synergy mechanism for different urban functional areas.

2. Data and Methods

2.1. Data Source

As shown in Figure 1, we used the CSU-RSISC10 dataset [44] as the main research dataset, which was collected via Google Earth in Santa Monica, a coastal city in Los Angeles, USA, that covers an area of approximately 20 km^2. It contains 288 sample-level images with a pixel resolution of 2000 $\times$ 2000 and a spatial resolution of 0.15 m per pixel. Each sample-level image was further divided into 400 nonoverlapping patch-level images of size 100 $\times$ 100. Figure 2 illustrates the two-level hierarchical structure of this dataset. The first part of the model uses unordered patch-level images as training data. The second part uses sample-level images containing spatial location information for training. In our

experiments, we chose 70% of the data as the training set and the remaining data as the test set. For the patch-level data, the distribution of the number of different UFZ categories in the training and test sets is shown in Table 1. In addition, we downloaded 22,162 POIs from the open-source OSM website and reclassified them into 9 classes: airport, industry, supermarket, retail, hotel, institution, public service, nature, and residence. The distribution of the number of each type of POI is shown in Table 2. Finally, according to the "Code for Classification of Urban Land Use and Planning Standards of Development Land (GB 50137-2011)" and the LBCS standards, we classified the urban functional zones into commerce, industry, residence, construction, institution, transport, open space, and water.

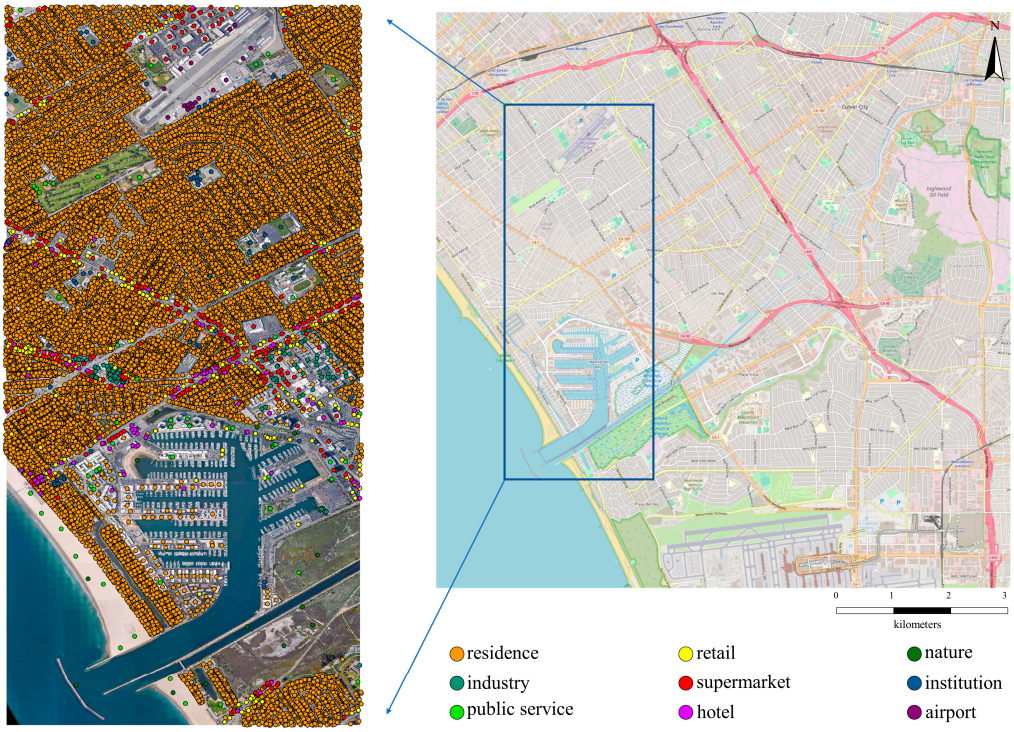

Figure 1. Overview of the dataset used in this study.

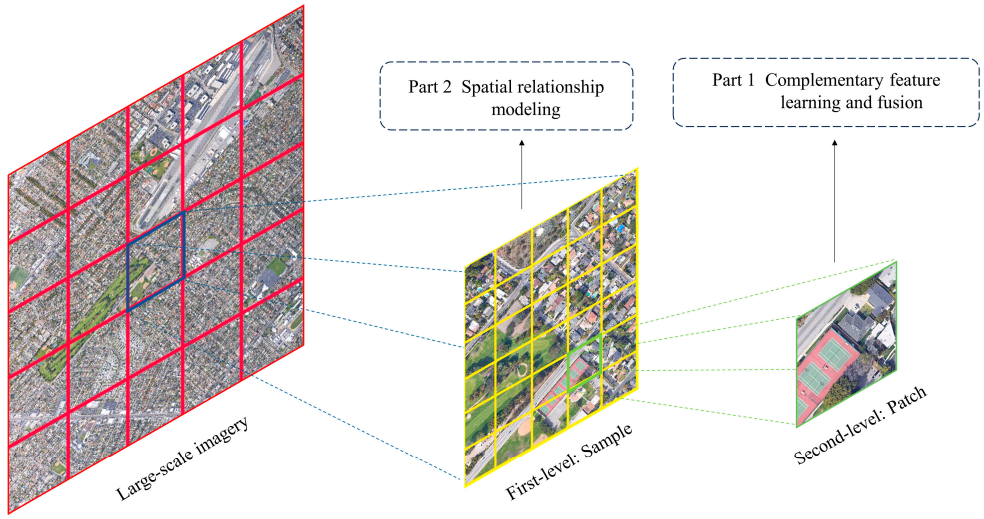

Figure 2. Two-level hierarchy of the dataset.

Table 1. Quantitative distribution of different UFZs in training and test sets.

Category	Commerce	Industry	Residence	Construction	Institution	Transport	Open Space	Water
Training set	6011	1396	30,407	346	1764	20,744	9418	10,714
Testing set	1269	698	15,127	206	570	8364	4811	3355

Table 2. Quantitative distribution of different POI categories.

Category	Residence	Retail	Supermarket	Nature	Hotel	Public Service	Institution	Industry	Airport
Number of POIs	21,045	447	310	13	91	56	62	118	20

2.2. Methods

The overall structure of the network framework established in this paper is shown in Figure 3. In the first part, we use a CNN to extract visual and social features while fusing them efficiently. In the second part, we design an architecture based on the CrossFormer network to extract local spatial relationships and global spatial relationships from the fused features generated in the first part and utilize them efficiently for the final classification output. The following subsections explain the details of the proposed framework.

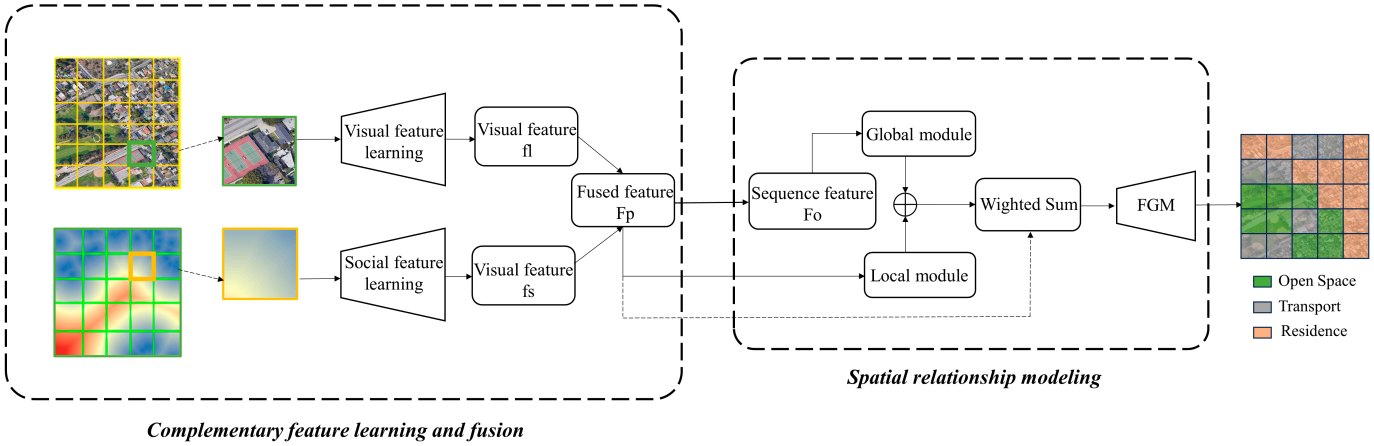

Figure 3. The overall structure of the proposed framework.

2.2.1. Complementary Feature Learning and Fusion

In this part, we use two CNNs to extract visual features and social features from RSIs and POIs. Then, the two types of features are fused efficiently to obtain fused features with strong representation ability based on the FFM.

Visual Feature Learning

First, we crop a large-scale three-band RGB RSI with a size of 3 × h × w according to the window step (h × w) to obtain a group of image units $n_j\left(j = 1, 2, 3, \ldots, \frac{H}{h} \times \frac{W}{w}\right)$, where each individual image unit represents a type of urban functional area. Then, ResNet-50 [46] is used as the backbone network to extract features from these image units, and a global average pooling operation is used to obtain the final visual feature $f_l \in \mathbb{R}^{2048\times1\times1}$.

Social Feature Learning

Many UFZs are visually similar; for example, the appearance of commercial buildings and institutional buildings may look extremely similar. Therefore, relying on remote-sensing imagery alone will not yield satisfactory results in many cases. We introduce POI data with rich social attributes as complementary features to RSI data and use a CNN with

an attention mechanism to capture possible co-occurrence relationships between urban functional zones and the corresponding points of interest.

Specifically, we first convert them into the corresponding hierarchical distance heatmap tensor according to the number of POI categories and input it into the convolutional neural network, as the CNN is better at processing two-dimensional continuous image data. Assume that the region contains M types of POIs, and the number of each POI category is n_u (u = 1, 2, . . ., M). For the uth-category point of interest, we calculate the minimum distance from pixel point (x_i, y_i) in the remote-sensing image to n_u POIs according to Equation (1) to generate the distance heatmap DP:

$$DP = MIN\left(\sqrt{(x_i - x_l)^2 + (y_i - y_l)^2}\right) \tag{1}$$

where (x_l, y_l) represents the spatial coordinates of the uth-category point of interest, $l = 1, 2, \ldots, n_u$. As a result, we obtain M types of distance heatmaps and then stack them to form a distance heatmap tensor $\mathrm{T} \in \mathbb{R}^{\mathrm{M}\times\mathrm{H}\times\mathrm{W}}$. Similar to the remote-sensing-image-cropping method, we crop the heatmap tensor T into $H/h \times W/w$ nonoverlapping patches: $t_j \in \mathbb{R}^{M\times\mathrm{h}\times\mathrm{w}}\left(j = 1, 2, \ldots, \frac{H}{h} \times \frac{W}{w}\right)$. Second, the relationships between POIs and UFZs are not one-to-one or one-to-many. An urban functional district may have multiple types of POIs, or a POI may appear in different urban functional zones simultaneously. For example, there will be retail POIs in both commercial and residential areas. To this end, inspired by Lu's approach [42], we adaptively weight t_j by adding a layer-weighted model (LWM) to the CNN to capture possible co-occurrence relationships between different types of POIs and UFZs. The structure of the LWM model is shown in Figure 4. We first use global average pooling encoding t_j to obtain an intermediate heatmap tensor $p_j \in R^{\mathrm{M}\times 1}$, and then two fully connected (FC) layers are used to obtain the interlayer weights:

$$w_j = \varphi\left(w_2\, \delta\left(w_1 p_j\right)\right) \in \mathbb{R}^M \tag{2}$$

where $\delta(x)$ is the ReLU activation function, and $\varphi(x)$ is the sigmoid function. The first FC layer has a learnable weight $w_1 \in \mathbb{R}^{D_{r1}\times M}$ to extend the dimension of the feature t_j, and the second FC layer has a learnable weight $w_2 \in \mathbb{R}^{M\times D_{r1}}$ to reduce the output dimension of the first FC layer. Then, we use the output w_j to adaptively weight the layers of t_j by performing layerwise multiplication between t_j and w_j:

$$\left(t'_j = w_j \times t_j\right) \in \mathbb{R}^{M\times h\times w} \tag{3}$$

where t'_j represents the heatmap tensor after adaptive weighting. Finally, we incorporate the LWM into the CNN to extract the social feature $f_s \in \mathbb{R}^{2048\times 1\times 1}$.

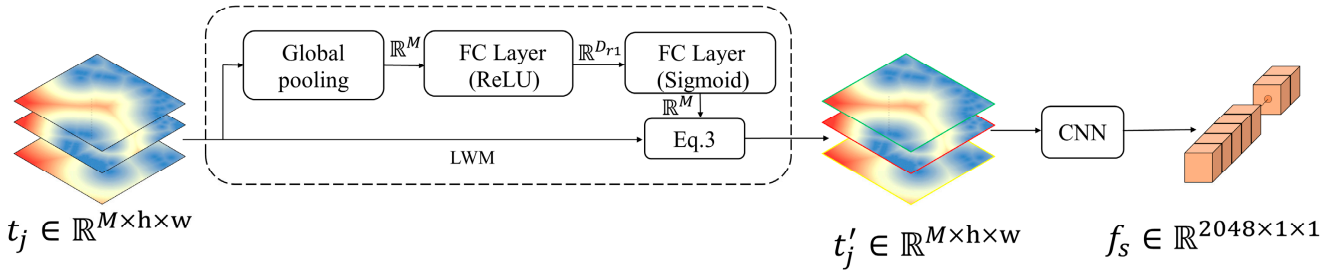

Figure 4. The structure of the layer-weighted model.

Complementary Feature Fusion

As we can imagine, visual features are effective for identifying certain urban functional areas, such as water and many open spaces. However, for commercial, industrial, and institutional areas, we need social features that contain rich socioeconomic attributes

for more accurate recognition, suggesting that visual and social features have different discriminative abilities for different UFZ categories. Therefore, similar to the Social Feature Learning section, we use an adaptive feature fusion module (FFM) for better feature fusion. As shown in Figure 5, specifically, the module can learn the adaptive fusion weights from the visual feature f_l and the social feature f_s, as shown in Equation (4):

$$w_f = \varphi(w_4\ \delta(w_3 f_c)) \tag{4}$$

where $f_c = \begin{bmatrix} f_l \\ f_s \end{bmatrix} \in \mathbb{R}^{4096}$. The first FC layer is a downsampling layer to reduce the output dimension, and its weight is $w_3 \in \mathbb{R}^{\frac{4096}{r} \times 4096}$; the second FC layer is a rescaling layer to compute the feature's weight factor, and r is the rescaling factor.

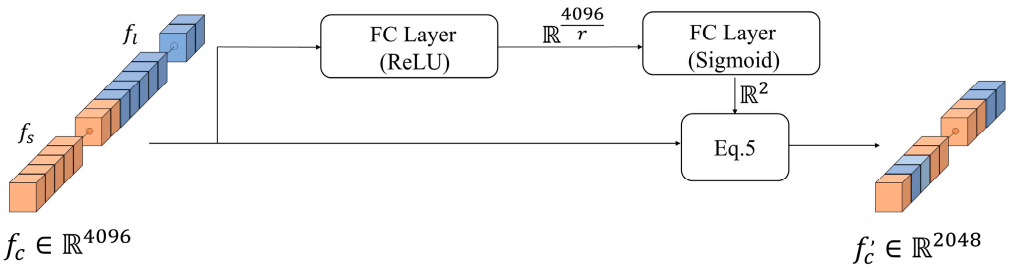

Figure 5. The structure of the adaptive feature fusion module.

$w_f = [w_l\ w_s]$ is a feature tensor of length 2, and w_l and w_s denote the ability of visual and social features to discriminate between different urban functional zones, respectively. Then, through Equation (5), we obtain the final fused feature f_c' by adaptively weighting the visual features and social features:

$$f_c{}' = w_f f_c = \left([w_l\ w_s] \begin{bmatrix} f_l \\ f_s \end{bmatrix} \right) \in \mathbb{R}^{2048} \tag{5}$$

2.2.2. Spatial Relationship Modeling

In the first part, we convert the input into a group of fused features $f_c'\left(c = 1,\ 2,\ 3, \ldots, \frac{H}{h} \times \frac{W}{w}\right)$. After that, we construct the feature tensor F_p according to the position relation of the tensor f_c', as defined in Equation (6):

$$F_p = \begin{pmatrix} f_1{}' & \cdots & f'_{\frac{W}{w}} \\ \vdots & \ddots & \vdots \\ f'_{\frac{W}{w} \times (\frac{H}{h} - 1) + 1} & \cdots & f'_{\frac{HW}{hw}} \end{pmatrix} \in \mathbb{R}^{2048 \times \frac{H}{h} \times \frac{W}{w}} \tag{6}$$

Each tensor f_c' in the feature tensor F_p corresponds to patch-level imagery clipping from large-scale imagery. Then, a simple convolutional network is used to extract local spatial relationships. Finally, we convert the feature F_p into a one-dimensional sequence F_o and input it into the global module to extract the global spatial relationships:

$$F_o = \left(f_1', f_2', f_3', \ \ldots\ , f'_{\frac{HW}{hw}} \right) \in \mathbb{R}^{L \times E} \tag{7}$$

where $L = \frac{H}{h} \times \frac{W}{w}$ is the sequence length, and E is the embedding dimension.

As shown in Figure 6, the local spatial relationships can be obtained through two groups of parallel convolution operations, and four groups of CrossFormer blocks (CF blocks) are used to obtain the global spatial relationships. The global features and local features are added, and the results are weighted sums with the original input F_p. Finally, a feature aggregation module (FGM) is used to obtain the final global–local context.

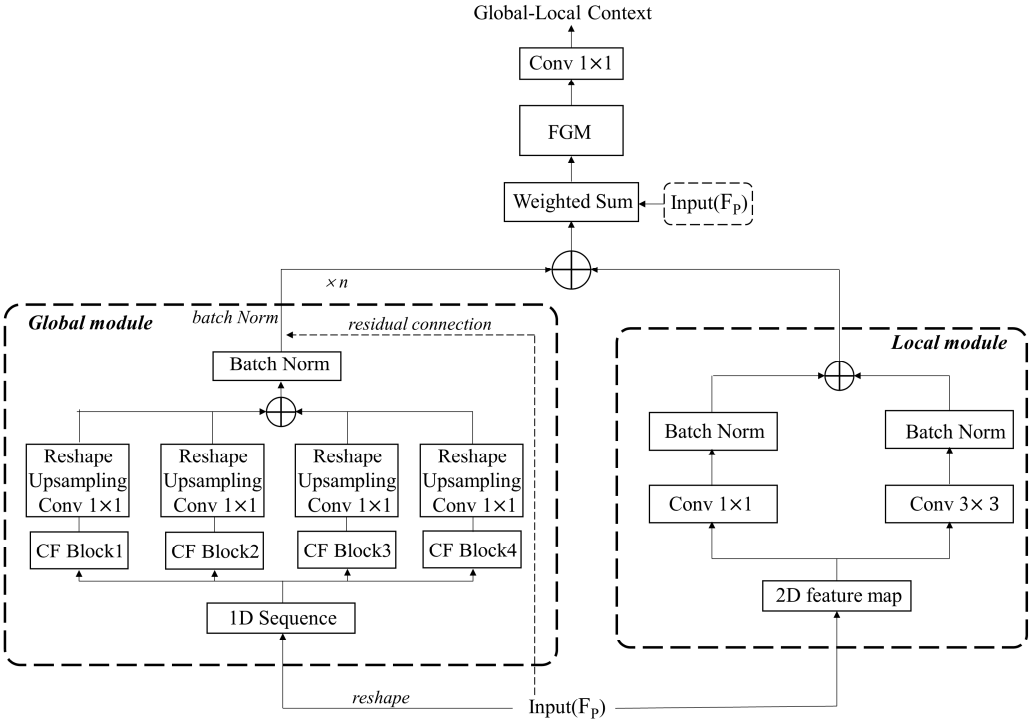

Figure 6. Illustration of the spatial-relationship-modeling module.

Local Spatial Relationship Modeling

While global spatial relationships are crucial to identifying complex UFZs, local information is also important for maintaining rich spatial details. As shown in the right half of Figure 6, we use two groups of parallel convolutions containing batch normalization operations to extract local information, with convolution kernel sizes of 1 and 3, and finally perform a sum operation.

Global Spatial Relationship Modeling

On a large scale, global information represents the relationships between different urban functional areas or a UFZ and its sub-UFZs. Some studies [19,47] have shown that vision-transformer-based structures have unique advantages in capturing global spatial relationships. In this paper, we present a network architecture designed to capture the global spatial relationships of the urban functional district based on CrossFormer. As shown in the left half of Figure 6, we first transform the original input F_p into a one-dimensional sequence and then input it into the CrossFormer blocks sequentially. As shown in Figure 7a, a CF block consists of a series of modules containing short-distance attention (SDA), long-distance attention (LDA), dynamic position bias (DPB), and multilayer perception (MLP). After that, we perform reshaping, upsampling, and convolution operations to restore the output to the size of the input. Finally, a residual connection is used to prevent model degradation, and we stack 6 global modules to obtain the final global features. In the following, we explain the details of the CrossFormer block.

(a) Dynamic Position Bias

The commonly used relative position bias (RPB) represents the relative embedding positions by adding a bias to the self-attention. The following equation represents the use of RPB to represent long-short-distance attention:

$$Attn = softmax\left(QK^T/\sqrt{d} + B\right)V \quad (8)$$

where $Q, K, V \in \mathbb{R}^{G^2 \times D}$ denote query, key, and value in the self-attention module, respectively, G is the group size, and D is the dimension of the embeddings. $\sqrt{d}$ is a constant, and $B \in \mathbb{R}^{G^2 \times G^2}$ is the RPB matrix. In the past, $B_{i,j} = B'_{\Delta x, \Delta y}$, where B' is a fixed-size matrix, and $(\Delta x, \Delta y)$ is the coordinate distance of the ith and jth embeddings. When the size of $(\Delta x, \Delta y)$ is larger than B', the size of the image or group is limited. The MLP-based module DPB is designed to solve the following problem:

$$B_{i,j} = DPB(\Delta x, \Delta y) \tag{9}$$

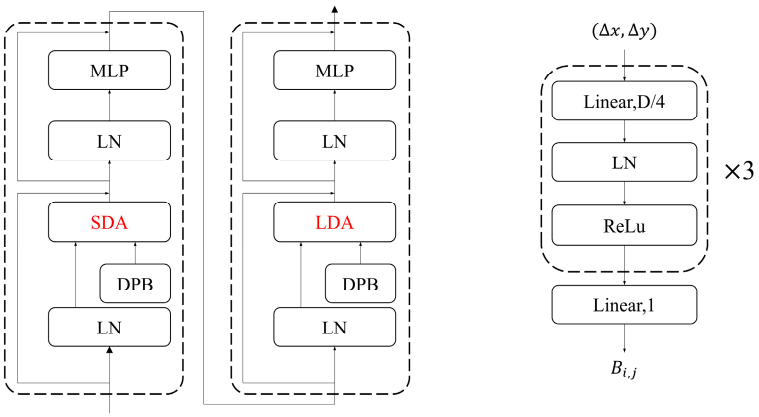

Figure 7. The structures of CF block and DPB. (**a**) CF block; (**b**) DPB.

As shown in Figure 7b, the structure comprises three FC layers containing layer normalization (LN) and ReLU activation, and the dimension of the middle layer is set to D/4. The output $B_{i,j}$ is a scalar representing the relative positions of the ith and jth embeddings. The DPB is a trainable module that can be optimized along with the entire model; it can handle arbitrary input group sizes without being limited by $(\Delta x, \Delta y)$.

(b) Long- and Short-Distance Attention

The self-attention module in the CrossFormer block is divided into two parts: short-distance attention (SDA) and long-distance attention (LDA). For SDA, we use a group of varying windows (G × G) to divide adjacent embeddings, and Figure 8a illustrates the case where G = 3. Unlike the fixed window of the Swin Transformer, SDA has a variable group size. In this experiment, the parameter G of the four CF blocks is {2, 4, 5, 4}. For LDA, given an input of M × M, the embeddings are sampled at a fixed interval I. As shown in Figure 8b (M = 9, I = 3), all of the embeddings belonging to the red boxes form a group, and the parts belonging to the purple boxes form another group. The width and height of the group can be computed by G = M/I. After grouping embeddings, both SDA and LDA compute the self-attention within their respective groups. As a result, the computational cost of the self-attention module will be reduced while capturing the fine-grained spatial relationships efficiently.

Feature Aggregation Module

The original feature F_P retains rich spatial details but lacks semantic attributes. Additionally, the global–local feature has fine-grained semantic information, but its spatial resolution is insufficient. Therefore, adding the two directly may reduce the classification accuracy [48]. In the method described in this paper, we use a feature aggregation module (FGM) to narrow the semantic gap between the two types of features to achieve more accurate UFZ recognition.

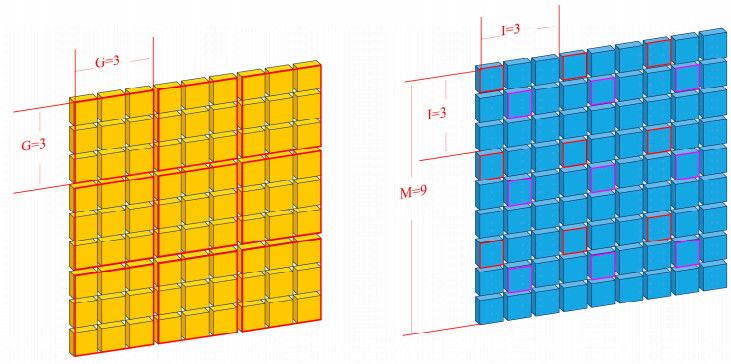

Figure 8. The structure diagrams of SDA and LDA. (**a**) SDA; (**b**) LDA.

First, a weighted-sum operation is performed on the two types of features, and the weights can be updated during model training to make full use of the rich spatial details and precise semantic attributes of different features. As shown in Figure 9, after a convolution operation (3×3), the fused features are input into the spatial path and channel path. Second, the two paths designed in the model help to strengthen the channel-based and space-based feature representation. Specifically, for the spatial path, the model uses a depth-separable convolution to produce a space-based attention feature $S \in \mathbb{R}^{h \times w \times 1}$, where h and w represent the spatial resolution of the feature map. After processing by the sigmoid function, matrix multiplication is used to obtain the path output. For the channel path, we first use the global average pooling operation to obtain the channel-based attention feature $C \in \mathbb{R}^{1 \times 1 \times c}$, where c represents the channel dimension. In addition, the rescaling operation consists of two convolution layers of 1×1, reducing and then restoring the dimension of the channel by a fixed factor. Similar to the spatial path, the sigmoid function and matrix multiplication are used to obtain the final path output. Finally, we use a convolution layer of 1×1 to obtain the FGM output.

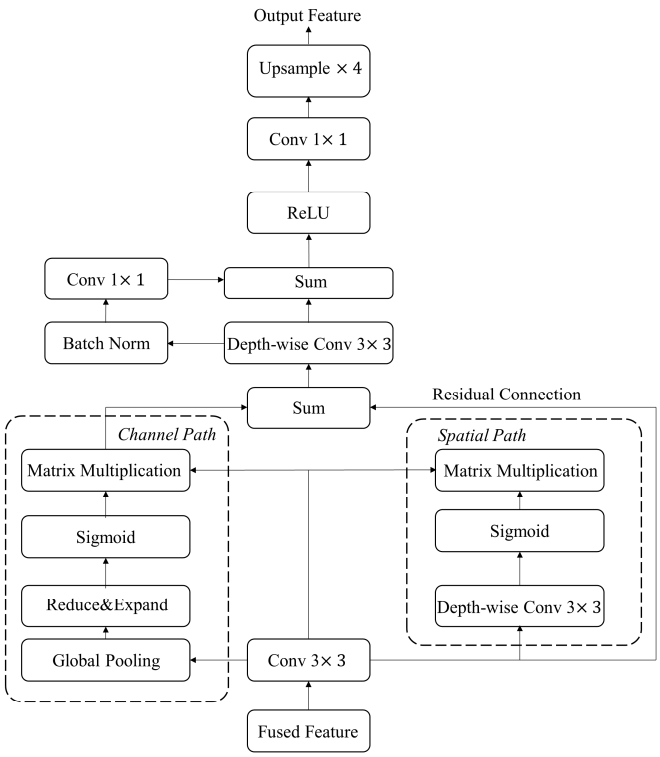

Figure 9. The structure of the FGM.

3. Experimental Analysis

3.1. Implementation Details

In the first part, we use ResNet-50 as the backbone network to extract visual features from the remote-sensing imagery. Meanwhile, the points of interest are transformed into a hierarchical distance heatmap tensor and input into the convolutional neural network to extract social features. Finally, a module based on an attention mechanism is used to efficiently fuse the two types of features. To train this part, we used cross-entropy loss (CE) and the Adam optimizer. The batch size was set to 16; the learning rate was initially set to 1×10^{-5}, and every five epochs, it became 0.98 times the original.

In the second part, we use two modules to extract global spatial relationships and local spatial relationships from the fused features, and then a feature aggregation module is used to obtain the final global–local context. After the model of the first part was trained, we saved the fused features obtained in this part to a local computer as input for the second part. Meanwhile, the loss function, optimizer, batch size, and learning rate adjustment strategies used in this part were exactly the same as in the first part.

We used the PyTorch deep-learning framework for training based on the Windows 10 operating system with an NVIDIA GeForce RTX 3090 graphics card (memory 24 GB). For the complementary feature-learning and fusion part, the model converged after 200 epochs; for the spatial-relationship-modeling part, the model converged after 150 epochs.

3.2. Evaluation Metrics

In the experiment, we used the kappa coefficient (*Kappa*) to evaluate the comprehensive performance of each model, the *F*1 score to measure the accuracy of each category, and the overall accuracy (*OA*) to evaluate the overall classification accuracy, as shown in the following equations:

$$\begin{aligned} OA &= \frac{TP+TN}{TP+TN+FP+FN} \\ precision &= \frac{TP}{TP+FP} \\ recall &= \frac{TP}{TP+FN} \\ F1 &= \frac{2\times precision\times recall}{(precision+recall)} \end{aligned} \tag{10}$$

$$\begin{aligned} p_e &= \frac{a1\times b1+a2\times b2+\cdots+aC\times bC}{n\times n} \\ Kappa &= \frac{p_o-p_e}{1-p_e} \end{aligned} \tag{11}$$

where *TP* denotes the number of pixels correctly classified into the positive class, *TN* denotes the number of pixels correctly classified into the negative class, *FP* denotes the number of pixels misclassified into the positive class, and *FN* denotes the number of pixels misclassified into the negative class. *Po* denotes the overall classification accuracy; $a1$, $a2$, …, aC denote the number of true samples in each category; $b1$, $b2$, …, bC denote the number of predicted samples in each category; and n is the total number of samples.

3.3. Experiments

3.3.1. Comparative Experiments for the Second Part

In the following experiment, we kept the structure of the first part unchanged and compared different models for the second part to further verify the importance of spatial relationships for urban functional zone recognition. The models we chose to compare include the hierarchical vision transformer networks based on the sparse attention mechanism Swin Transformer (SWINT) [49] and the pyramid vision transformer (PVT) [50], the long short-term memory network (LSTM) [51], and the gated recurrent neural network (GRU) [52] based on time-series modeling. For a fair comparison, we stack six global modules and one local module for all experimental models.

As shown in Figure 10, the area classified by our method is purer and the road structure is more complete compared with other methods. The main reason is that our model can

obtain richer composition patterns of spatial relationships by adding the variable group size and long-short-distance attention. As shown in the red circles in the first row, our method accurately identifies most of the institutional areas and has fewer misclassified spots, with a clearer road network compared to the results obtained by the GRU and LSTM. As marked by the red circles in the fourth row, our method obtains better results than SWINT and PVT due to the consideration of finer spatial relationships. In the sixth row, the classification results of our method are purer than those of the SWNT and PVT. Although the GRU and LSTM have no obvious misclassified spots in this region, the urban functional area edge recognized by them is not precise enough.

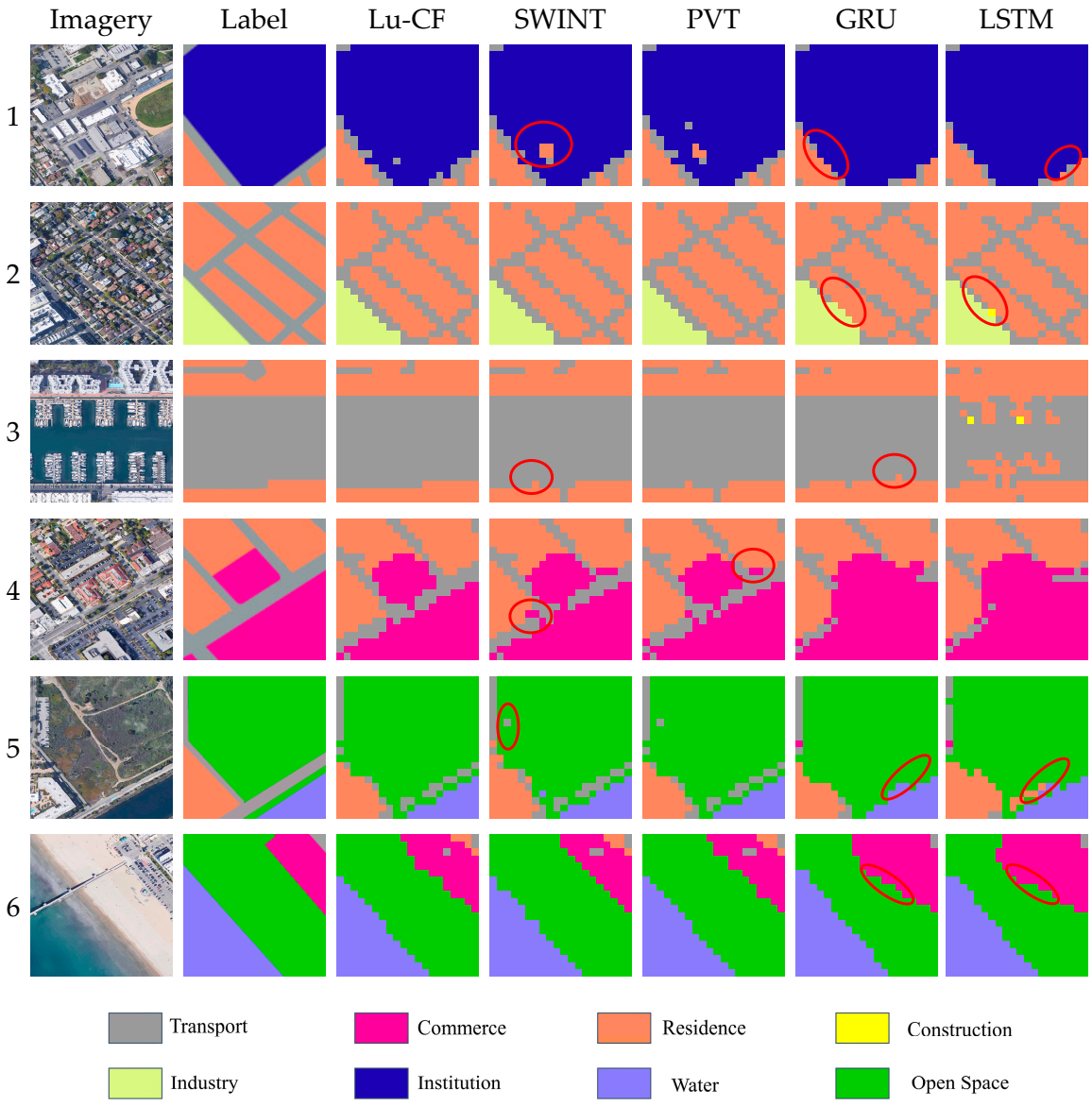

Figure 10. The test results of different models for the second part.

The quantitative evaluation results of the models are shown in Table 3. The GRU and LSTM can only capture the long-range dependence information in the horizontal and vertical directions, so they are the least effective. Our method improves *Kappa* by 1.03% over the second-best method and 7.04% over the worst method, and the overall accuracy is improved by at least 1.8% over other methods. In addition, the *F*1 scores of each UFZ category can further prove the advantage of our method. Our model has the best performance in six categories, especially the *F*1 score of the transport category,

which reaches 0.8820. Meanwhile, the accuracy performance is not satisfactory in the construction category due to the samples being scarce. For the other two categories, the accuracy gap between our method and the best method is less than 1%. The qualitative and quantitative evaluation results show that our model can make full use of the fine-grained spatial relationships for more accurate urban functional zone recognition.

Table 3. Quantitative evaluation results of different models for the second part.

Models	F1 Scores								*Kappa*	*OA*
	Transport	**Commerce**	**Residence**	**Construction**	**Industry**	**Institution**	**Open Space**	**Water**		
Lu-CF	0.8820	0.7865	0.9300	0.6528	0.8661	0.7514	0.9125	0.9596	0.8697	0.9136
SWINT	0.8796	0.7458	0.9280	0.6213	0.8669	0.7349	0.9109	0.9600	0.8594	0.8956
PVT	0.8794	0.7338	0.9276	0.6471	0.8690	0.7202	0.9105	0.9571	0.8587	0.8924
LSTM	0.7805	0.6578	0.9146	0.4512	0.8392	0.7041	0.8772	0.9474	0.7993	0.8451
GRU	0.8196	0.6707	0.9225	0.5325	0.8139	0.6938	0.8915	0.9476	0.8276	0.8762

3.3.2. Comparative Experiments with Other Methods

In this part, we used three related methods and a benchmark model for comparative experiments. The first one is the RPFM (Remote Sensing Images and Point-of-Interest Fused Model) [41], which also uses the image patch as the basic mapping unit, and it represents the relationship between POIs and urban region functions based on the distance metric. The second one is DHAO (deep integration of high-resolution imagery and OpenStreetMap) [53], which uses statistical features from POI data and visual features from RSI data for urban functional zone classification, and the road network is used as the basic mapping unit. The third one is the SO-CNN (Super-Object-based CNN) [21], which uses only remote-sensing image data to identify UFZs, and the super object is used as the basic mapping unit. In addition, the benchmark model is the widely used UNet network [20], which only utilizes image data to identify urban functional areas.

On the one hand, social perception information is crucial for accurate UFZ recognition, and our method can sufficiently capture it through the complementary feature-learning and fusion module. By learning rich social features, our approach can identify institutional areas, commercial areas, and open spaces more accurately. The results obtained by the SO-CNN are not accurate enough because only RSI data are considered. Although the RPFM and DHAO also use both POI data and RSI data, they simply incorporate the statistical characteristics of POI data and therefore fail to take full advantage of their social attributes. As a result, the inadequate representation can produce misclassified results, such as the red circles in Figure 11. We further give two specific examples. The region in the fourth row of Figure 11 consists of a large number of commercial areas as well as some residential areas. Accurately identifying commercial areas has always been a difficult problem in the UFZ classification task. Nevertheless, our method correctly identified almost all commercial areas in the region and has fewer misclassification spots. In the fifth row, our model accurately identified the area where the beach meets the ocean, while all other models misclassified this area as a transport category. On the other hand, spatial relationship features are also important for accurate UFZ classification. As marked by the yellow circles in Figure 11, the results obtained by our method are purer and more continuous due to the consideration of the spatial position relationships between different patches, while other methods only utilize the features of a single UFZ. In addition, the benchmark model obtained the worst results due to the model structure and use of a single data type.

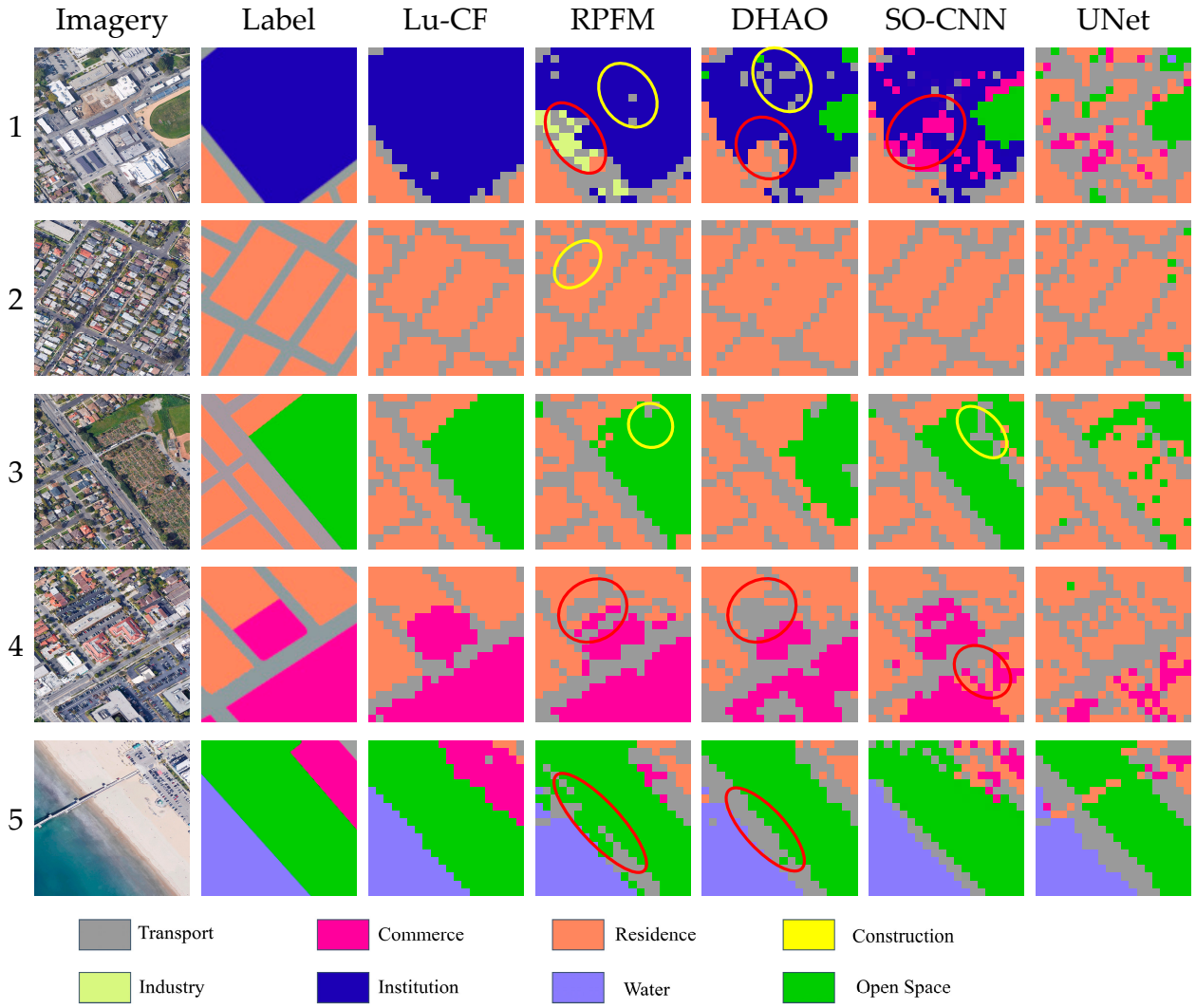

Figure 11. The test results of different models.

The quantitative evaluation results are shown in Table 4, which shows that our method obtains the highest scores on most of the metrics. The kappa value of our method is at least 2.84% higher than that of other methods, and it has an overall accuracy improvement of 3.84% over the second-best method. In addition, our model achieves the highest *F*1 scores on seven UFZ categories, with the scores improved by at least 6.50% and 3.61% for the commercial and institutional areas, respectively. The qualitative and quantitative evaluation results show that our model has a better overall performance in the urban functional zone recognition task.

Table 4. Quantitative evaluation results of different models.

Models	F1 Scores								*Kappa*	*OA*
	Transport	Commerce	Residence	Construction	Industry	Institution	Open Space	Water		
Lu-CF	0.8820	0.7865	0.9300	0.6528	0.8661	0.7514	0.9125	0.9596	0.8697	0.9136
RPFM	0.8613	0.7215	0.9063	0.6276	0.8479	0.7153	0.8995	0.9432	0.8413	0.8752
DHAO	0.8143	0.6736	0.8965	0.5928	0.8035	0.6814	0.8719	0.9361	0.8167	0.8567
SO-CNN	0.8332	0.6961	0.9325	0.6016	0.8113	0.6976	0.8817	0.9456	0.8325	0.8643
UNet	0.7516	0.6198	0.8746	0.4332	0.7652	0.5678	0.7764	0.9175	0.7361	0.7623

3.3.3. Ablation Experiment

In this section, we analyze the specific contribution of each module in the proposed framework to the final urban functional zone recognition accuracy through a series of ablation experiments. The results of the ablation experiments are shown in Table 5. We conducted experiments by continuously adding new modules to verify the contribution of different modules to the UFZ recognition accuracy. For example, using both RSI data and POI data (Stage 2) improves the *Kappa* value by 0.0793 and *OA* value by 0.0779 compared to using only remote-sensing imagery (Stage 1), and the use of the LWM module (Stage 3) increases the *Kappa* value to 0.8246; meanwhile, the addition of the FFM (Stage 4) increases the *Kappa* value from 0.8246 to 0.8379. The spatial relationship modules also have an important impact on improving the urban functional district classification accuracy. After adding the local module and global module (Stage 7), the *Kappa* value increases from 0.8379 to 0.8654. Finally, our model obtains the optimal performance (*Kappa* = 0.8697, *OA* = 0.9136) after adding the FGM (Stage 8). It is worth noting that including only the local module (Stage 6) leads to a greater accuracy improvement compared to including only the global module (Stage 5), suggesting that the local spatial relationship is also crucial to accurate UFZ classification. The results of the ablation experiments show that all of the modules of our framework have a positive effect on improving the urban functional area recognition accuracy, and using POI data leads to the greatest improvement in accuracy; therefore, we discuss the relationship between points of interest and urban functional zone recognition accuracy in the following section.

Table 5. The results of the ablation experiment. RSI: use of remote-sensing imagery. POI: use of distance heatmap tensor of POIs. LWM: layer-weighted module. FFM: feature fusion module. LOCAL: local spatial relationship modeling. Global: global spatial relationship modeling. FGM: feature aggregation module.

Stage	Strategy							*Kappa*	*OA*
	RSI	POI	LWM	FFM	LOCAL	GLOBAL	FGM		
1	√							0.7364	0.7639
2	√	√						0.8157	0.8418
3	√	√	√					0.8246	0.8579
4	√	√	√	√				0.8379	0.8713
5	√	√	√	√		√		0.8548	0.8925
6	√	√	√	√	√			0.8637	0.9012
7	√	√	√	√	√	√		0.8654	0.9058
8	√	√	√	√	√	√	√	0.8697	0.9136

4. Discussion

In this section, we discuss the contributions of different point-of-interest categories to the UFZ classification accuracy and the synergy mechanism between RSI data and POI data in the urban functional area recognition task. Finally, we visualize the layer activation of visual features.

4.1. Specific Impact of POIs on UFZ Recognition

The relationship between POIs and UFZs is not one-to-one or one-to-many and is complex in many cases. To determine which point-of-interest categories have a more important impact on the urban functional zone classification accuracy, we deleted one POI category at a time and repeated the experiment. As shown in Table 6, we find that regardless of which POI types are deleted, the final recognition accuracy decreases, indicating that all POIs have a positive effect on improving the urban functional area classification accuracy. When we remove POIs in the institution, residence, industry, and public service categories, the model accuracy decreases the most, suggesting that these POI categories have a relatively strong association with the urban functional district. In contrast, deleting retail, hotel, and

supermarket POIs has less of an impact on the classification accuracy, mainly because these points of interest can occur in almost all urban functional area categories, and there is no clear representative relationship between them and the urban functional zone.

Table 6. Impact of removing different types of POIs on model accuracy.

Drop POI Category	Institution	Residence	Industry	Nature	Airport	Public Service	Retail	Hotel	Supermarket
Kappa	0.8391	0.8345	0.8456	0.8479	0.8527	0.8443	0.8613	0.8634	0.8625
Decreasing rate of Kappa (%)	3.52	4.05	2.77	2.51	1.95	2.96	0.97	0.72	0.83
Using all POIs	*Kappa* = 0.8697								

In addition, we investigated the issue of the cost of using both remote-sensing images and points of interest simultaneously versus training with more remote-sensing imagery alone in the case of obtaining close classification accuracy. As Table 7 shows, training with both RSIs and POIs results in higher accuracy than training with only twice as much RSI data, which is close to the accuracy obtained when training with only three times as many remote-sensing images, further demonstrating the advantages of using the point-of-interest data.

Table 7. The advantages of using POI data.

RSI	POI	F1 Score								*Kappa*
		Transport	Commerce	Industry	Residence	Construction	Institution	Water	Open Space	
27,000	√	0.8339	0.6051	0.5642	0.8655	0.3156	0.3912	0.8965	0.8749	0.7257
54,000	×	0.8271	0.3925	0.2313	0.8552	0.3119	0.1756	0.9125	0.7912	0.6913
80,800	×	0.8615	0.4349	0.3723	0.8964	0.3586	0.3146	0.9324	0.8153	0.7294

4.2. Synergy Mechanism of POIs and RSIs in UFZ Recognition

The discriminative abilities of visual and social features for different urban functional zone categories are different; as we can imagine, visual features are more important for identifying water, while social features may be more crucial to identifying commercial or institutional areas. To further clarify this synergy mechanism, as shown in Figure 12, for patch-level data, we take the weights of visual feature w_l and social feature w_s to generate the point (w_l, w_s) and visualize the two-dimensional display of the discriminative ability of the different features for each UFZ category. We observe two phenomena. First, even for the same UFZ category, the fusion weights change dynamically, and the distribution of fusion weights for all urban functional areas tends to be relatively dispersed and concentrated. Second, visual features are more important for identifying urban functional zones in the transport and water categories, while social features are more crucial for recognizing the commerce, industry, institution, open space, and residence classes. However, for the residence and open space categories, although the social features are dominant, the complementary role of visual features becomes increasingly obvious. For the transport category, the rule is the opposite. For the construction area, the two types of features have a relatively equivalent status.

Based on the above discussion, we conjecture that this synergy mechanism is manifested by taking one feature as the main fusion object and the other as an auxiliary object. Nevertheless, for some urban functional area categories, when the discriminative ability of the dominant feature reaches a peak or even declines, the counterpart of the complementary feature continues to improve and approaches the dominant feature. As shown in Figure 13, we give two examples to further illustrate this synergy mechanism. In Figure 13a, it is difficult to determine whether the sports field belongs to an institutional area, residential area, or open space based on visual features alone, but the surrounding residential POIs provide

strong social signals indicating that the field belongs to a residential area; in Figure 13b, a large area of green space provides strong visual information indicating that the area belongs to the open space, while the points of interest are scarce in this area, and it will produce incorrect results if we only rely on the surrounding residential POIs. As a result, our model tends to give the social features a larger weight in Figure 13a and the visual features a larger weight in Figure 13b. Therefore, based on the different features obtained from RSIs and POIs, the synergy mechanism is a dynamic process in specific urban functional zones, and the experiments prove that the model in this paper can efficiently utilize this synergy mechanism for more accurate urban functional area recognition.

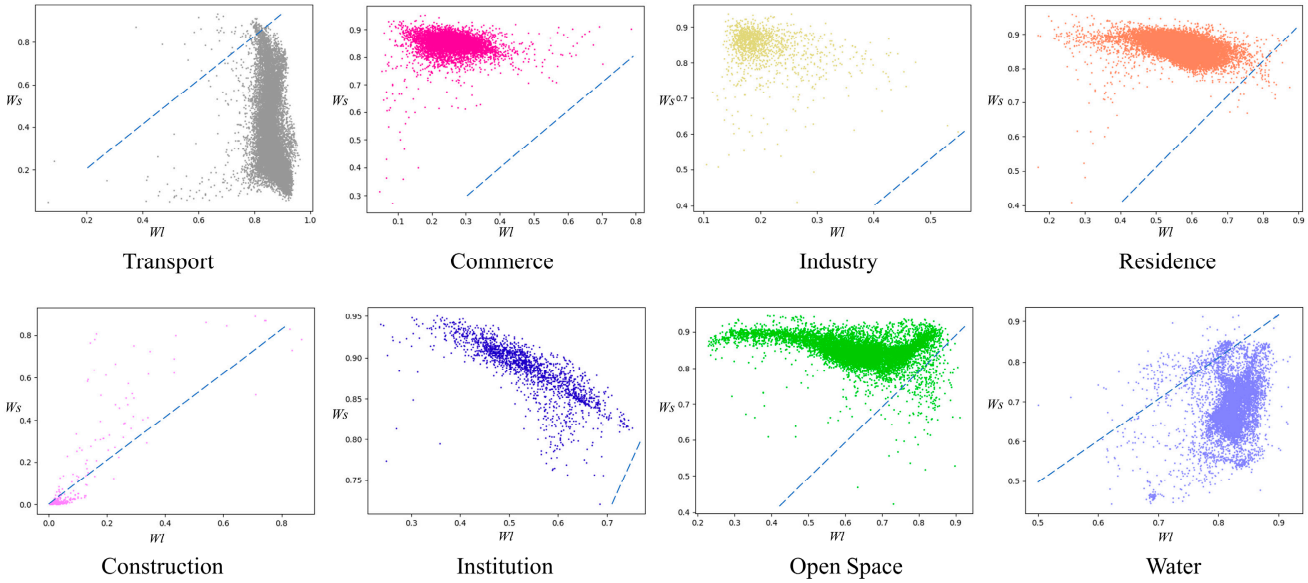

Figure 12. Two-dimensional visualization of discriminative ability of different features for different UFZ categories. *Ws* represents the weight of social features, *Wl* represents the weight of visual features, and the blue dotted line denotes the case where *Ws* = *Wl*. When the point is above the dotted line, social features are more important. When the point is below the dotted line, visual features are more important.

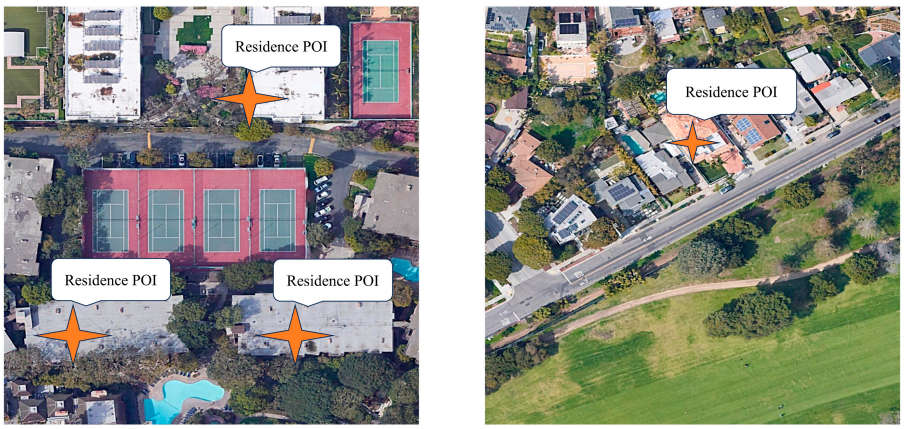

Figure 13. A case of the synergy mechanism. (**a**) Social features are more important; (**b**) visual features are more important.

4.3. Layer Activation Visualization of Visual Features

In the above sections, we focused on the importance of social features for urban functional zone classification; however, visual features are also crucial in almost all relevant tasks. For this purpose, we use the Grad-CAM [54] method to visualize the layer activation of the network for input during the visual feature extraction phase. As shown in Figure 14, we find that in the shallow stage (Layer 1 or Layer 2), the network tends to focus on the important regions of the image at a larger scale. When the number of layers deepens, the network focuses on a specific region of the image at a relatively small scale. Meanwhile, for urban functional districts containing more buildings, the edge and corner information of some buildings is more important; for the institutional area, we chose a relatively representative image and can see that the network mainly focuses on the areas of different surface materials on the roof; for green spaces belonging to the open space category, the network focuses on the texture information of the image; for water, the network tends to focus on the edge region or texture information of the image. As a result, visual features with strong representation ability lay a solid foundation for accurate urban functional zone recognition.

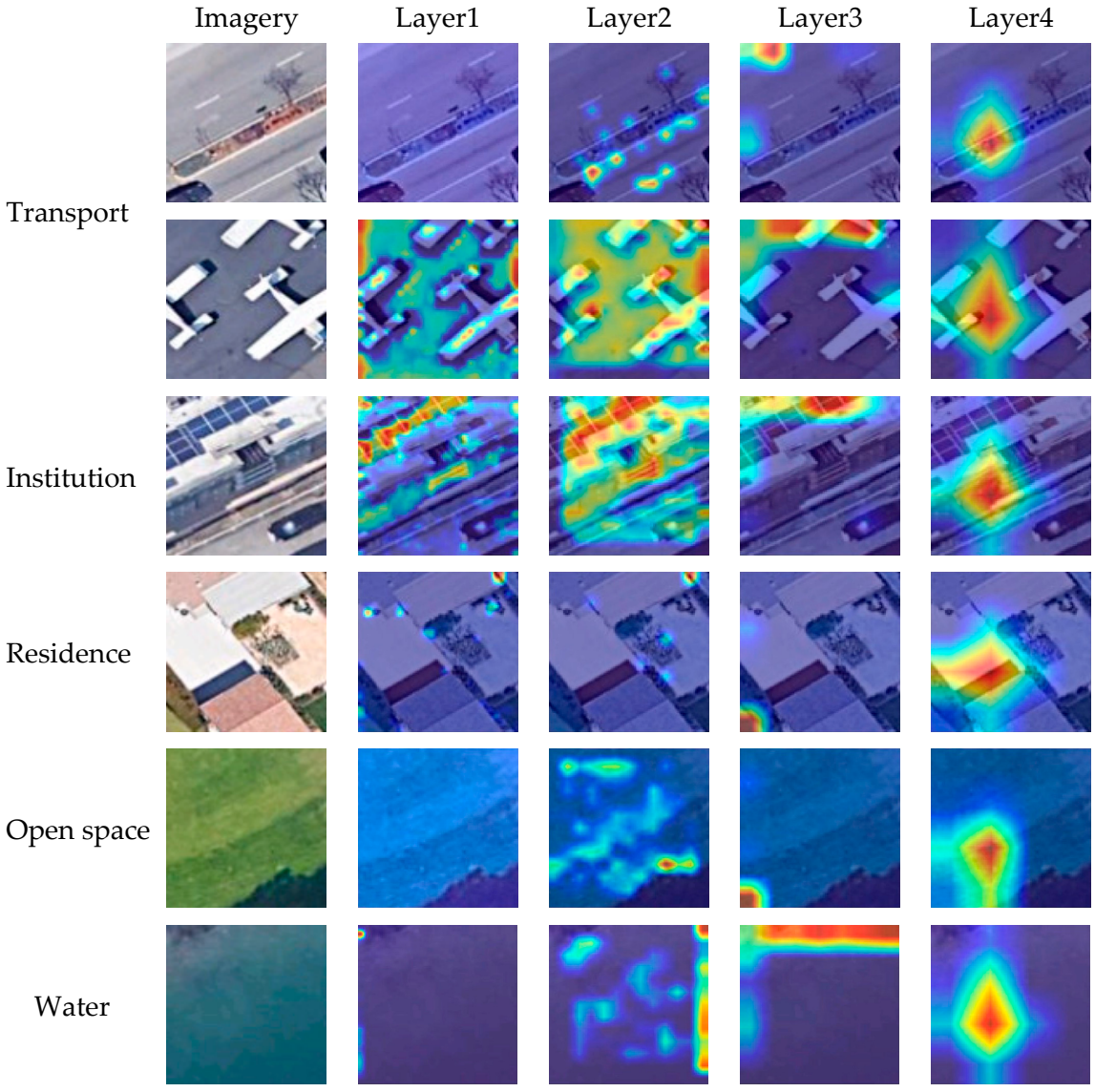

Figure 14. Layer activation visualization of visual features.

5. Conclusions

In this study, a unified deep-learning framework was designed for learning visual, social, and spatial relationship features simultaneously from remote-sensing imagery data and point-of-interest data for more accurate urban functional area recognition. In the complementary feature-learning and fusion module, we use two convolutional neural networks to extract visual features and social features from RSIs and POIs, respectively, and then the feature fusion module is used to fuse them efficiently. In the spatial-relationship-modeling module, taking the fused features obtained in the first part as input, we designed a new network structure based on the CrossFormer to extract the global and local spatial information of the urban functional zone distribution, and then a feature aggregation module is used to utilize the two spatial relationships efficiently. The comparative experimental results for the second part show that the *Kappa* value (0.8697) of our model on the test dataset is 1.03% higher than that of the second-best method (0.8594) and 7.04% higher than that of the worst method (0.7993). Additionally, it obtains the highest *F*1 scores in six urban functional zone categories, proving that the proposed model can effectively utilize the spatial relationship composition patterns of different urban functional zones to obtain more accurate results. Meanwhile, the results of comparative experiments with other methods show that the *Kappa* and *OA* values are improved by 2.84% and 3.84%, respectively, over the second-best method, proving that our method can effectively utilize the synergy mechanism based on different situations and integrates the visual features, social features, and spatial relationship features for more rational and effective urban functional area recognition tasks. In addition, we investigate the influence of different point-of-interest categories on the urban functional zone classification accuracy, the synergy mechanism of POIs and RSIs in the UFZ classification task, and the layer activation visualization of visual features. In particular, we demonstrate that the synergy mechanism is a dynamic process based on the discriminative ability of complementary features from RSIs and POIs in different UFZ categories, and our method can capture this mechanism to obtain a better urban functional zone recognition result. The framework in this paper is instructive for accurate urban functional area recognition using multimodal data. In the future, we will attempt to use more representative datasets and investigate the impact of different mapping units on the UFZ classification accuracy to achieve more efficient and accurate urban functional zone recognition tasks.

Author Contributions: Conceptualization, Haiqing Xu and Mingyang Yu; methodology, Haiqing Xu; software, Haiqing Xu; validation, Shuai Xu, Mingyang Yu and Hongling Yin; formal analysis, Haiqing Xu; investigation, Fangliang Zhou; resources, Mingyang Yu; data curation, Haiqing Xu; writing—original draft preparation, Haiqing Xu and Mingyang Yu; writing—review and editing, Mingyang Yu; visualization, Haiqing Xu; supervision, Mingyang Yu; project administration, Mingyang Yu; funding acquisition, Mingyang Yu. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China National Key R&D Program during the 13th Five-year Plan Period, grant number 2019YFD1100800, and the National Natural Science Foundation of China, grant number 41801308.

Data Availability Statement: The RSI dataset can be obtained from [44]. The POI dataset can be downloaded from https://export.hotosm.org/, accessed on 15 June 2023.

Acknowledgments: The authors thank the managing editor and anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tu, W.; Hu, Z.; Li, L.; Cao, J.; Jiang, J.; Li, Q.; Li, Q. Portraying urban functional zones by coupling remote sensing imagery and human sensing data. *Remote Sens.* **2018**, *10*, 141. [CrossRef]
2. Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping urban land use by using landsat images and open social data. *Remote Sens.* **2016**, *8*, 151. [CrossRef]

3. Zhang, Z.; Wang, Y.; Liu, Q.; Li, L.; Wang, P. A CNN based functional zone classification method for aerial images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10–15 July 2016; pp. 5449–5452.
4. Ge, P.; He, J.; Zhang, S.; Zhang, L.; She, J. An integrated framework combining multiple human activity features for land use classification. *ISPRS Int. J. Geoinf.* **2019**, *8*, 90. [CrossRef]
5. Song, J.; Tong, X.; Wang, L.; Zhao, C.; Prishchepov, A.V. Monitoring finer-scale population density in urban functional zones: A remote sensing data fusion approach. *Landsc. Urban Plan.* **2019**, *190*, 103580. [CrossRef]
6. Yu, B.; Wang, Z.; Mu, H.; Sun, L.; Hu, F. Identification of urban functional regions based on floating car track data and POI data. *Sustainability* **2019**, *11*, 6541. [CrossRef]
7. Zhang, Y.; Li, Q.; Huang, H.; Wu, W.; Du, X.; Wang, H. The combined use of remote sensing and social sensing data in fine-grained urban land use mapping: A case study in Beijing. *Remote Sens.* **2017**, *9*, 865. [CrossRef]
8. Banzhaf, E.; Netzband, M. Monitoring urban land use changes with remote sensing techniques. In *Applied Urban Ecology: A Global Framework*; Wiley: Hoboken, NJ, USA, 2011; pp. 18–32.
9. Herold, M.; Couclelis, H.; Clarke, K.C. The role of spatial metrics in the analysis and modeling of urban landuse change. *Comput. Environ. Urban Syst.* **2005**, *29*, 369–399. [CrossRef]
10. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Dalal, N.; Bill, T. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
12. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
13. Li, M.; Stein, A.; Bijker, W.; Zhan, Q. Urban land use extraction from very high resolution remote sensing imagery using a Bayesian network. *ISPRS J. Photogramm. Remote Sens.* **2016**, *122*, 192–205. [CrossRef]
14. Zhang, X.; Du, S.; Wang, Q. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sens. Environ.* **2018**, *212*, 231–248. [CrossRef]
15. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–16 July 2017; pp. 1251–1258.
16. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014; pp. 1–14.
17. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2 August 2017; p. 30.
19. Wang, L.; Li, R.; Zhang, C.; Fang, S.; Duan, C.; Meng, X.; Atkinson, P.M. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 196–214. [CrossRef]
20. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
21. Zhou, W.; Ming, D.; Lv, X.; Zhou, K.; Bao, H.; Hong, Z. SO–CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sens. Environ.* **2020**, *236*, 111458. [CrossRef]
22. Du, S.; Du, S.; Liu, B.; Zhang, X. 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.* **2021**, *261*, 112480. [CrossRef]
23. Voltersen, M.; Berger, C.; Hese, S.; Schmullius, C. Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level. *Remote Sens. Environ.* **2014**, *154*, 192–201. [CrossRef]
24. Peng, F.; Weng, Q. A time series analysis of urbanization induced land use and land cover change and its impact on land surface temperature with Landsat imagery. *Remote Sens. Environ.* **2016**, *175*, 205–214.
25. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007. [CrossRef]
26. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
27. Zhang, X.; Du, S.; Wang, Q. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. *ISPRS J. Photogramm. Remote Sens.* **2017**, *132*, 170–184. [CrossRef]
28. Zhang, C.; Sargent, I.; Pan, X.; Li, H.; Gardiner, A.; Hare, J.; Atkinson, P.M. Joint deep learning for land cover and land use classification. *Remote Sens. Environ.* **2019**, *221*, 173–187. [CrossRef]
29. Yang, H.; Wolfson, O.; Zheng, Y.; Capra, L. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol.* **2014**, *5*, 1–55. [CrossRef]
30. Liu, X.; He, J.; Yao, Y.; Zhang, J.; Liang, H.; Wang, H.; Hong, Y. Classifying urban land use by integrating remote sensing and social media data. *Int. J. Geograph. Informat. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
31. Jia, Y.; Ge, Y.; Ling, F.; Guo, X.; Wang, J.; Wang, L.; Chen, Y.; Li, X. Urban land use mapping by combining remote sensing imagery and mobile phone positioning data. *Remote Sens.* **2018**, *10*, 446. [CrossRef]

32. Tu, W.; Cao, J.; Yue, Y.; Shaw, S.-L.; Zhou, M.; Wang, Z.; Chang, X.; Xu, Y.; Li, Q. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *Int. J. Geograph. Informat. Sci.* **2017**, *31*, 2331–2358. [CrossRef]
33. Gao, S.; Janowicz, K.; Couclelis, H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Trans. GIS* **2017**, *21*, 446–467. [CrossRef]
34. Cao, R.; Zhu, J.; Tu, W.; Li, Q.; Cao, J.; Liu, B.; Zhang, Q.; Qiu, G. Integrating aerial and street view images for urban land use classification. *Remote Sens.* **2018**, *10*, 1553. [CrossRef]
35. Zhu, Y.; Newsam, S. Land use classification using convolutional neural networks applied to ground-level images. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; ACM: New York, NY, USA, 2015; pp. 61:1–61:4.
36. Tu, W.; Cao, R.; Yue, Y.; Zhou, B.; Li, Q.; Li, Q. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. *J. Transp. Geogr.* **2018**, *69*, 45–57. [CrossRef]
37. Liu, Y.; Wang, F.; Xiao, Y.; Gao, S. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landsc. Urban Plann.* **2012**, *106*, 73–87. [CrossRef]
38. Qian, Z.; Liu, X.; Tao, F.; Zhou, T. Identification of urban functional areas by coupling satellite images and taxi GPS trajectories. *Remote Sens.* **2020**, *12*, 2449. [CrossRef]
39. Cao, R.; Tu, W.; Yang, C.; Li, Q.; Liu, J.; Zhu, J.; Zhang, Q.; Li, Q.; Qiu, G. Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS J. Photogramm. Remote Sens.* **2020**, *163*, 82–97. [CrossRef]
40. Song, J.; Lin, T.; Li, X.; Prishchepov, A.V. Mapping Urban Functional Zones by Integrating Very High Spatial Resolution Remote Sensing Imagery and Points of Interest: A Case Study of Xiamen, China. *Remote Sens.* **2018**, *10*, 1737. [CrossRef]
41. Xu, S.; Qing, L.; Han, L.; Liu, M.; Peng, Y.; Shen, L. A New Remote Sensing Images and Point-of-Interest Fused (RPF) Model for Sensing Urban Functional Regions. *Remote Sens.* **2020**, *12*, 1032. [CrossRef]
42. Lu, W.; Tao, C.; Li, H.; Qi, J.; Li, Y. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sens. Environ.* **2022**, *270*, 112830. [CrossRef]
43. Bao, H.; Ming, D.; Guo, Y.; Zhang, K.; Zhou, K.; Du, S. DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. *Remote Sens.* **2020**, *12*, 1088. [CrossRef]
44. Tao, C.; Lu, W.; Qi, J.; Wang, H. Spatial information considered network for scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 984–998. [CrossRef]
45. Wang, W.; Chen, W.; Qiu, Q.; Chen, L.; Wu, B.; Lin, B.; He, X.; Liu, W. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *arXiv* **2023**, arXiv:2303.06908.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Zheng, F.; Lin, S.; Zhou, W.; Huang, H. A Lightweight Dual-branch Swin Transformer for Remote Sensing Scene Classification. *Remote Sens.* **2023**, *15*, 2865. [CrossRef]
48. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
49. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
50. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
51. Chung, J.; Gulcehre, C.; Cho, K.H.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
52. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
53. Zhao, W.; Bo, Y.; Chen, J.; Tiede, D.; Blaschke, T.; Emery, W.J. Exploring semantic elements for urban scene recognition: Deep integration of high-resolution imagery and OpenStreetMap (OSM). *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 237–250. [CrossRef]
54. Selvaraju, R.; Cogswell, M.; Das, A.; Vedantam, A.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.