International Journal of Geo-Information

MDPI

Article

Urban Land Use Classification Model Fusing Multimodal Deep Features

Yougui Ren [1], Zhiwei Xie [2,*] and Shuaizhi Zhai [2]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China; 1810600@stu.neu.edu.cn
[2] School of Transportation and Geomatics Engineering, Shenyang Jianzhu University, Shenyang 110168, China; zsz2152@sina.com
* Correspondence: zwxrs@sjzu.edu.cn

Abstract: Urban land use classification plays a significant role in urban studies and provides key guidance for urban development. However, existing methods predominantly rely on either raster structure deep features through convolutional neural networks (CNNs) or topological structure deep features through graph neural networks (GNNs), making it challenging to comprehensively capture the rich semantic information in remote sensing images. To address this limitation, we propose a novel urban land use classification model by integrating both raster and topological structure deep features to enhance the accuracy and robustness of the classification model. First, we divide the urban area into block units based on road network data and further subdivide these units using the fractal network evolution algorithm (FNEA). Next, the K-nearest neighbors (KNN) graph construction method with adaptive fusion coefficients is employed to generate both global and local graphs of the blocks and sub-units. The spectral features and subgraph features are then constructed, and a graph convolutional network (GCN) is utilized to extract the node relational features from both the global and local graphs, forming the topological structure deep features while aggregating local features into global ones. Subsequently, VGG-16 (Visual Geometry Group 16) is used to extract the image convolutional features of the block units, obtaining the raster structure deep features. Finally, the transformer is used to fuse both topological and raster structure deep features, and land use classification is completed using the softmax function. Experiments were conducted using high-resolution Google images and Open Street Map (OSM) data, with study areas on the third ring road of Shenyang and the fourth ring road of Chengdu. The results demonstrate that the proposed method improves the overall accuracy and Kappa coefficient by 9.32% and 0.17, respectively, compared to single deep learning models. Incorporating subgraph structure features further enhances the overall accuracy and Kappa by 1.13% and 0.1. The adaptive KNN graph construction method achieves accuracy comparable to that of the empirical threshold method. This study enables accurate large-scale urban land use classification with reduced manual intervention, improving urban planning efficiency. The experimental results verify the effectiveness of the proposed method, particularly in terms of classification accuracy and feature representation completeness.

Keywords: land use classification; convolutional neural network (CNN); graph neural network (GNN); transformer

check for updates

Citation: Ren, Y.; Xie, Z.; Zhai, S. Urban Land Use Classification Model Fusing Multimodal Deep Features. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 378. https://doi.org/10.3390/ijgi13110378

Academic Editors: Wolfgang Kainz and Godwin Yeboah

Received: 11 August 2024
Revised: 12 October 2024
Accepted: 28 October 2024
Published: 30 October 2024

Copyright: © 2024 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Land use classification plays a pivotal role in numerous fields, including urban and regional planning, environmental sensitivity analysis, and hazard monitoring [1–3]. Key performance indicators (KPIs) based on land use classification data can significantly enhance the effectiveness of urban and regional planning. By classifying and monitoring different functional zones, urban planners can allocate land resources, construct infrastructure, and define functional zones more precisely, thereby promoting sustainable urban

development [4,5]. Analyzing land use categories and changes allows for better monitoring and management of urban environmental challenges, especially in regions undergoing rapid industrialization and urbanization. Land use classification can help identify high-pollution areas and implement environmental protection measures [6,7]. Additionally, land use classification aids in assessing the externalities of transportation and optimizing traffic management and logistics systems to reduce environmental and social impacts [8]. The relationships between land use types and human activities are inextricably linked. In urban areas, industrial activities are no longer dispersed but tend to be concentrated in specific areas. Furthermore, human activities are changing the terrestrial environment at an unprecedented rate, leading to the creation of specific functional zones for housing, offices, education, and research, as well as recreation and entertainment. These behaviors have triggered changes in land use classification and urban sprawl, becoming major drivers of global environmental change [9]. Urban sprawl presents multifaceted challenges to urban planning. It is typically associated with low-density development, leading to the inefficient use of urban space. Inefficient land use not only increases the costs of municipal infrastructure but also contributes to environmental degradation [10]. Due to the discontinuous nature of urban expansion, municipal infrastructure (such as water, electricity, public transportation, etc.) is often unable to effectively cover all areas [11]. This results in an over-provision of infrastructure in some regions while others remain underserved. Urban sprawl exacerbates socioeconomic stratification, with the wealthy and the poor residing in different urban areas. Typically, low-income groups live in areas far from city centers, with limited access to quality public services and employment opportunities, while the wealthy reside in areas with better infrastructure. This segregation intensifies social inequality and increases the daily commuting burden for low-income individuals [12]. Although many cities have developed master plans to control urban expansion, the effectiveness of these plans is often limited, and they have generally failed to curb urban sprawl [10]. Governments face challenges in coordinating land use and expansion control policies, particularly in terms of cross-departmental and cross-administrative planning coordination. Quantitative methods can be applied to assess land use types and effectively capture and guide regional planning. With rapid population growth and continuous economic and industrial progress, particularly in developing countries where changes in land use type have intensified, it is imperative to assess and analyze the human living environment, accounting for a multitude of factors [13–15]. The study of land use classification has become a significant focus within global land environment research. Land use classification not only enhances the theoretical framework of urban planning but also provides urban managers with a crucial decision-making basis, ultimately leading to a more rational and scientific urban spatial layout.

The field of land use classification focuses on analyzing the utilization and functional attributes of land. Studies in this area aim to analyze the economic attributes of society through land use. The use of remote sensing image data or social sensing data at the block scale can reveal the functional attributes of the parcels contained in a block analysis unit. This approach has become the mainstream method for studying urban land use classification, with the block serving as the fundamental research unit. For example, Gu [16] and Gao et al. [17] employed multisource social sensing data and identified urban functional area categories within their respective study areas through the use of kernel density clustering and Dirichlet distribution models. Yao [18], Zhai [19], and Hu et al. [20] employed a word embedding model in conjunction with point of interest (POI) data to facilitate the classification of urban land use. Furthermore, Pennington et al. [21] employed POI along with a global a priori statistical information model to regulate the relative weights of words, thereby achieving more accurate land use classification. The aforementioned methods, which utilize social sensing data, are insufficient for large-scale land use classification because of their limited coverage, lack of real-time and periodic data, and overly complex data processing processes.

Remote sensing images can be acquired at regular intervals, enabling the objective identification of large-scale ground object classification. This makes remote sensing a more suitable approach for large-scale urban land use classification. There are two common approaches for extracting deep features of block image objects: using convolutional neural networks (CNNs) with element-by-element convolution and using encoder embedding in graph neural networks (GNNs). The deep features obtained via these two methods are typically highly semantically expressive. Zhang et al. [22] constructed an urban topological network based on remote sensing images, extracted semantic features of street-scale image scenes, and introduced a method based on a two-layer spatial correlation graph neural network structure to achieve urban block scene classification. Chen et al. [23] proposed a semantically enhanced graph convolutional network (GCN) to facilitate the multiscale embedding of block units, thereby enabling the completion of land use classification for the study area. Liu et al. [24] introduced an attention mechanism and channel attention, which they used to construct a multiscale semantic segmentation model for blocks. This model aggregates multiscale features to complete the classification of block scenes in remote sensing images. Fu et al. [25] proposed an exemplar feature GCN approach for semi-supervised classification that improves the accuracy of block scene classification while preserving the local geometric feature embedding of the data. The aforementioned methods extract embedded features of remote sensing images through graph neural networks and their extensions, retaining the spatial and spectral information of the images. However, these methods often ignore the semantic information at the pixel level in the images and utilize only spatial or spectral similarity to construct a graph structure. This approach is insufficient for exploring the features of remote sensing images. Convolutional neural networks are capable of performing pixel-by-pixel deep feature extraction by adjusting the size, style, and number of convolutional kernels, as well as the parameters of pooling and fully connected layers [26]. Zhang et al. [27] employed ultrahigh-resolution images to enhance the representation of spatial and spectral features of image objects, utilizing a joint deep learning framework comprising a combination of multilayer perceptron and convolutional neural networks. Weng et al. [28] used high-resolution remote sensing images in conjunction with convolutional neural networks to enhance the expression of the deep features of image objects for land use classification. Yu et al. [29] used a combined convolutional neural network model to integrate the multisource semantic information of image objects effectively to complete the urban block scene classification of high-resolution image datasets. The above methods use convolutional neural networks and their extensions to introduce deep semantic information for block-scale image objects.

Single-model semantic features are often insufficient for fully expressing the characteristics of image objects. The combination of the GNN and CNN can often be used to capture multiple types of deep features, which can then be employed to enhance the expression of image objects. Liang et al. [30] used a dual-stream architecture that combines CNN global features and GCN object features to classify block scenes in two datasets. This approach improves the refinement of features and places the joint features in the fully connected layer of the CNN. Li et al. [31] employed a multilabel remote sensing image scene classification framework that integrated a CNN and a GNN to extract joint deep features of image objects and classify block scenes via graph attention networks (GATs). Although these methods employ GNNs and CNNs to extract deep features, they utilize a simplistic vector concatenating approach in the feature fusion process, which limits the ability to effectively consider the correlation between the deep features of multiple models, resulting in a need for further improvement in the feature fusion effect.

Urban land use information can help researchers and decision-makers better understand the processes, trends, and potential issues of urbanization, thereby formulating effective policies and planning measures [32]. However, manual statistics remain the primary method for obtaining land use information in current urban planning research [33]. This method is low in automation and highly subjective. Remote sensing technology provides an effective means for automatically acquiring large-scale urban land use infor-

mation, significantly improving the timeliness and objectivity of this data and enhancing the reliability of data support for urbanization-related studies [34]. Additionally, traditional land use classification methods based on remote sensing data often rely on manual features, while multimodal deep feature fusion can automatically learn spatial features, capturing both micro and macro patterns of urban structures. This approach facilitates the automated classification and extraction of complex urban forms and functional zoning, aiding in the identification of dynamic processes such as urban expansion, suburbanization, and industrial clustering [14]. Urban planning and land use classification are not static; dynamic changes in the urban development process, such as land reallocation and the transformation and expansion of functional zones, are crucial for understanding and predicting urban processes. Multimodal fusion can reduce human intervention and labor costs, enabling efficient automated land use classification that assists urban planners in making more intelligent decisions [35].

The objective of this study is to improve the accuracy of urban land use classification by effectively mining and utilizing multiple types of deep features from block-scale remote sensing images, thereby providing high-quality land use data support for urban planning and urbanization-related research. To achieve this, we propose a multimodal deep feature fusion land use classification model tailored for block-scale analysis. This model employs a KNN graph construction method with self-adjusting fusion coefficients, incorporating subgraph structural features into block-scale image objects to enhance the topological structural deep feature representation of GCN. Additionally, it combines the raster structural deep features from VGG-16 and integrates these two types of deep features using a transformer model, aiming to enhance land use classification accuracy at the block scale. This method effectively integrates multidimensional composite deep features across different scales, revealing complex urban spatial patterns and land use distribution characteristics. The main contributions of this paper are as follows:

- We incorporate the transformer model into the multimodal deep feature fusion process, facilitating the use of multimodal deep features in attending to the correlation information between different model deep features during fusion and thereby enhancing the efficacy of multimodal deep feature fusion.
- We develop a methodology for extracting structural features of internal subgraphs of block image objects. This method involves the multiscale segmentation of each block image object and the generation of a subgraph structure on the basis of the positional information of the newly created image object. The structural features of the subgraphs are subjected to analysis and extraction with the objective of enhancing the feature representation of the image objects, which in turn serves to augment the topological structure deep feature representation of the GCN.
- We propose an adaptive graph structure construction model based on composite similarity to enhance the expression of the edge weights in the graph structure by employing spatial and spectral composite similarity. The introduction of an adaptive adjustment parameter mechanism optimizes the process of seeking the optimal solution.

2. Main Methods

The transformer is employed to fuse GCN topological structure deep features and VGG-16 raster structure deep features, thereby enhancing the feature expression refinement and land use classification accuracy of block-scale image objects. The process is illustrated in Figure 1. First, the images are cropped in accordance with the fundamental units of the block, and the traditional features of the image objects at the block scale are identified. Second, the raster structure deep features of the segmented image are extracted via VGG-16. The traditional features of the image objects are subsequently combined with the subgraph features through vector concatenation, and the optimal graph structure is selected to extract the topological structure deep features via the GCN. The final stage of the land use classification process is the fusion of the two deep feature vectors through the transformer.

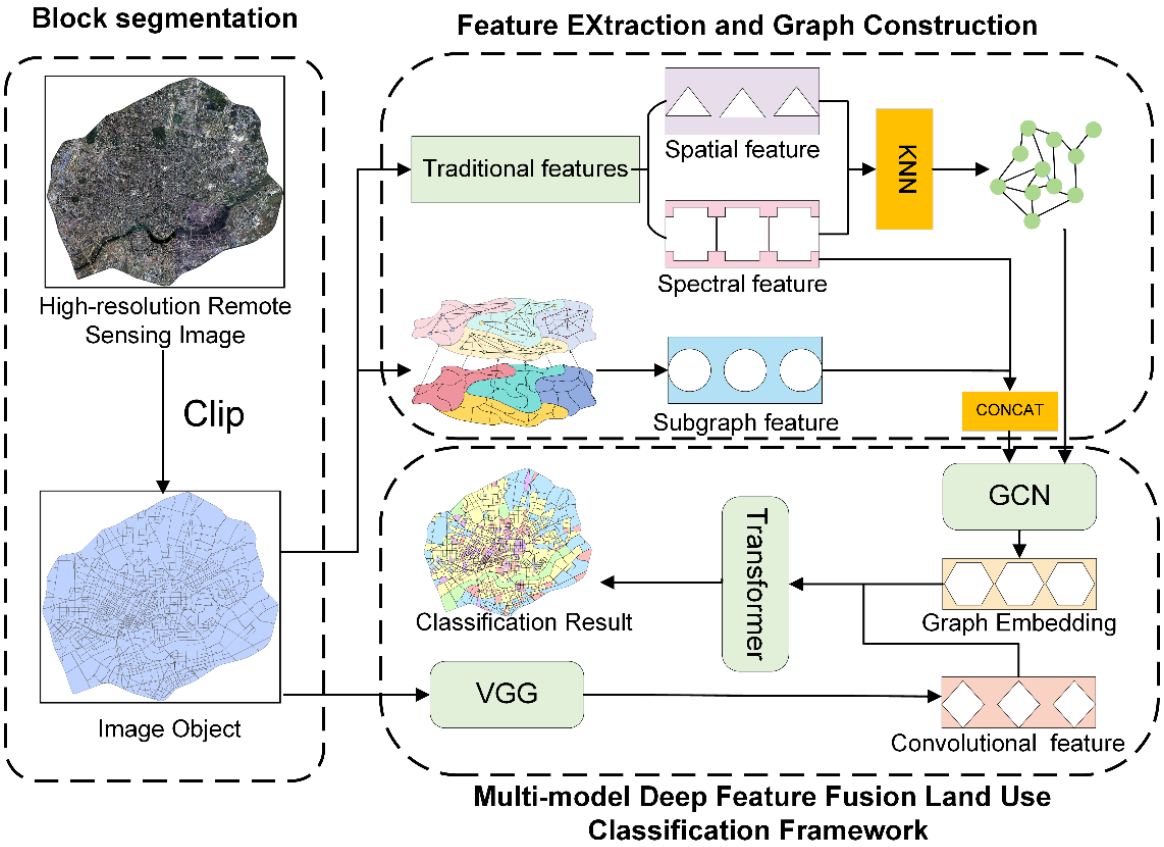


Figure 1. The methodological flow of the proposed approach.

2.1. Multiscale Land Use Analysis Unit Construction

The collection of street segmentation data is a crucial stage in urban land use studies. Commonly employed methods include collaboration with government departments to obtain data, utilization of high-resolution remote sensing images and nighttime images to identify roads, and the utilization of Open Street Map (OSM) open-source geographic data as the basis for segmentation. Previous studies have demonstrated the efficacy of utilizing OSM data to extract urban streets, and the accessibility of OSM data reduces the financial burden of the study while enhancing the generalizability of the method [36,37]. The incorporation of a limited degree of human intervention into OSM data can result in a more accurate representation of the actual urban spatial layout [38,39]. In light of the aforementioned considerations, 2021 road network data for Shenyang and Chengdu were obtained from the official OSM website, serving as the foundation for block delineation.

A selection of primary, secondary, and tertiary roads within the OSM road network was identified, and the topology was corrected. The endpoints of each road were subsequently connected to other roads through operations such as pruning and extension. Buffer zones were generated on the basis of different classes of roads; the road space was created and then removed to obtain the traffic analysis zone (TAZ). In accordance with the National Standard of the People's Republic of China, "Code for classification of urban land use and planning standards of development land GB50137" [40], we developed five land use categories: natural, public, residential, industrial, and commercial.

2.1.1. Traditional Feature Construction for Block Image Objects

The Open Map Service application programming interface (Google Earth API), provided by Google Earth, was employed to utilize Python scripts for batch downloading 2021 high-resolution remote sensing images in vector format. The geographic coordinates of the downloaded images were used to complete the stitching process. Each image was segmented via TAZ to obtain block-scale image objects, and the traditional feature space

of the image objects was constructed by extracting spectral features and texture features. Spectral features represent the fundamental features that define the spectral information of an image, including the spectral mean, standard deviation, and skewness. The analysis of image texture features is of paramount importance in the field of high-resolution image analysis. The gray-level cooccurrence matrix (GLCM) and gray-level difference vector (GLDV) are extensively employed in the domain of land cover classification of remote sensing images [39,41]. As illustrated in Table 1, a total of 13-dimensional features were employed, including the spectral mean, standard deviation, skewness, GLCM homogeneity, GLCM dissimilarity, and GLDV contrast in each band. These features constitute the traditional features of block image objects.

Table 1. Traditional feature system for block image objects.

Feature Type	Statistical Method
Spectral features	Red band mean Green band mean Blue band mean Brightness Red band standard deviation Green band standard deviation Blue band standard deviation Red band skewness Green band skewness Blue band skewness
Textural features	GLCM dissimilarity GLCM homogeneity GLDV contrast

The min–max normalization method was employed to standardize the traditional features of block image objects x_i, achieving a unified magnitude for all features [42]. The process of min–max normalization is also referred to as linear function normalization. The formula for this calculation is as follows:

$$x_i^{new} = \frac{x_i - x_i^{Min}}{x_i^{Max} - x_i^{Min}} \quad (1)$$

where x_i^{new} is the normalized traditional features of block image object, x_i is the traditional features of the original block image object, x_i^{Min} and x_i^{Max} denote the maximum and minimum values of the feature values of class i.

2.1.2. Subgraph Structural Feature Construction for a Block Image Object

Traditional methods primarily focus on extracting grayscale features from remote sensing images of surface-covered objects. However, these methods are often inadequate for analyzing spatial features, such as topological and distance features, which arise because of the spatial autocorrelation of surface-covered objects. Based on the original image objects, multi-scale image segmentation is performed internally using the fractal network evolution approach (FNEA) [43]. FNEA divides the image into objects based on the principle of minimizing heterogeneity, utilizing both spectral and shape heterogeneity. The segmentation scale, compactness, and shape factor are critical parameters influencing the segmentation results. Since the effectiveness of the segmentation depends on the spatial resolution of the image and the size of the features, there is currently no effective method for automatically determining the parameters. Therefore, based on expert experience, the segmentation scale is set to 30, while the compactness and shape factor are set to the algorithm-recommended values of 0.5 and 0.1, respectively. The subgraph structure is constructed from the secondary image objects obtained through segmentation, and subgraph metrics are analyzed and extracted to supplement the feature space of the original

image objects. The method for constructing the subgraph structure is illustrated in Figure 2.

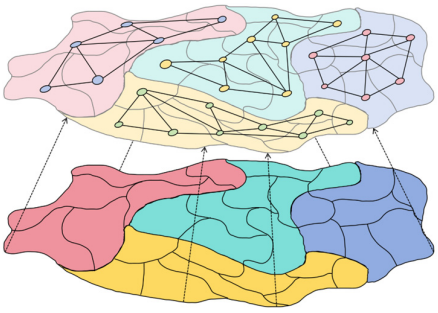


Figure 2. Schematic diagram of subgraph structure construction.

The subgraph structure is constructed on the basis of whether there are shared edges between secondary image objects. Subgraph features include distance features, spectral features, and subgraph indicator features. Two distinct types of distance features can be identified: the number of newly generated secondary image objects for each original image object and the average distance from the center of mass of the newly generated secondary image objects to the center of mass of the original image object. The latter measure represents the degree of dispersion. The spectral features comprise the calculated variance and standard deviation between the gray values of the bands within the secondary image object and the original image object. The subgraph indicator feature set comprises the number of nodes in the subgraph structure, high degree of clustering, average degree, average clustering coefficient, average path length, graph diameter, and small-world effect.

The degree of a node is defined as the number of edges directly connected to it. The degree of node i is denoted as k_i. The average degree is defined as the arithmetic mean of the degrees of all nodes in the graph. The average degree of the graph G can be calculated via the adjacency matrix $A = (a_{ij})_{N\times N}$ and the number of edges M. The formula is as follows:

$$\overline{k} = \frac{1}{N}\sum_{i,j=1}^{N} a_{ij} \tag{2}$$

where N represents the total number of nodes in the graph. The path with the least number of edges between two nodes in the graph is called the shortest path. The distance d_{ij} is defined as the number of edges of the shortest path of nodes i and j. The average path length $\overline{S}$ is the average of the distances of any two nodes in the graph with the following formula:

$$\overline{S} = \frac{1}{\frac{1}{2}N(N-1)}\sum_{i>=j} d_{ij} \tag{3}$$

The diameter D of the graph, defined as the maximum value of the distance between any two nodes in the graph, is given by the following formula:

$$D = \max(d_{ij}) \tag{4}$$

The clustering coefficient indicates the probability that any pair of nodes in the block of a given node has a connected edge, and the average clustering coefficient is the average of the clustering coefficients of all the nodes in the graph, with the following formula:

$$\overline{C} = \frac{1}{N}\sum_{i=1}^{N} \frac{2E_i}{k_i(k_i-1)} \tag{5}$$

where E_i is the number of edges that actually exist between the neighboring nodes of the point and $k_i(k_i-1)/2$ represents the maximum number of edges possible for these neighboring nodes. To quantify the small-world effect of a graph, the original definition of

a small-world graph can be used to compare the average clustering coefficient $\overline{C}$ and path length L of the given graph with those of an equivalent random graph. The small-world effect metric ω is formulated as follows:

$$\omega = \frac{L_r}{L} - \frac{\overline{C}}{\overline{C}_l} \tag{6}$$

where $\overline{C}_l$ is the average clustering coefficient of the equivalent raster graph and L_r is the feature path length of the equivalent random graph. In this work, the 16-dimensional features presented in Table 2 are employed as subgraph structural features of block image objects.

Table 2. Subgraph structural feature system for block image objects.

Subgraph Structural Feature Type	Statistical Method
Spectral features	Standard deviation of red band mean Variance of the red band mean Standard deviation of green band mean Variance of the green band mean Standard deviation of blue band mean Variance of the blue band mean Standard deviation of brightness Variance of brightness
Subgraph indicator features	High degree of clustering Average degree Average clustering coefficient Average path length Graph diameter Small-world effect
Distance features	Nodes number Dispersion

2.2. Self-Adjusting Fusion Coefficients for the KNN Graph Construction Method

The node similarity S employed by traditional KNN graph construction methods is typically the spatial Euclidean distance or spectral similarity between nodes. To address the issue of insufficient expressiveness of single similarity measures for hyperspectral images, Yang et al. developed a node similarity calculation method that integrates spectral and spatial distance features within the traditional KNN graph construction framework [44]. This approach was subsequently employed for hyperspectral image classification. The formula for the similarity of this node is as follows:

$$s_{ij} = \omega \times spe_{ij} + (1 - \omega) \times dis_{ij} \tag{7}$$

where s_{ij} is the similarity of nodes i and j, ω is the fusion coefficient of spectral features, and $(1 - \omega)$ is the fusion coefficient of spatial distance features. The optimal values of the closest number K and the fusion coefficient ω need to be determined through multiple experiments. The fusion coefficient ω relies on manual adjustment, which is highly subjective and affects the algorithm's stability.

To address the limitations of relying on manually adjusted fusion coefficients, we developed a compositional method for adaptive fusion coefficient adjustment. The spectral feature similarity Spe and spatial distance similarity Dis are computed by utilizing the center of mass coordinates of the image objects in each block as spatial features and the conventional feature vectors in each block as spectral features, respectively. The probabilistic prediction vectors $P_{Spe} = [P_1, P_2, P_3, P_4, P_5]$ and $P_{Dis} = [P_1, P_2, P_3, P_4, P_5]$ of both are obtained after they are output through the fully connected layer of the GCN. The probabilistic predicted values of the five types of block scenes, designated P_1, P_2, P_3, P_4 and

P_5, are presented. The spectral feature similarity Spe and the spatial distance similarity Dis are calculated as follows:

$$Dis = \sqrt{(d_{i1} - d_{j1})^2 + (d_{j2} - d_{j2})^2} \tag{8}$$

$$Spe = \sqrt{(r_{i1} - r_{j1})^2 + \cdots + (r_{iM} - r_{jM})^2} \tag{9}$$

where $D_i = \{d_{i1}, d_{i2}\}^T$ is the set of spatial coordinates corresponding to node v_i, d_{i1} is the longitude of node v_i, d_{i2} is the latitude of node v_i, $R_i = \{r_{i1}, \cdots, r_{iM}\}^T$ is the set of spectral features corresponding to node v_i, and M is the number of spectral features.

The formula for the probabilistic prediction vector P, derived from a mixture of the two similarities Spe and Dis, is as follows:

$$P = \omega \times P_{Spe} + (1 - \omega) \times P_{Dis} \tag{10}$$

where P represents the probabilistic prediction vectors of the five types of block scenes to be classified. The probabilistic prediction vectors P_1, P_2, P_3, P_4 and P_5 contained in the probabilistic prediction vectors P are converged to 1 by controlling the spectral fusion coefficients ω to be varied automatically in equal intervals. The optimal fusion coefficients ω are determined quickly by comparing the probabilistic prediction values computed by different coefficients and selecting the most accurate coefficient P.

2.3. GCN-Based Deep Feature Extraction of Topological Structures

A GCN is a type of graph neural network that extracts the embedding vectors of nodes by performing convolution operations on the graph structure and generates a new representation of the node embedding vectors through the interactions between these embedding vectors [45]. These embedding vectors are able to encode the local structural information of the nodes, which in turn can be used for downstream tasks such as node classification and graph classification.

GCNs can be applied to a wide range of graph-structured data because of their high scalability and flexibility, enabling them to adapt to different graph sizes and complexities. Furthermore, GCNs can be integrated with other neural networks and machine learning models to address more complex problems. The GCN learns the embedding vectors of nodes by exploiting the structural properties of the graph and the interactions between the nodes. These vectors capture the roles and relationships of the nodes in the graph, providing a rich representation of the features for subsequent data analysis and comprehension. The embedding formula of the GCN is represented as follows:

$$h'(v) = \sigma\left(\sum_{u \in N(v)} W \cdot h(u)\right) \tag{11}$$

where $h'(v)$ denotes the embedding vector of the updated node v, $N(v)$ is the set of neighboring nodes of node v, W is a learnable weight matrix used to adjust the importance of different features and generate new feature representations, σ is an activation function, usually a rectified linear unit (ReLU) or sigmoid function, which introduces nonlinearity, allowing the network to capture more complex feature interactions, and $h(u)$ is a feature vector reflecting the initial state of node u in the graph.

The block graph structure, constructed via the KNN composition method with self-adjusting fusion coefficients, the conventional feature vectors of each block, and the sub-graph structure feature vectors, are input into the GCN for the convolution operation, enabling the extraction of a deep feature representation of the topology of each block.

2.4. VGG-16-Based Deep Feature Extraction of Raster Structures

The images were cropped by TAZ to obtain block-scale image blocks, and we employed VGG-16 to convolve each image block and extract the embedding vectors. VGG is a deep convolutional neural network developed by the Visual Geometry Group at the University of Oxford. This network is employed primarily for image recognition tasks [46]. As illustrated in Figure 3, the VGG-16 model is typically employed as a feature extractor to transform an image into a fixed-length vector representation that captures the most salient visual characteristics of the image.

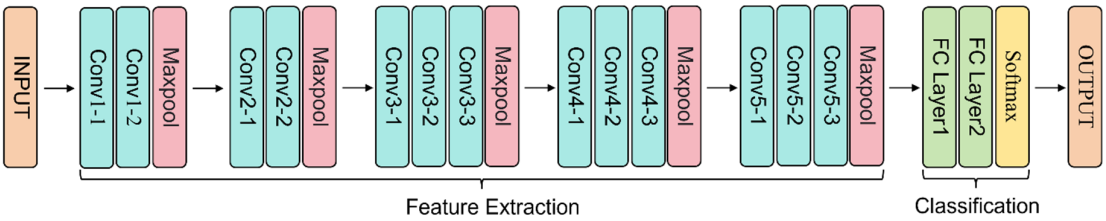


Figure 3. VGG-16 structure.

In this instance, the initial convolutional layer comprises 64 channels, with the number of channels increasing by a factor of three following each maximum pooling layer until reaching 512 channels. Finally, two fully connected layers, each comprising 4096 nodes, are connected, after which classification is performed by a classifier. This approach introduces the concept of grouping multiple convolutional layers with smaller convolutional kernels in lieu of a single convolutional layer with larger convolutional kernels, thereby expanding the size of the receptive field.

In the forward propagation phase of VGG-16, the pooling layers, particularly the maximum pooling layer, reduce the spatial size of the deep feature maps, thereby reducing the dimensionality of the data. This dimensionality reduction operation serves to reduce the amount of computation required while simultaneously facilitating the extraction of more representative features. However, this process inevitably entails the loss of some spatial information. In the context of feature extraction, the output of one of the fully connected layers or the final pooling layer of VGG-16 is typically employed as the embedding vector. In practice, the parameters of the VGG-16 model are typically pretrained and employed to extract deep features from new images. The TAZ-cropped image was input into VGG-16 for the convolution operation, and the vector output from the final fully connected layer was employed as the deep feature of the raster structure of the block unit.

2.5. Multimodal Deep Feature Transformer Fusion and Classification

The transformer was initially proposed by Vaswani et al. and subsequently applied to the field of machine translation. The aim of the transformer is to utilize encoder and decoder architectures based on the self-attention mechanism to achieve sequence-to-sequence transformation [47]. The fundamental elements of the transformer include multihead self-attention (MSA), MLP, and layer normalization (LN).

In the multihead self-attention module, the input $X \in R^{n \times d}$ is transformed into three parts, namely, the query $Q \in R^{n \times d_k}$, the key $K \in R^{n \times d_k}$, and the value $V \in R^{n \times d_v}$, where n is the sequence length, and d, d_k and d_v are the dimensions of the input, query (key) and value, respectively. When scale dot product notation is applied to Q, K and V, the formula is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

Finally, a linear layer is employed to generate the output. Multihead self-attention divides the query, keys, and values into h parts, performs the attention function in parallel, and then concatenates and linearly projects the output values of each attention header to

form the final output. The MLP is applied between the self-attention layers to perform feature transformations and nonlinear processing with the following formulas:

$$MLP(X) = FC(\sigma(FC(X))), \quad FC(X) = XW + b \tag{13}$$

where W and b are the weights and skewness of the fully connected layer, respectively, and where $\sigma(\cdot)$ is an activation function such as ReLU. Let $C = \{c_i\}_i$ be the set of feature vectors. The scalar dot product attention layer formula is expressed as follows:

$$y_i = \sum_{c_j \in C} \rho\Big(\varphi(c_i)^\top \psi(c_j) + \zeta\Big)\alpha(c_j) \tag{14}$$

where y_i is the output feature, φ, ψ and α are pointwise feature transformations such as linear projection or MLP, ζ is a positional coding function, and ρ is a normalization function such as softmax. The scalar attention layer computes the scalar product between the φ and ψ transformed features and uses its output as an attention weight on the aggregation of the α transformed features.

The GCN topological structure deep features and VGG-16 raster structure deep features of the block units are input into the transformer in turn. The potential relationships between the deep features of the two models are considered through the transformer's self-attention mechanism, and the feature fusion operation is completed through the transformer's encoder. Finally, the land use classification is completed through the use of the softmax function.

3. Results

The experiments described in this paper were conducted using a Windows 11 computer operating system(Microsoft Corporation, Redmond, WA, USA) and an AMD Ryzen 7 5800H with a Radeon Graphics CPU @ 3.20 GHz processor running on 16 GB of RAM (Advanced Micro Devices, Inc., Santa Clara, CA, USA). Additionally, an NVIDIA GeForce RTX 3060 laptop GPU, with 6 GB of dedicated GPU memory (NVIDIA Corporation, Santa Clara, CA, USA), was employed for graphics processing. The method in this paper was implemented via TensorFlow 2.15.0 and PyTorch 2.0. A total of 30% of the dataset was selected as the training set, with the remaining dataset serving as the test set.

3.1. Results of the Construction of Multiscale Land Use Analysis Units

We focus on the areas within the third ring road of Shenyang, Liaoning Province, and the fourth ring road of Chengdu, Sichuan Province, in China. Urban ring roads are generally important demarcations in a city's spatial structure. The third and fourth ring roads serve as significant transportation and spatial boundaries, encompassing different functional zones from the city core to peripheral areas, delineating the spatial land use distribution across urban and suburban regions. The ring road structures of Shenyang and Chengdu are relatively clear, making them natural spatial boundaries that facilitate comparisons between the functional characteristics of core, transitional, and suburban areas. Both Shenyang and Chengdu are major central cities in China, with large-scale urban areas that exhibit typical characteristics of medium to large cities. Within Shenyang's third ring road, various land use types are present, including commercial, industrial, and residential zones. In Chengdu, the area within the fourth ring road encompasses a wide range of functional zones, from old city districts, commercial areas, and technology parks to cultural and creative zones and residential areas. Chengdu has undergone rapid land use transformation in recent years, particularly with the development of high-tech industries and urban renewal, providing a unique perspective for studying changes in urban land use.

Preprocessed OSM data were employed to delineate distinct buffer zones classified according to road type. Road spaces within all buffer sections were removed to obtain the block analysis units (TAZs). Land use labels were assigned to each TAZ through visual interpretation to create reference data for block land use classification of the third ring roads in Shenyang and the fourth ring roads in Chengdu. The results and reference data for

the third ring road TAZ construction in Shenyang can be observed in Figure 4a,c, whereas the results and reference data for the fourth ring road TAZ construction in Chengdu can be observed in Figure 4b,d.

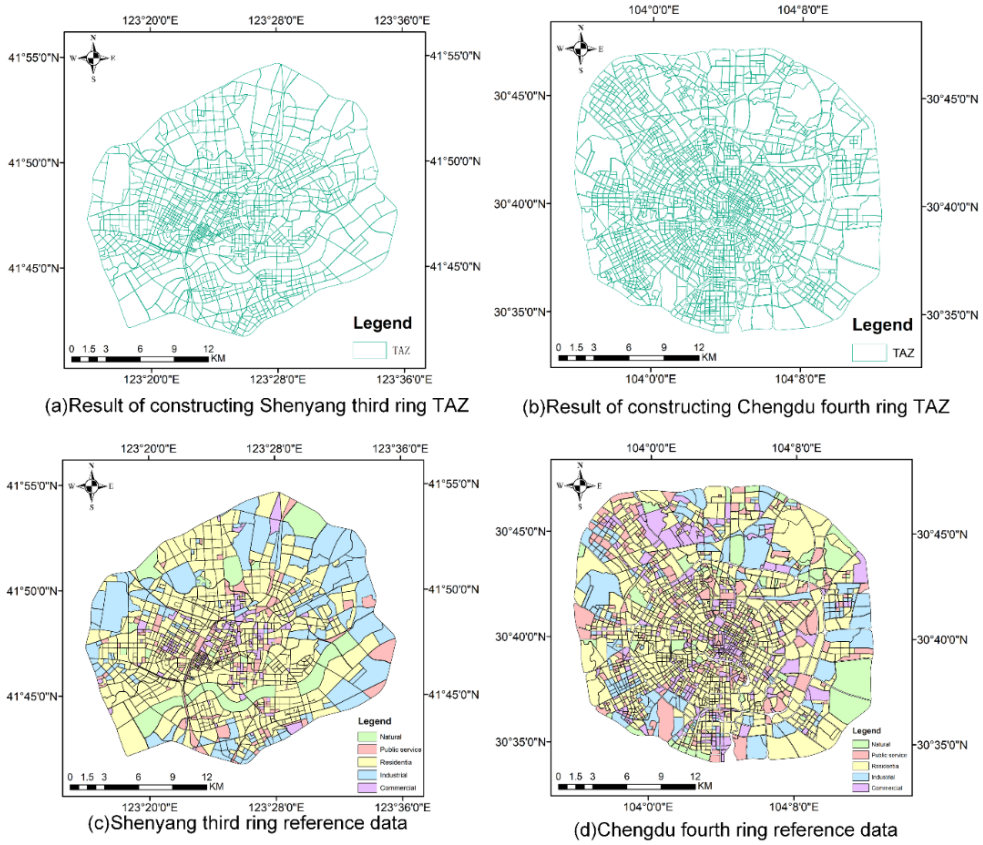


Figure 4. Data preprocessing.

The third ring of Shenyang and the fourth ring of Chengdu comprise 1146 and 1857 TAZs, respectively. The reference data for these areas include 31 and 60 natural categories, 218 and 332 public categories, 546 and 900 residential categories, 149 and 172 industrial categories, and 202 and 303 commercial categories, respectively.

3.2. Image Segmentation and Feature Extraction Results

The two segmentations lead to 47,689 and 31,792 image objects from Shenyang and Chengdu, respectively. The image objects resulting from the second segmentation were employed as nodes, and subsequently, the subgraph structure was constructed on the basis of the presence or absence of a common edge between the newly generated secondary image objects within each block-scale image object. The constructed subgraph structure is illustrated in Figure 5.

In a separate analysis, features, including both traditional features and subgraph structural features, were extracted for block-scale image objects. The traditional features were divided into spectral and textural features. The subgraph structural features include distance features, spectral features, and subgraph indicator features. The traditional feature dimensions of the experimental data of the third ring of Shenyang and the fourth ring of Chengdu were both 13 dimensions, subgraph structural features were added to construct the feature set of the optimized image objects, and the feature dimensions were both 29 dimensions.

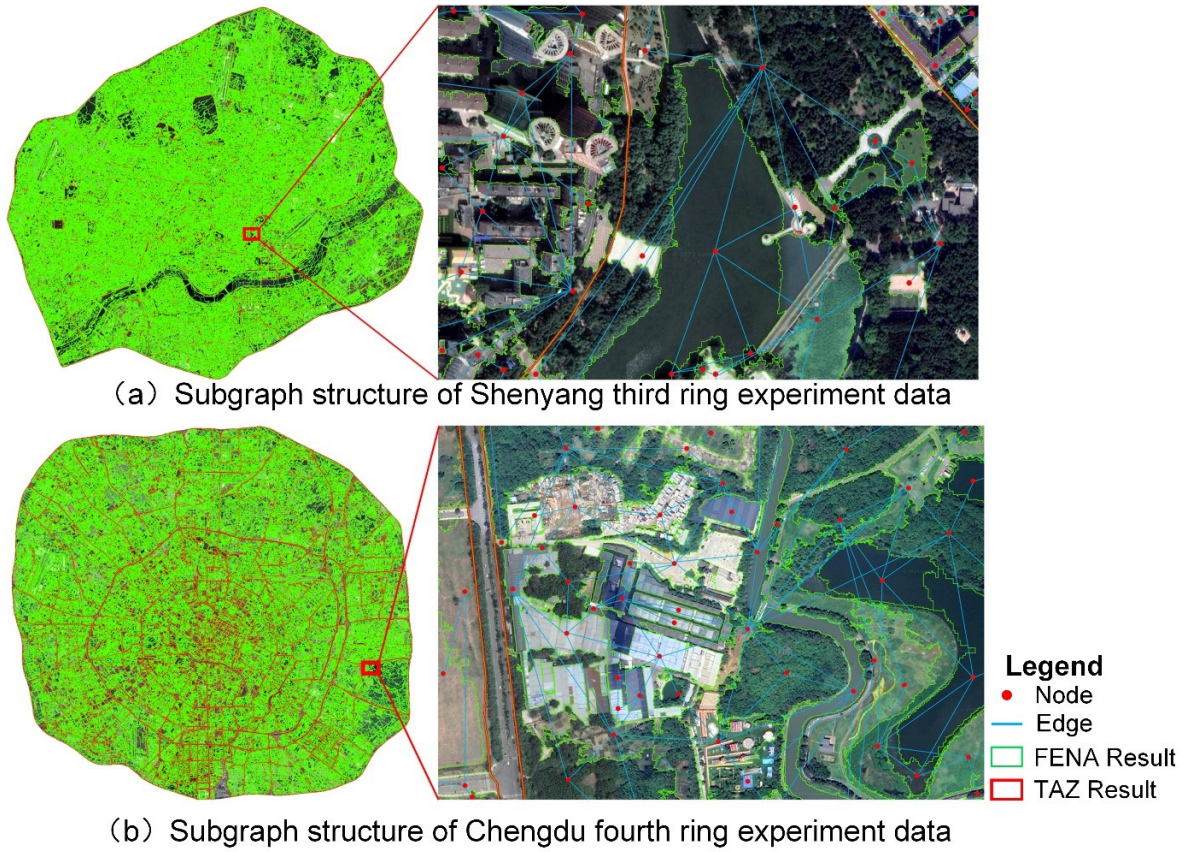


Figure 5. Image segmentation and subgraph construction results.

3.3. Transformer Fusion Embedding Vectors for Land Use Classification Results

In this part of the experiment, we compared the block-scale oriented multimodal deep feature fusion land use classification method proposed in this paper with a classification method using a single model. When the embedding feature vectors of image objects were extracted via the GCN, the feature space of the image objects was enriched by improving the image graph structure and adding subgraph structure features, resulting in a more refined expression of the embedding feature vectors. The land use classification results obtained through the fusion of optimized GCN topological structure deep features and VGG-16 raster structure deep features via the transformer are presented in Figures 6 and 7. For comparison, the classification results of the unoptimized GCN and VGG-16 are shown in Table 3.

Table 3. Accuracy evaluation before and after optimization in Shenyang and Chengdu.

Data	Accuracy Indicators	Different Methods		
		Transformer	GCN	VGG-16
Shenyang Third Ring	Natural Producer's Accuracy	80.65%	16.13%	74.19%
	Natural User's Accuracy	92.59%	33.33%	85.19%
	Public Producer's Accuracy	61.01%	0.92%	23.85%
	Public User's Accuracy	73.48%	15.38%	61.90%
	Residential Producer's Accuracy	92.31%	86.26%	81.32%
	Residential User's Accuracy	75.45%	52.10%	71.50%
	Industrial Producer's Accuracy	71.81%	58.39%	83.22%
	Industrial User's Accuracy	84.25%	61.70%	61.69%
	Commercial Producer's Accuracy	53.47%	15.84%	50.00%
	Commercial User's Accuracy	75.52%	43.84%	47.42%
	Overall Accuracy	76.53%	52.09%	64.92%
	Kappa	0.6417	0.1941	0.4777

Table 3. *Cont.*

Data	Accuracy Indicators	Different Methods		
		Transformer	GCN	VGG-16
Chengdu Fourth Ring	Natural Producer's Accuracy	53.33%	0.00%	50.00%
	Natural User's Accuracy	82.05%	0.00%	65.22%
	Public Producer's Accuracy	59.34%	2.11%	36.14%
	Public User's Accuracy	61.56%	24.14%	42.40%
	Residential Producer's Accuracy	89.80%	94.95%	88.79%
	Residential User's Accuracy	79.16%	55.85%	69.93%
	Industrial Producer's Accuracy	58.72%	19.19%	37.79%
	Industrial User's Accuracy	60.12%	38.37%	56.03%
	Commercial Producer's Accuracy	46.86%	8.25%	31.02%
	Commercial User's Accuracy	68.60%	42.37%	60.65%
	Overall Accuracy	73.29%	54.12%	63.97%
	Kappa	0.5689	0.0907	0.3911

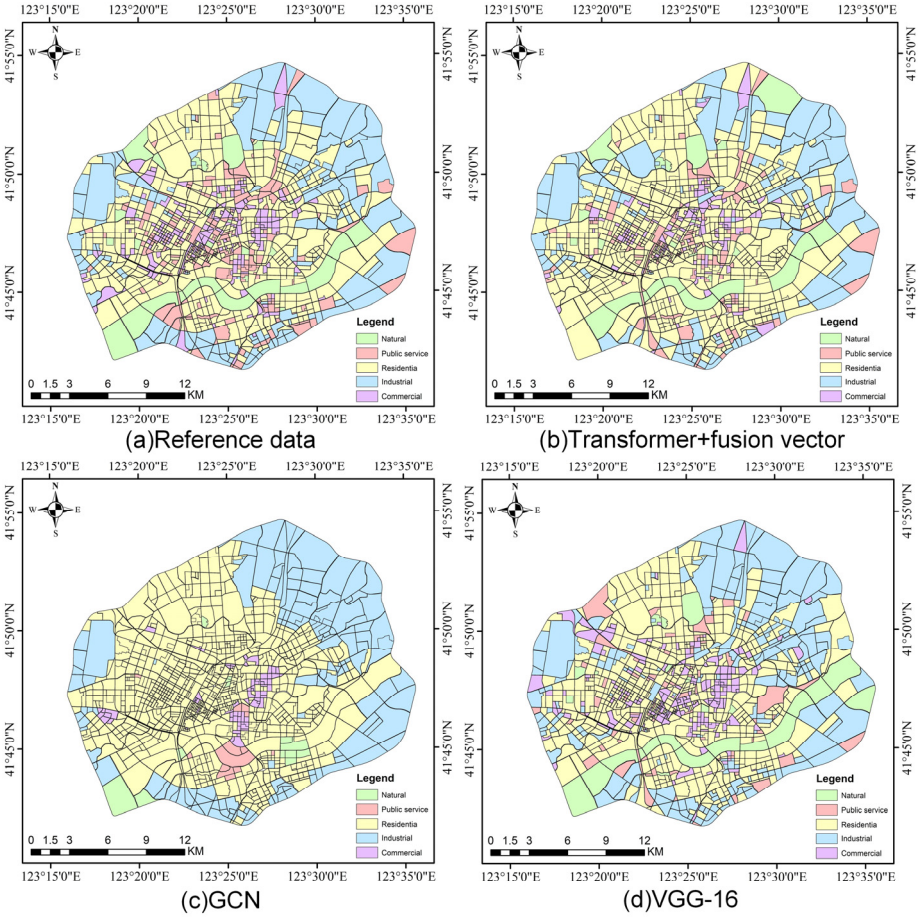


Figure 6. Comparison of the classification results of different methods for the Shenyang third ring dataset.

The land use classification results achieved by fusing optimized GCN and VGG-16 embedding feature vectors via the transformer model are shown in Figures 6 and 7. Table 3 shows the superior performance of the transformer fusion embedding vector approach compared with that via unoptimized GCN and VGG-16 classification approaches. For the natural category in the Shenyang third ring road dataset, the producer's accuracy increased from 16.13% and 74.19% for the GCN and VGG-16, respectively, to 80.65% for the proposed method, and the user's accuracy improved from 33.33% and 85.19% to 92.59%. For the public category, the producer's accuracy increased from 0.92% and 23.85% to

61.01%, and the user's accuracy considerably increased from 15.38% and 61.90% to 73.48%. For the residential category, the producer's accuracy improved from 86.26% and 81.32% to 92.31%, and the user's accuracy improved from 52.10% and 71.50% to 75.45%. For the industrial category, the user's accuracy notably improved, increasing from 61.70% and 61.69% to 84.25%. Conversely, the producer's accuracy decreased from 83.22% to 71.81% in comparison with that of the VGG-16 method. However, the producer's accuracy showed an encouraging enhancement, increasing from 58.39% to 71.81% in comparison with the unoptimized GCN approach. For the commercial category, the producer's accuracy increased from 15.84% and 50.00% to 53.47%, and the user's accuracy increased from 43.84% and 47.42% to 75.52%. The overall accuracy on the Shenyang dataset increased from 52.09% and 64.92% to 76.53%, and the kappa coefficient increased from 0.1941 and 0.4777 to 0.6417. On the Chengdu fourth ring dataset, the transformer fusion embedding vector method improved the producer's accuracy for the natural category by 53.33% and 3.33% in comparison with that of the unoptimized GCN and VGG-16 methods, respectively. The producer's accuracy for the public category increased by 57.23% and 23.2%, respectively. However, for the residential category, the producer's accuracy was reduced from 94.95% to 89.8% in comparison to that of the unoptimized GCN method but improved from 88.79% to 89.8% in comparison with that of the VGG-16 method. The improvements in the user's accuracy were 23.31% and 9.23%, respectively. For the industrial category, the producer's accuracy increased by 39.53% and 20.93%, respectively, while for the commercial category, the producer's accuracy increased by 38.61% and 15.84%, respectively. The overall accuracy on the Chengdu dataset increased from 54.12% and 63.97% to 73.29%, whereas the kappa coefficient improved from 0.0907 and 0.3911 to 0.5689, respectively.

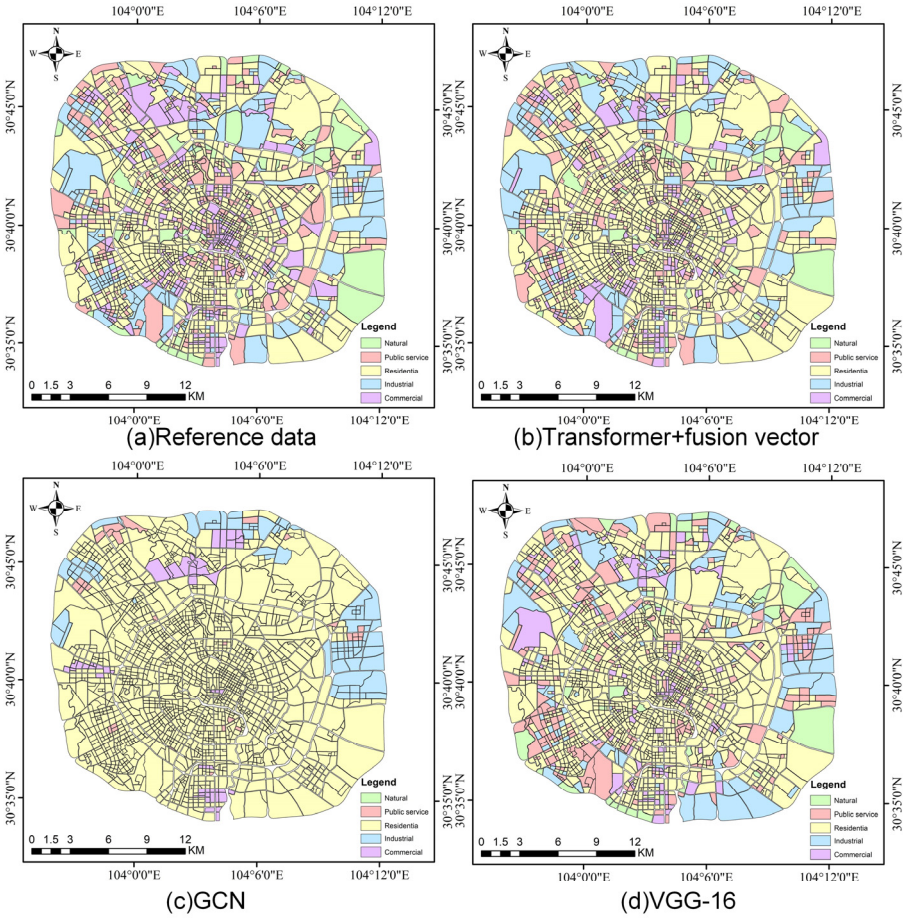


Figure 7. Comparison of the classification results of the different methods for the Chengdu fourth ring dataset.

3.4. Results of Self-Adjusting Fusion Coefficients for KNN Graph Construction

In this phase of the experiment, the KNN graph construction method with self-adjusting spectral coefficients was employed to construct the image graph structure. The spectral parameter interval ω was set to 0.1, while the closest proximity indices K were set to 5, 10, 15, and 20. This approach allowed for the automatic identification of the optimal graph structure. The hyperparameters of the GCN were set uniformly for all the experiments. The accuracies of the results obtained with different ω and K values for the third and fourth rings of Shenyang and Chengdu are presented in Table 4.

Table 4. Accuracy rating table for different ω and K.

Indicator		Shenyang Third Ring				Chengdu Fourth Ring			
		K = 5	K = 10	K = 15	K = 20	K = 5	K = 10	K = 15	K = 20
Overall Accuracy	$\omega = 0$	52.09%	51.75%	51.40%	50.17%	54.12%	53.80%	53.58%	53.53%
	$\omega = 0.1$	57.94%	56.20%	55.85%	54.62%	57.67%	56.27%	56.06%	56.22%
	$\omega = 0.2$	58.99%	56.46%	55.67%	55.67%	57.62%	55.95%	56.27%	54.66%
	$\omega = 0.3$	60.82%	56.02%	57.16%	57.42%	57.67%	56.44%	56.22%	56.00%
	$\omega = 0.4$	60.47%	57.94%	58.81%	58.12%	57.30%	56.00%	55.79%	55.41%
	$\omega = 0.5$	59.69%	58.03%	57.77%	57.24%	56.97%	55.47%	55.74%	55.79%
	$\omega = 0.6$	59.42%	56.89%	56.89%	57.24%	57.08%	56.38%	56.33%	55.79%
	$\omega = 0.7$	60.21%	57.07%	57.33%	55.85%	56.44%	56.11%	55.95%	55.95%
	$\omega = 0.8$	57.94%	58.64%	56.54%	56.89%	57.08%	56.38%	55.74%	54.60%
	$\omega = 0.9$	60.47%	57.94%	56.54%	55.76%	56.92%	56.44%	56.17%	55.74%

The analysis of the data in Table 4 reveals that, regardless of whether the data are from the Shenyang third ring or the Chengdu fourth ring, the overall accuracy tends to decrease with increasing K. Consequently, $K = 5$ represents the index of the most proximate optimal graph construction. When $\omega = 1$, that is, when only spectral features were employed for framing, the classification was more robust than that when only distance features were used for framing ($\omega = 0$). Furthermore, all the aforementioned laws remain valid when K varies. The utilization of spectral and distance features in conjunction with the construction of a graph structure typically yielded superior outcomes in land use classification compared with those when utilizing distance or spectral features alone. In conclusion, the optimal graph construction parameters for the experimental data of the Shenyang third ring and the Chengdu fourth ring are $\omega = 0.3$ and $K = 5$, respectively. Furthermore, the land use classification effect of using this composition through the GCN was superior to that of using the distance feature or spectral feature graph construction alone.

3.5. GCN Land Use Classification Results Using Subgraph Structural Features

In addition to the impact of the graph structure on the outcome of the graph embedding extracted by the GCN, the initial feature expression of the image object is a significant factor influencing the efficacy of land use classification in the block of the GCN. The two sets of experimental data were subjected to further segmentation (both segmentation scales were 500), resulting in the generation of 47,689 and 31,792 image objects, respectively. This additional segmentation led to an enhancement of the feature space from its original 13 dimensions to 29 dimensions. To determine whether the optimized graph structure and subgraph features can improve the accuracy of block land use classification in the GCN and whether the combination of both can yield further improvements, we conducted a comparative experiment.

Table 5 reveals that the addition of subgraph features to the Shenyang data had no effect on the producer's accuracy for the natural category. However, there was a notable enhancement in the user's accuracy. The producer's accuracy for the public category increased by 14.68%, whereas the user's accuracy improved by 31.84%. The producer accuracies for the residential and industrial categories decreased by 12.63% and 2.01%, respectively. While there was a slight increase in the user's accuracy for the residential

category, a decrease was observed for the industrial category. The most notable impact was observed in the commercial category, where the producer's accuracy increased by 29.7%. When combining the optimized graph structure with the subgraph features, the producer's accuracy for the public category improved by 27.99% compared to that using the optimized graph structure alone. Furthermore, the inclusion of subgraph features led to an additional 25.23% improvement in the producer's accuracy compared with that using subgraph features alone. Compared with the results obtained when the diagram structure was optimized as a standalone measure, the producer's accuracy for the nature category increased by 3.23%, and the user's accuracy increased by 6.92%. The producer's accuracy for both residential and industrial applications improved when subgraph features were incorporated, although a decrease was observed when the graph structure was optimized. However, the user's accuracy increased in both cases. Conversely, for the commercial category, there was a decrease in the producer's accuracy, but the user's accuracy was consistent with the aforementioned findings. Overall, the accuracy was improved by 1.75% and 8.73% when the graph structure was optimized and the subgraph features were incorporated, respectively. Additionally, the kappa values improved by 0.0527 and 0.1507, respectively.

Table 5. Evaluation of feature comparative experiment accuracy.

Data	Accuracy Indicators	GCN			
		ω = 0.3 + Subgraph Feature	ω = 0 + Subgraph Feature	ω = 0.3	ω = 0
Shenyang Third Ring	Natural Producer's Accuracy	64.52%	16.13%	61.29%	16.13%
	Natural User's Accuracy	80.00%	45.45%	73.08%	33.33%
	Public Producer's Accuracy	40.83%	15.60%	12.84%	0.92%
	Public User's Accuracy	41.20%	47.22%	44.44%	15.38%
	Residential Producer's Accuracy	80.40%	73.63%	86.63%	86.26%
	Residential User's Accuracy	71.97%	58.86%	64.62%	52.10%
	Industrial Producer's Accuracy	69.80%	56.38%	71.14%	58.39%
	Industrial User's Accuracy	68.87%	58.74%	63.47%	61.70%
	Commercial Producer's Accuracy	32.18%	45.54%	35.15%	15.84%
	Commercial User's Accuracy	45.14%	38.82%	44.94%	43.84%
	Overall Accuracy	62.57%	53.84%	60.82%	52.09%
	Kappa	0.4418	0.2911	0.3891	0.1941
Chengdu Fourth Ring	Natural Producer's Accuracy	40.00%	8.33%	43.33%	0.00%
	Natural User's Accuracy	55.81%	41.67%	48.15%	0.00%
	Public Producer's Accuracy	25.90%	12.95%	9.64%	2.11%
	Public User's Accuracy	30.71%	30.50%	27.35%	24.14%
	Residential Producer's Accuracy	87.68%	88.28%	92.63%	94.95%
	Residential User's Accuracy	66.31%	60.15%	62.55%	55.85%
	Industrial Producer's Accuracy	18.60%	36.05%	18.60%	19.19%
	Industrial User's Accuracy	34.78%	39.24%	38.55%	38.37%
	Commercial Producer's Accuracy	19.80%	13.86%	21.12%	8.25%
	Commercial User's Accuracy	45.11%	45.16%	46.72%	42.37%
	Overall Accuracy	57.62%	55.25%	57.67%	54.12%
	Kappa	0.2696	0.1908	0.2315	0.0907

Following the incorporation of the subgraph features on an individual basis into the Chengdu test dataset, the increase in the producer's accuracy was more pronounced for the industrial and public categories, with improvements of 16.86% and 10.84%, respectively. For the natural and commercial categories, comparatively modest improvements of 8.33% and 5.61%, respectively, were observed. The user's accuracy generally increased across all four categories. For the residential category, the user's accuracy increased by 4.3%. However, there was a concurrent decrease in the producer's accuracy. For the public category, the combination of the optimized graph structure subgraph features, and $\omega = 0.3$ led to notable improvements of 12.95% and 16.26%, respectively. Despite a decrease in the producer's

accuracy for the natural and residential categories relative to $\omega = 0.3$, there were notable increases in the user's accuracy of 7.66% and 3.76%, respectively. The classification accuracy for the industrial category was comparable to that observed at $\omega = 0.3$ but inferior to that when only subgraph features were utilized. For the commercial category, an intermediate level of accuracy was achieved.

In conclusion, the land use classification effect was enhanced to varying degrees for both sets of experimental data after the addition of subgraph features or the utilization of an optimized graph structure. The concurrent application of an optimized graph structure and subgraph features can increase the precision of the GCN for land use classification at the block scale, with a more pronounced improvement observed compared with that using an optimized graph structure and subgraph features alone.

4. Discussion

4.1. Ablation Experiments and Analysis

To further examine the efficacy of employing deep feature vectors in land use classification through the transformer, we compared two approaches: the aforementioned land use classification approach, which integrates optimized GCN topological structure deep features and VGG-16 raster structure deep features through the transformer, and the land use classification method that fuses optimized GCN topological structure deep features or VGG-16 raster structure deep features through the transformer. A comparative experiment on land use classification was conducted via the transformer in conjunction with different deep features. The effects of the three classification methods on the two sets of experimental data were then analyzed.

Table 6 shows the accuracy results of the ablation experiments from the Shenyang third ring road and Chengdu fourth ring road datasets. On the Shenyang third ring dataset, the method of fusing deep features via the transformer was observed to improve the producer's accuracy for the natural category by 12.91% and 6.46% and the user's accuracy by 14.81% and 24.94% in comparison with those of the method of using the optimized GCN topological structure deep features and VGG-16 raster structure deep features alone, respectively. For the public category, the producer's accuracy improved by 36.7% and 2.75%, respectively, and the user's accuracy improved by 16.49% and 3.7%, respectively. The producer's accuracy for the residential category exhibited improvements of 13.92% and 3.3%, respectively. For the industrial category, the proposed method reduced the producer's accuracy error from 82.55% to 71.81% in comparison to the method utilizing optimized GCN topological structure deep features and from 69.80% to 71.81% in comparison with the method employing VGG-16 raster structure deep features. These improvements in accuracy were particularly notable in the context of the user's accuracy, with increases of 24.25% and 6.05%, respectively. For the commercial category, the producer's accuracy improved by 4.46% and 1.98%, respectively, and the user's accuracy improved by 29.47% and 3.8%, respectively. The overall accuracy on the Shenyang dataset was increased by 13.35% and 2.88%, respectively, whereas the kappa coefficient increased by 0.1867 and 0.0407, respectively. On the Chengdu fourth ring dataset, the method of fusing embedding vectors via the transformer improved the producer's accuracy for the natural category by 8.33% and 5% and the user's accuracy by 22.05% and 13% in comparison with the method of using the optimized GCN topological structure deep features and VGG-16 raster structure deep features alone, respectively. The producer's accuracy for the public category improved by 31.33% and 2.11%, respectively, whereas the user's accuracy improved by 28.11% and 1.81%, respectively. For the industrial category, the producer's accuracy improved by 36.05% and 2.32%, respectively, and the user accuracy improved by 19.07% and 4.05%, respectively. For the commercial category, the producers' accuracy improved by 25.08% and 1.32%, respectively, and the user's accuracy improved by 19.71% and 2.89%, respectively. The overall accuracy on the Chengdu dataset improved by 14% and 1.72%, respectively, and the kappa coefficient improved by 0.269 and 0.0262, respectively.

Table 6. Comparative experimental accuracy evaluation of Transformer model.

Data	Accuracy Indicators	Fusion Vectors	Transformer GCN Vectors	VGG-16 Vectors
Shenyang Third Ring	Natural Producer's Accuracy	80.65%	67.74%	74.19%
	Natural User's Accuracy	92.59%	77.78%	67.65%
	Public Producer's Accuracy	61.01%	24.31%	58.26%
	Public User's Accuracy	73.48%	56.99%	69.78%
	Residential Producer's Accuracy	92.31%	78.39%	89.01%
	Residential User's Accuracy	75.45%	70.63%	74.54%
	Industrial Producer's Accuracy	71.81%	82.55%	69.80%
	Industrial User's Accuracy	84.25%	60.00%	78.20%
	Commercial Producer's Accuracy	53.47%	49.01%	51.49%
	Commercial User's Accuracy	75.52%	46.05%	71.72%
	Overall Accuracy	76.53%	63.18%	73.65%
	Kappa	0.6417	0.4550	0.6010
Chengdu Fourth Ring	Natural Producer's Accuracy	53.33%	45.00%	48.33%
	Natural User's Accuracy	82.05%	60.00%	69.05%
	Public Producer's Accuracy	59.34%	28.01%	57.23%
	Public User's Accuracy	61.56%	33.45%	59.75%
	Residential Producer's Accuracy	89.80%	88.48%	88.38%
	Residential User's Accuracy	79.16%	67.18%	78.55%
	Industrial Producer's Accuracy	58.72%	22.67%	56.40%
	Industrial User's Accuracy	60.12%	41.05%	56.07%
	Commercial Producer's Accuracy	46.86%	21.78%	45.54%
	Commercial User's Accuracy	68.60%	48.89%	65.71%
	Overall Accuracy	73.29%	59.29%	71.57%
	Kappa	0.5689	0.2999	0.5427

This experiment aims to provide a comprehensive analysis of the impact of different deep features on land use classification through the transformer. To this end, the classification details of various classification methods were mapped and evaluated, with a particular focus on the performance differences observed in the handling of specific scene features. The local detail maps of the comparison experiments for the third ring of Shenyang and the fourth ring of Chengdu data are presented in Figure 8 and Figure 9, respectively.

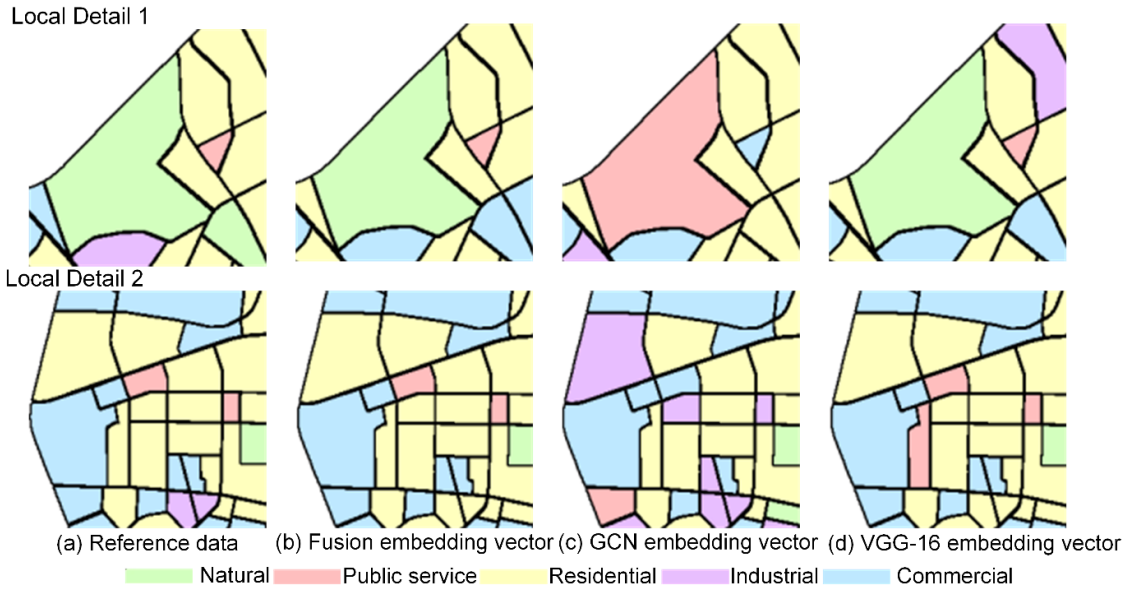


Figure 8. Localized details of the Shenyang third ring ablation experiment.

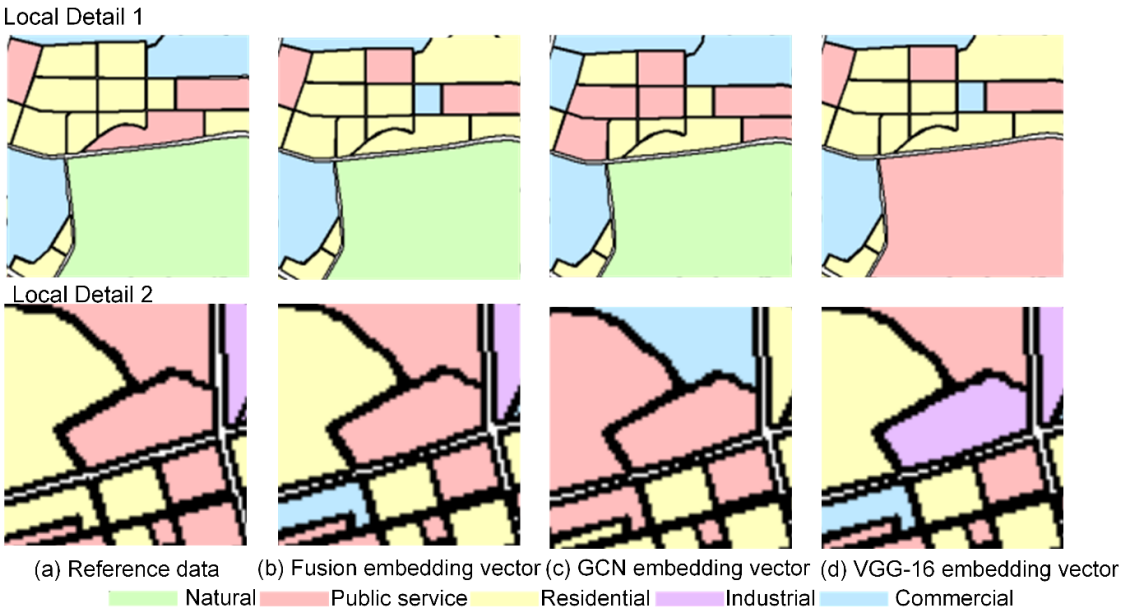


Figure 9. Localized details of the Chengdu fourth ring ablation experiment.

Figure 8 shows that classification utilizing GCN topological structure deep features was inadequate in local detail 1 of the Shenyang third ring comparison experiment, with six misclassified image objects. The utilization of VGG-16 raster structure deep features and fusion embedding vectors yielded comparable and enhanced outcomes relative to the employment of GCN topological structure deep features across four distinct classification error scenarios. In local detail 2 of the Shenyang third ring comparison experiment, the classification of the fused embedding vectors was demonstrably superior to the other two, with only two industrial areas not being successfully identified. As illustrated in Figure 9, the utilization of GCN topological structure deep features in local detail 1 of the fourth ring comparison experiment in Chengdu was observed to be ineffective for classification, resulting in the misclassification of seven image objects. Then, VGG-16 raster structure deep features were used, leading to the misclassification of five image objects. Conversely, the optimal approach was to use fused embedding vectors, which misclassified only four image objects. In local detail 2 of the Chengdu fourth ring comparison experiment, the classification results using GCN topological structure deep features were cluttered; there were three classification errors when VGG-16 raster structure deep features were used, including one residence and two public areas, whereas there were only two classification errors when fusion embedding vectors were used, including one residence and one public area. In general, the classification outcomes of fusing the embedding vectors with the transformer were superior to those of using the optimized GCN topological structure deep features or VGG-16 raster structure deep features with the transformer alone.

4.2. Parametric Analysis Experiment

4.2.1. GCN Parameter Experiment

In the context of extracting embedding feature vectors via the GCN model, the function of the graph convolution layer is to aggregate information pertaining to neighboring nodes of varying orders. The impact of feature extraction can be changed by modifying the number of layers in the graph convolution layer. To identify the optimal GCN structure, we set the number of layers of the GCN to 1, 2, 3, 4, 5, and 6 and employed a statistical approach to evaluate the accuracy of the model with varying parameters on the test set. As illustrated in Figure 10, the utilization of two convolutional layers was demonstrated to increase the feature extraction precision. The extraction of features via a single graph convolutional layer was prone to instability. The inclusion of an excessive number of convolutional

layers may result in the over-aggregation of neighboring node information, which in turn may impact the extraction of context-embedding information. It is crucial to preserve the classification effect of the GCN while maximizing the extraction of contextual embedding information. To this end, we employed a two-layer GCN to extract the topological structure deep features of block-scale image objects.

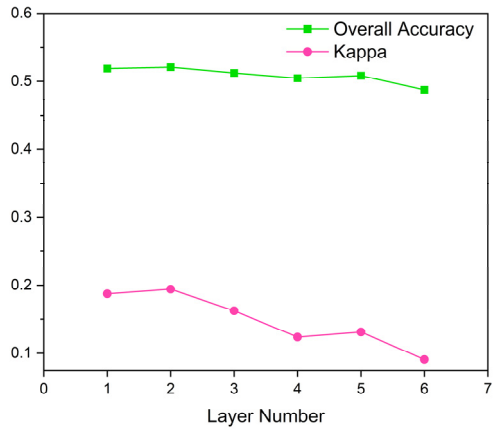


Figure 10. Results of different convolution layers of the GCN.

4.2.2. VGG-16 Parameter Experiment

The accuracy of feature vector extraction via the VGG-16 model was contingent upon the batch size employed. A reduction in the batch size may result in a more challenging training process for the model, whereas an increase in the batch size may lead to a greater demand for memory resources. The batch sizes were set to 32, 64, 128, 256, and 512, with the objective of counting the model accuracies of the test set under different parameters. As illustrated in Figure 11, the overall accuracy and kappa coefficient were greater at batch sizes of 256 than at batch sizes of 32, 64, 128, and 512. Therefore, VGG-16, with a batch size of 256, was selected for the extraction of raster structure deep features of block-scale image objects.

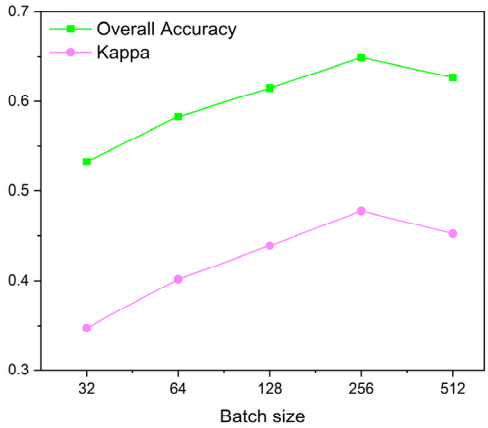


Figure 11. Results of different batch sizes of VGG-16.

4.2.3. Transformer Parameter Experiment

As the number of encoder layers in the transformer increases, the network model capacity also increases, thereby allowing for increased expressive power. Nevertheless, this approach presents certain challenges, including slower model convergence and vanishing gradients. It is thus imperative to select the optimal number of encoder layers when the transformer is used for land use classification. The number of encoder layers was set to 1, 2, 3, 4, 5, and 6 to determine the classification accuracy of the transformer on the test set under different parameters. As illustrated in Figure 12, the overall accuracy and kappa

coefficient were greater when the number of encoder layers was set to 4. Consequently, the number of layers in the encoder was fixed at 4 in the transformer fusion embedding feature vectors for land use classification.

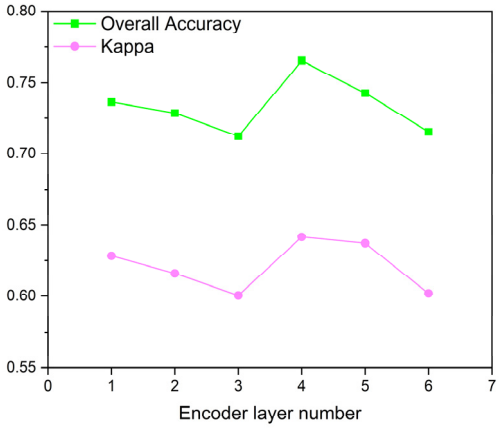


Figure 12. Results of different encoder layers of the transformer.

5. Conclusions

The majority of current land classification methods adopt either the deep features of the raster structures using convolutional neural networks or the deep features of topological structures using graph neural networks. However, these approaches are inadequate for fully describing the semantic information present in remote sensing images. This paper presents a land use classification method at the block scale. This method fuses deep features of GCN topological structures and VGG-16 raster structures via a transformer. The TAZ is employed for the segmentation of high-resolution remote sensing images of the third ring of Shenyang and the fourth ring of Chengdu, with the objective of obtaining block-scale image objects. The optimal graph structure is identified via a GCN, with subgraph features incorporated and the classification effect of the GCN improved. Furthermore, the topological structure's deep features should be extracted. For the segmented image, the raster structure deep features are extracted via a VGG-16 convolution operation. Land use classification is accomplished by fusing the two deep features through the transformer. The experimental results on two datasets demonstrate that the method proposed in this paper effectively incorporates the subgraph structure features into the traditional feature space, enhances the expressive capacity of the traditional KNN graph structure through the KNN composition method with self-adjusting fusion coefficients, and efficiently completes the fusion of the deep features of the two models via the transformer. Furthermore, the classification accuracy is superior to that of a single model by strengthening the deep features of the model and enhancing the representation ability of effective information. These approaches enhance the capacity of the model to identify and characterize the effective information present within the deep features.

Nevertheless, the methodology described in this paper can be further refined. To fully leverage the strengths of diverse models in feature extraction and classification, three independent models are employed to achieve multimodal deep feature fusion for land use classification. Although this multimodal integration approach demonstrated satisfactory classification results, the capacity to integrate the GCN, VGG-16, and transformer into a unified model would facilitate the forward propagation of embedding features. This approach can be achieved by end-to-end transfer of gradient information between the original image and the high-level feature representations, thus further optimizing the feature representation of the nodes. The application of the self-attention mechanism in the transformer to focus on the multimodal deep information within a single block unit represents an effective approach to feature fusion. However, the fusion process does not fully consider the multimodal deep information of neighboring block units. The spatial features between the block units were insufficiently considered. Consequently, we will

consider combining the GNN and the self-attention mechanism to fully exploit the role of spatial features in the feature fusion process and further improve the accuracy of land use classification. The study areas selected in this paper are all located in the northern hemisphere, and due to the large differences in the physical and economic environments of northern hemisphere and southern hemisphere cities, the land use methods of the northern hemisphere cities may not be directly applied to the southern hemisphere cities, and we will consider the issue of land use classification for southern hemisphere cities in subsequent studies.

Author Contributions: Conceptualization, Yougui Ren; methodology, Yougui Ren; software, Yougui Ren; validation, Yougui Ren; formal analysis, Yougui Ren and Zhiwei Xie; investigation, Yougui Ren, Zhiwei Xie and Shuaizhi Zhai; resources, Yougui Ren; data curation, Zhiwei Xie and Shuaizhi Zhai; writing—original draft preparation, Yougui Ren; writing—review and editing, Zhiwei Xie and Shuaizhi Zhai; visualization, Shuaizhi Zhai; supervision, Yougui Ren; project administration, Yougui Ren; funding acquisition, Yougui Ren; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basic Scientific Research Project of the Educational Department of Liaoning Province of China (Grant No. LJKMZ20220946 and JYTMS20231602), Social Science Planning Fund of Liaoning Province of China (Grant No. L23BGL011).

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to restrictions imposed by privacy approval and informed consent agreements regarding study participants.

Acknowledgments: The authors thank the reviewers and editors for their careful review and constructive suggestions, which helped to improve the manuscript. We thank the data distribution agency for providing the publicly available data used in this study. The high-resolution image data were obtained from the official Google Earth website of Google, Inc. Road network data were obtained from OpenStreetMap.

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

References

1. Zheng, K.; Wang, H.; Qin, F.; Han, Z. A land use classification model based on conditional random fields and attention mechanism convolutional networks. *Remote Sens.* **2022**, *14*, 2688. [CrossRef]
2. Srivastava, P.K.; Han, D.; Rico-Ramirez, M.A.; Bray, M.; Islam, T. Selection of classification techniques for land use/land cover change investigation. *Adv. Space Res.* **2012**, *50*, 1250–1265. [CrossRef]
3. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land use classification in remote sensing images by convolutional neural networks. *arXiv* **2015**, arXiv:1508.00092.
4. Weiland, U.; Kindler, A.; Banzhaf, E.; Ebert, A.; Reyes-Paecke, S. Indicators for sustainable land use management in Santiago de Chile. *Ecol. Indic.* **2011**, *11*, 1074–1083. [CrossRef]
5. Morrison, N. Developing Indicators to Evaluate the Effectiveness of Land Use Planning. In *Planning in the UK*; Routledge: Abingdon, UK, 2018; pp. 91–102.
6. Yu, H.; Zheng, C. Environmental regulation, land use efficiency and industrial structure upgrading: Test analysis based on spatial durbin model and threshold effect. *Heliyon* **2024**, *10*, e26508. [CrossRef]
7. Ma, W.; Jiang, G.; Li, W.; Zhou, T. How do population decline, urban sprawl and industrial transformation impact land use change in rural residential areas? A comparative regional analysis at the peri-urban interface. *J. Clean. Prod.* **2018**, *205*, 76–85. [CrossRef]
8. Silvestri, B.; Roccotelli, M.; Iavaniglio, R.P.; Ranieri, L. Optimization of goods relocation in urban store networks with an incentive strategy. *IET Collab. Intell. Manuf.* **2021**, *3*, 13–22. [CrossRef]
9. Talukdar, S.; Singha, P.; Mahato, S.; Pal, S.; Liou, Y.-A.; Rahman, A. Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sens.* **2020**, *12*, 1135. [CrossRef]
10. Ewing, R.; Hamidi, S. Compactness versus sprawl: A review of recent evidence from the United States. *J. Plan. Lit.* **2015**, *30*, 413–432. [CrossRef]
11. Tian, L.; Li, Y.; Yan, Y.; Wang, B. Measuring urban sprawl and exploring the role planning plays: A shanghai case study. *Land Use Policy* **2017**, *67*, 426–435. [CrossRef]
12. Maurício, P.; Mirian Vizintim Fernandes, B. Urban sprawl and the challenges for urban planning. *J. Environ. Prot.* **2012**, *2012*, 22651.

13. Macarringue, L.S.; Bolfe, É.L.; Pereira, P.R.M. Developments in Land Use and Land Cover Classification Techniques in Remote Sensing: A Review. *J. Geogr. Inf. Syst.* **2022**, *14*, 1–28. [CrossRef]
14. Vali, A.; Comai, S.; Matteucci, M. Deep learning for land use and land cover classification based on hyperspectral and multispectral earth observation data: A review. *Remote Sens.* **2020**, *12*, 2495. [CrossRef]
15. Luus, F.P.; Salmon, B.P.; Van den Bergh, F.; Maharaj, B.T.J. Multiview deep learning for land-use classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2448–2452. [CrossRef]
16. Gu, Y.; Jiao, L.; Dong, T.; Wang, Y.; Xu, G. Spatial distribution and interaction analysis of urban functional areas based on multi-source data. *Geomat. Inf. Sci. Wuhan Univ.* **2018**, *43*, 1113–1121.
17. 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]
18. Yao, Y.; Li, X.; Liu, X.; Liu, P.; Liang, Z.; Zhang, J.; Mai, K. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 825–848. [CrossRef]
19. Zhai, W.; Bai, X.; Shi, Y.; Han, Y.; Peng, Z.-R.; Gu, C. Beyond Word2vec: An approach for urban functional region extraction and identification by combining Place2vec and POIs. *Comput. Environ. Urban Syst.* **2019**, *74*, 1–12. [CrossRef]
20. Hu, S.; He, Z.; Wu, L.; Yin, L.; Xu, Y.; Cui, H. A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Comput. Environ. Urban Syst.* **2020**, *80*, 101442. [CrossRef]
21. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
22. Zhang, Y.; Liu, P.; Biljecki, F. Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 153–168. [CrossRef]
23. Chen, Y.; Zhao, P.; Lin, Y.; Sun, Y.; Chen, R.; Yu, L.; Liu, Y. Semantic-Enhanced Graph Convolutional Neural Networks for Multi-Scale Urban Functional-Feature Identification Based on Human Mobility. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 27. [CrossRef]
24. LIU, L.; MA, Z.; SUN, Y.; LI, W.; XIANG, Z. Semantic Segmentation of Street View and Multi-dimensional Feature Identification of City Based on MS-DeepLabV3+. *Geomat. Inf. Sci. Wuhan Univ.* **2024**, *49*, 343–354.
25. Fu, S.; Liu, W.; Zhang, K.; Zhou, Y. Example-feature graph convolutional networks for semi-supervised classification. *Neurocomputing* **2021**, *461*, 63–76. [CrossRef]
26. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
27. 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]
28. Weng, Q.; Mao, Z.; Lin, J.; Guo, W. Land-use classification via extreme learning classifier based on deep convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 704–708. [CrossRef]
29. Yu, J.; Zeng, P.; Yu, Y.; Yu, H.; Huang, L.; Zhou, D. A combined convolutional neural network for urban land-use classification with GIS data. *Remote Sens.* **2022**, *14*, 1128. [CrossRef]
30. Liang, J.; Deng, Y.; Zeng, D. A deep neural network combined CNN and GCN for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 4325–4338. [CrossRef]
31. Li, Y.; Chen, R.; Zhang, Y.; Zhang, M.; Chen, L. Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network. *Remote Sens.* **2020**, *12*, 4003. [CrossRef]
32. Nuissl, H.; Siedentop, S. Urbanisation and land use change. In *Sustainable Land Management in a European Context. Human-Environment Interactions*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 75–99.
33. 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. Geogr. Inf. Sci.* **2017**, *31*, 1675–1696. [CrossRef]
34. Yin, J.; Dong, J.; Hamm, N.A.; Li, Z.; Wang, J.; Xing, H.; Fu, P. Integrating remote sensing and geospatial big data for urban land use mapping: A review. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102514. [CrossRef]
35. Zhou, G.; Zhang, J.; Li, C.; Liu, Y. Spatial Pattern of Functional Urban Land Conversion and Expansion under Rapid Urbanization: A Case Study of Changchun, China. *Land* **2022**, *11*, 119. [CrossRef]
36. Zhou, Q.; Wang, S.; Liu, Y. Exploring the accuracy and completeness patterns of global land-cover/land-use data in OpenStreetMap. *Appl. Geogr.* **2022**, *145*, 102742. [CrossRef]
37. Xu, Y.; Zhou, B.; Jin, S.; Xie, X.; Chen, Z.; Hu, S.; He, N. 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.* **2022**, *95*, 101807. [CrossRef]
38. Huang, W.; Zhang, D.; Mai, G.; Guo, X.; Cui, L. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 134–145. [CrossRef]
39. He, C.; Wei, A.; Shi, P.; Zhang, Q.; Zhao, Y. Detecting land-use/land-cover change in rural–urban fringe areas using extended change-vector analysis. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 572–585. [CrossRef]
40. The Ministry of Housing and Urban-Rural Development of the People's Republic of China. *GB 50137-2011*; Code for Classification of Urban Land Use and Planning Standards of Development Land. China Architecture Publishing & Media Co., Ltd.: Beijing, China, 2011.

41. Zhang, X.; Cui, J.; Wang, W.; Lin, C. A study for texture feature extraction of high-resolution satellite images based on a direction measure and gray level co-occurrence matrix fusion algorithm. *Sensors* **2017**, *17*, 1474. [CrossRef]
42. Patro, S.; Sahu, K.K. Normalization: A preprocessing stage. *arXiv* **2015**, arXiv:1503.06462. [CrossRef]
43. Yang, Y.; Li, H.; Han, Y.; Gu, H. High resolution remote sensing image segmentation based on graph theory and fractal net evolution approach. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 197–201. [CrossRef]
44. Yang, P.; Tong, L.; Qian, B.; Gao, Z.; Yu, J.; Xiao, C. Hyperspectral image classification with spectral and spatial graph using inductive representation learning network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 791–800. [CrossRef]
45. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
46. Tammina, S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *Int. J. Sci. Res. Publ.* **2019**, *9*, 143–150. [CrossRef]
47. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.