

Geo-scenes dissecting urban fabric: Understanding and recognition combining AI, remotely sensed data and multimodal spatial semantics

Hanqing Bao^{a,*}, Lanyue Zhou^a, Lukas W. Lehnert^a

^a Department of Geography, Ludwig-Maximilians-University Munich 80333 Munich, Germany

ARTICLE INFO

Keywords:

Urban fabric
Geo-scenes
Spatial relationship
Distribution pattern
Multimodal deep learning

ABSTRACT

Urban fabric represents the intersection of spatial structure and social function. Analyzing its geographic components, functional semantics, and interactive relationships enables a deeper understanding of the formation and evolution of urban geo-scenes. Urban geo-scenes (UGS), as the fundamental units of urban systems, play a vital role in balancing and optimizing spatial layout, while enhancing urban resilience and vitality. Although multimodal spatial data are widely used to describe UGS, conventional approaches that rely solely on visual or social features are insufficient when addressing the complexity of modern urban systems. The spatial relationships and distributional patterns among urban elements are equally crucial for capturing the full semantic structure of urban geo-scenes. In parallel, most deep learning models still face limitations in effectively mining and fusing such diverse information. To address these challenges, we propose a multimodal deep learning framework for UGS recognition. Guided by the concepts of urban fabric and spatial co-location patterns, our method dissects the internal structure of geo-scenes and constructs a bottom-up urban fabric graph model to capture spatial semantics among geographic entities. Specifically, we employ a customized SE-DenseNet branch to extract deep physical and visual features from high-resolution satellite imagery, along with social semantic information from auxiliary data (e.g., POIs, building footprint coverage). A semantic fusion module is further introduced to enable collaborative interaction among multi-modal and multi-scale features. The framework was validated across four Chinese cities with varying sizes, economic levels, and cultural contexts. The proposed method achieved an overall accuracy of approximately 90%, outperforming existing state-of-the-art multimodal approaches. Moreover, ablation studies conducted in three cities of different scales confirm the critical role of urban fabric in UGS recognition. Our results demonstrate that the joint modeling of visual appearance, functional attributes, and spatial semantics offers a novel and more comprehensive understanding of urban geo-scenes.

1. Introduction

The global urbanization process is accelerating, and by 2050, nearly 70 % of the world's population is projected to reside in urban areas. Cities have become central hubs of human activity and act as catalysts for socio-economic development and civilizational progress (Kong et al., 2024).

Urban geo-scenes (UGS) represent the fundamental functional units of urban systems, encompassing residential neighborhoods, commercial districts, educational institutions, and public parks. Accurately identifying and interpreting UGS enables optimized spatial planning, efficient resource allocation, and enhanced urban vitality—key components of sustainable urban development (Chen and Huang, 2024; Su et al., 2024; Zhang et al., 2022).

From an urban dissection perspective, the city can be regarded as an organic and intricate system (Fig. 2 (A)) (Chen et al., 2021). In this analogy, UGS function as the “organs” of the city, while geo-objects within the scenes—such as buildings, roads, infrastructure, and vegetation—serve as the structural “tissues” that define urban functions, spatial distribution, and organizational structure (Zhang et al., 2023). This layered composition constitutes the urban fabric, reflecting the dynamic interplay among urban components that collectively define the structure and function of the urban system (Çalışkan et al., 2022; Levy, 1999). Street intersections and building layouts establish the rhythm and flow of spatial movement, while public spaces and parks are strategically embedded within the urban fabric to foster social interaction and cultural expression (Moudon, 1997).

In this study, we conceptualize the urban fabric as a structured,

* Corresponding author.

E-mail addresses: hanqing09.bao@gmail.com (H. Bao), lanyue1109@gmail.com (L. Zhou), lehnert.lu@lmu.de (L.W. Lehnert).

<https://doi.org/10.1016/j.isprsjprs.2025.10.011>

Received 6 May 2025; Received in revised form 8 October 2025; Accepted 11 October 2025

Available online 16 October 2025

0924-2716/© 2025 The Author(s). Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

multimodal representation of urban space, capturing not only the physical configuration of geo-elements but also their spatial interrelations and functional semantics. Rather than treating UGS as isolated units, we model them as relational entities embedded within a geospatial system, where function emerges through spatial interaction. Specifically, the urban fabric is characterized along three interlinked dimensions:

1. **Geo-objects:** Including built features (e.g., building height), infrastructure (e.g., road networks), and natural components (e.g., vegetation, water bodies), which collectively define the physical environment
2. **Spatial structure and connectivity:** Referring to the arrangement and interrelation of these elements, captured through topological relations such as adjacency, co-location, and accessibility.
3. **Functional attributes:** Derived from land-use semantics, reflecting human-centered roles such as residential, commercial, industrial, and educational uses

This framing aligns with a geographic perspective, viewing UGS as emergent phenomena shaped by the spatial configuration and interaction of multiple elements. It captures micro-level urban heterogeneity and overcomes the limitations of conventional scene classification, which often relies solely on surface-level visual features. By integrating functional roles and spatial relationships (e.g., co-location, connectivity, distribution), this approach enables a deeper and context-aware understanding of urban systems.

How can we integrate multimodal geospatial data to effectively understand and recognize the composition of UGS? Addressing this core question requires integrating visual patterns, human activity semantics, and spatial co-location structures into a unified representation of the urban fabric. This study designed a UGS recognition framework based on multimodal deep learning from the perspective of urban fabric dissection. Unlike earlier methods that rely solely on visual features (Bao et al., 2020), our framework leverages the essential components of urban fabric—visual features, spatial distribution, and functional

relationships—to provide a richer understanding of urban systems. The framework consists of three key components: bottom-up spatial modeling of geo-objects, multimodal feature extraction, and semantic fusion. We construct a self-designed Multimodal Spatial Semantic Fusion Network (MSSFNet) that effectively integrates heterogeneous data to achieve comprehensive recognition of UGS. In summary, the main contributions and innovations of this paper are as follows:

1. Building on the concept of urban fabric dissection, we perform bottom-up spatial relationship modeling of urban elements to achieve a fundamental and interpretable understanding of UGS, filling a gap in prior research and supporting urban planning and sustainable development.
2. We propose a multimodal spatial semantic fusion network that integrates multi-source geospatial data, jointly considering visual features, functional, and spatial relationship features. This design mitigates feature fragmentation and enhances cross-modal interaction, effectively improving UGS understanding and recognition.
3. We develop two complementary network modules: SE-DenseNet for deep physical feature extraction and EA-GAT for spatial semantic reasoning. These complementary networks enhance the modeling of complex UGS and improve recognition accuracy by jointly capturing visual and spatial characteristics.
4. The proposed framework is validated across four Chinese cities with varying scales and socio-economic contexts, demonstrating its effectiveness and robustness. Ablation experiments and result analysis further reveal how urban fabric structure influences UGS recognition.

The remainder of this paper is organized as follows: In [Section 2](#), we briefly review related studies on spatial understanding and functional recognition of UGS. [Section 3](#) elaborates on the proposed framework for UGS understanding and recognition from the perspective of urban fabric dissection. [Section 4](#) introduces the study area and experimental data used. Experimental results and analyses are presented in [Section 5](#). [Section 6](#) provides a summary of key conclusions and their impacts.

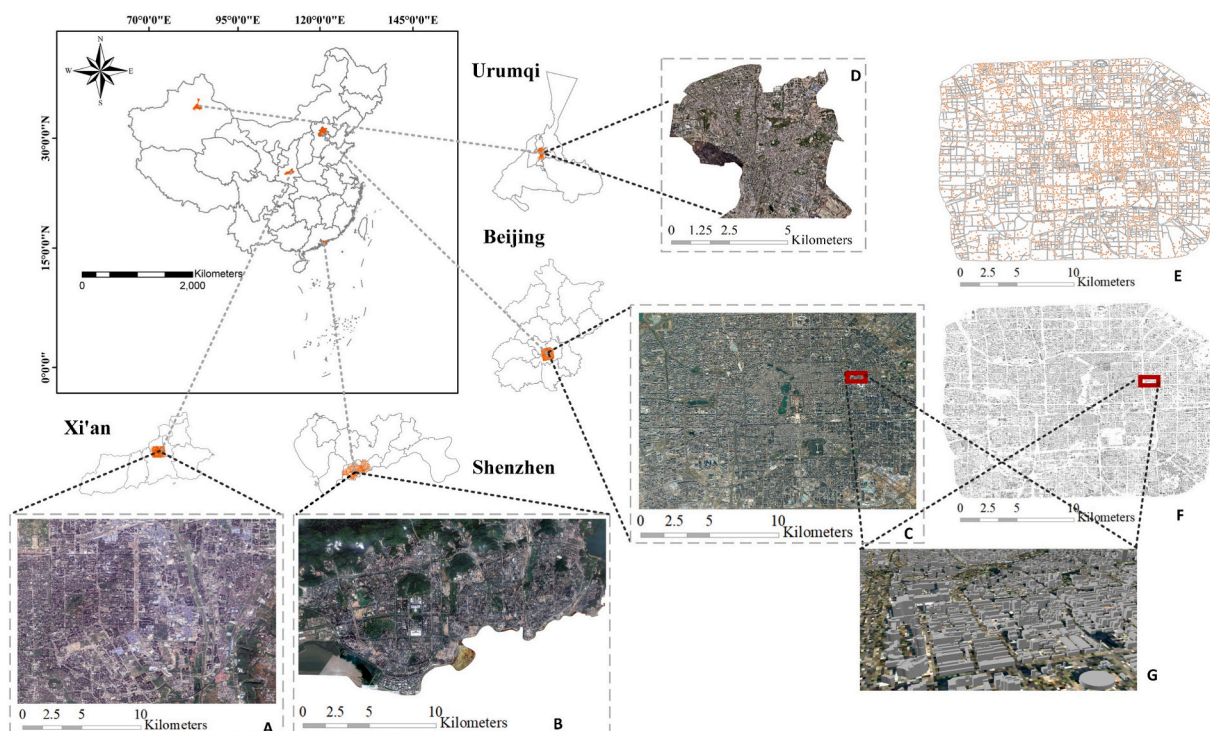


Fig. 1. Locations of four research areas and their RS images.

2. Background and related work

The development of UGS research is driven by both Earth observation technology and dynamic social activities. From the perspectives of data sources and recognition methods, existing studies can be categorized into three categories: satellite imagery-based research, social perception data-based research, and multimodal data fusion approaches (Li et al., 2024a).

2.1. Research based on satellite imagery

With the rapid evolution of Earth observation technologies, satellite imagery has progressed from coarse monitoring to fine-grained analysis, significantly advancing UGS recognition. However, medium- and low-resolution imagery often poses challenges due to mixed pixels and spectral confusion—particularly in heterogeneous urban environments (Du et al., 2021; Zhou et al., 2023). In contrast, high-resolution (HR) and very high-resolution (VHR) remote sensing data (GeoEye, WorldView, QuickBird, and the Gaofen series etc.) allow for a finer-scale distinction of information, enabling the understanding of complex surface structures and achieving precise recognition of urban functions (Li et al., 2023; Yuan et al., 2022).

High spatial resolution imagery supports the generation of multi-scale features of ground objects—from macro-level spatial patterns to micro-level structural details. Traditional methods utilize shallow features (such as spectral, shape, texture features and commonly used indices) for direct classification of UGS (Wen et al., 2020). However, these low-level features are only effective in simple scenes and are insufficient to represent and interpret the complexity of modern land cover types (Du et al., 2020; Lv et al., 2021; Zhang et al., 2022). In recent years, deep learning models based on deep neural architectures have demonstrated significant potential in the remote sensing field for

capturing high-level spatial features, addressing the complexity of urban geo-object semantics and distributions (Du et al., 2024; Zheng et al., 2024). Convolutional neural networks (CNNs) and Transformers have been widely applied to extract intricate spatial representations from high-resolution imagery, outperforming traditional methods in urban geo-scene classification (Bai et al., 2024; Fan et al., 2022; Liu et al., 2022; Zhu et al., 2022). Additionally, numerous variants and multimodal extensions have emerged, including: convolutional modules refined for deeper feature extraction (Chen et al., 2022a), attention mechanisms introduced to focus on key features (Chen et al., 2022b; Ouyang et al., 2025), graph-based convolutions emphasizing neighborhood information and element relationships (Chen et al., 2025; Ma et al., 2024), contrastive learning models capturing intrinsic data structures and feature representations (Gong et al., 2025; Guo et al., 2024a), and Vision Transformers (ViT) establishing long-range dependencies through global information (Bai et al., 2025; Xu et al., 2024; Zhou et al., 2024). These advancements have achieved notable success in improving the understanding and recognition of UGS through remote sensing.

2.2. Research based on social perception data

With the rapid advancement of sensor technologies and information and communication technology (ICT), social sensing data has become an important resource for quantifying and interpreting UGS (Lu et al., 2022; Su et al., 2024; Zhong et al., 2023). In contrast to satellite imagery—which primarily captures physical characteristics—social perception data offers a complementary lens for understanding urban functional semantics (Chen et al., 2024; Li et al., 2024). Various forms of human activity data, including transportation trajectories, social media content (e.g., Twitter, Facebook, Weibo), and mobile signaling records, capture functional patterns of urban spaces through latent human

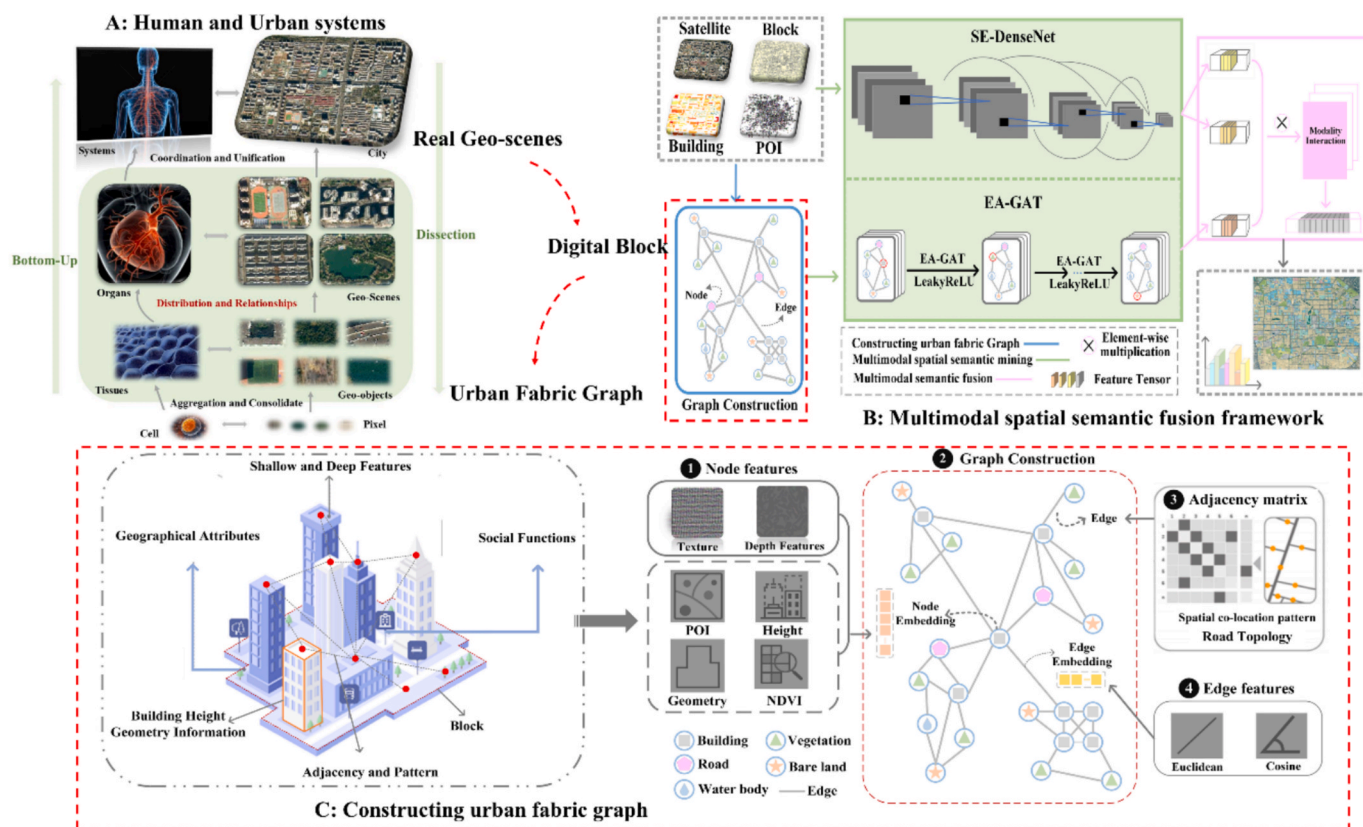


Fig. 2. Multimodal spatial semantic framework for Geo-scenes dissecting urban fabric. (A: Human body and urban system; B: Multimodal semantic fusion model; C: Construction of urban fabric graph).

activities (Hu et al., 2023; Hu et al., 2021; Sun et al., 2025). Simultaneously, data sources such as points of interest (POIs), street view imagery, and nighttime light intensity enable the identification of UGS based on the spatial distribution of urban functions (Fan et al., 2025; Huang et al., 2022; Zhang et al., 2020).

These new data types have also led to the emergence of innovative methodological frameworks. For example: multilayer perceptron models enable the exploration of latent urban functional distributions (Huang et al., 2024; Zhang et al., 2018); natural language processing (NLP) models can mine text-based social sensing data (Lin et al., 2025; Lore et al., 2024); spatial relational data effectively utilize critical information through spatial topology in graph convolutional network (GCN) simulations (Lei et al., 2024; Wang et al., 2025; Zhang et al., 2024); and the self-attention mechanism and multilayered architecture of Transformers make them well-suited to handle the multimodal and spatial dependency relationships within social perception data (Cheng et al., 2023; Liu et al., 2024a). The integration of these AI techniques with social sensing data enables the discovery of high-level latent features in human activity, ultimately supporting the mapping and interpretation of urban functions.

2.3. Research on multimodal data fusion

UGS understanding and recognition depend not only on the static physical attributes of geo-objects and dynamic human activity semantics, but also on the complex spatial structures and interactions among urban elements (Liang et al., 2023). Relying on single-source data or unimodal models often leads to fragmented insights—failing to simultaneously capture physical patterns, social meanings, and spatial relationships within urban systems. To address this, recent studies have proposed comprehensive, accurate, and flexible frameworks that integrate multimodal data and multi-model fusion strategies (Guo et al., 2024b; Hong et al., 2023; Ouyang et al., 2023).

Deep learning techniques help bridge modality gaps and reduce heterogeneity across datasets, making them ideal for integrating physical, social, and spatial signals. Consequently, growing research efforts have focused on combining satellite imagery with social sensing data using deep learning, in order to jointly represent physical structures, socio-economic behaviors, and complex spatial interactions—ultimately improving UGS recognition performance (Li et al., 2024a). Among these, the fusion of satellite images with POI data has become a standard approach for UGS classification (Huang et al., 2023). Moreover, emerging human-centered geospatial data sources—such as building footprints (BF), location-based services (LBS), and street view images (SVIs)—have further enriched the modeling of urban semantics and spatial relationships (Liu et al., 2024b). Notably, the introduction of CLIP has enabled the seamless integration of text-based social perception data with image-based satellite observations, expanding the possibilities for UGS analysis (Huang et al., 2024).

In summary, existing studies demonstrate that multimodal data fusion significantly enhances UGS recognition. However, two key challenges remain: (1) the spatial distribution and relationship descriptions are often vague and lack interpretability, relying solely on latent features within deep learning networks to represent spatial semantics (Bao et al., 2024); (2) merely overlaying multimodal data lacks a fundamental scientific explanation of urban spatial and functional structure (Zhang et al., 2023). Thus, To resolve these limitations, we advocate for a bottom-up modeling approach grounded in urban fabric dissection, which emphasizes relational understanding among geo-objects and offers more interpretable insight into urban structure and functionality (Zhang et al., 2018).

3. Study area and data sets

The functional layout of Chinese cities exhibits a highly complex, multi-layered, and multifunctional spatial pattern, which poses

significant challenges for urban planning and management.

Fig. 3 illustrates four representative study areas that differ in urbanization level and functional composition. These areas cover eastern, southern, western, and northern regions of China, reflecting both coastal–inland geographical contrasts and regional cultural variations. In terms of modernization level and socioeconomic factors, Beijing and Shenzhen are classified as mega first-tier cities, while Xi'an and Urumqi are classified as first- and second-tier cities, respectively.

Beijing, located in North China, is the nation's political, cultural, and international exchange hub. It is recognized as a world-class city, featuring ultra-modern infrastructure and exerting significant economic influence. Its urban structure is characterized by the interweaving of tradition and modernity.

Shenzhen, in southern China near Hong Kong, is a port city that emerged with China's economic reform era. It has rapidly evolved into one of the fastest-growing international metropolises, distinguished by its strong emphasis on modernization and technological innovation.

Xi'an, situated in central China, marks the starting point of the ancient Silk Road. Its urban form is characterized by a harmonious integration of historical legacy and modern urban lifestyle.

Urumqi, one of China's most inland cities, serves as a gateway city in Western China with a significant role in foreign trade. Its urban character combines modern infrastructure with rich ethnic cultural elements.

These four study areas were selected after careful consideration to test and assess the effectiveness and transferability of the model proposed in this study. By encompassing cities with varying degrees of urbanization, economic development, and cultural backgrounds, the experiment aims to validate the applicability of the urban fabric concept in understanding and recognizing UGS.

The satellite imagery used in this study was acquired from the SuperView-1 satellite, offering a spatial resolution of 0.5 m. POI data were collected from Amap (<https://lbs.amap.com/>), and urban building footprint and height data were sourced from the 3D-GloBFP dataset (<https://zenodo.org/records/11397015>) (Che et al., 2024). Table 1 summarizes the dataset specifications. Additionally, EULUC-China and OSM products were used for auxiliary verification.

4. Methodology

The proposed multimodal spatial semantic fusion framework for urban geo-scene (UGS) recognition consists of three stages (Fig. 2. (B)). First, multimodal data are utilized to construct a bottom-up Urban Geo-scene Fabric Graph (UGFG). Secondly, the dual-branch SE-DenseNet is employed to extract deep visual features and social-functional semantics, while the EA-GAT module reconstructs the spatial dependencies inherent in the urban fabric. Finally, a multimodal semantic fusion module integrates deep convolutional representations with transformer-based encoding to achieve comprehensive cross-modal semantic interaction and fusion.

4.1. Constructing urban fabric graph

Urban geo-scenes are composed of diverse geo-objects whose intricate spatial arrangements and interrelations form the “urban fabric.” Like a textile woven from colorful threads, this urban fabric emerges from the physical, spatial, and social components that collectively define a city's character. Following this concept, we decompose urban geo-scenes to construct a bottom-up texture map of urban geo-objects. As illustrated in Fig. 2. (C), urban geo-scenes are modeled not as isolated geographic entities, but as interconnected spatial networks whose interactions shape the city's spatial patterns and functional dynamics.

4.1.1. Defining the urban fabric graph

We choose the adaptive-scale segmentation algorithm based on average local variance, which has an advantage in accurately expressing

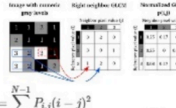
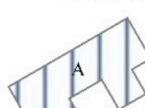


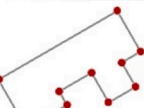

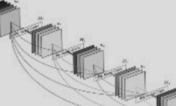
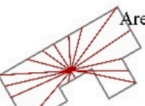
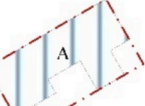
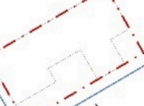


Texture Features	Area	Perimeter	Compactness	Node	NDVI and NDBI
 $Com = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2$ $Ent = \sum_{i,j=0}^{N-1} P_{ij}(-\ln P_{ij})$ $Hom = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$			 $Compactness = 4\pi A/P^2$	 $Node\ count = \sum n$	 $NDVI = \frac{NIR - Red}{NIR + Red}$ $NDBI = \frac{SWIR - NIR}{SWIR + NIR}$
Deeper Features	Shape Index	Regularity	Length-Width Ratio	Orientation	POI Kernel Density
 $F = Sigmoid(W \cdot Conv(X) + b)$	 $Shape\ index = \sum_{i=1}^n \frac{r_i}{\sum_{i=1}^n r_i} - \frac{100}{m}$	 $Regularity = A/A_{MBG}$	 $Length-Width\ ratio = L/W$	 $Orientation = \csc\left(\frac{x}{\sqrt{x^2 + y^2}}\right)$	 $f(x, y) = \frac{1}{n \cdot h^2} \sum_{i=1}^n pop_i K_0\left(\frac{dist_i}{h}\right)$ $K_0(t) = \frac{3}{\pi} (1 - t^2)^2$

Fig. 3. Detailed description and equations of multimodal features.

different scales of urban elements (Bao et al., 2024). Compared with the referenced publication, our segmentation focuses on the urban interior by refining categories for buildings, structures, and roads, while retaining urban green spaces and water bodies to better align with urban surface recognition and urban fabric modeling.

The centroids of segmented geo-objects are simplified into graph nodes, and spatial or functional linkages between them are defined as edges. The urban geo-scenes fabric map is represented as: $G_{UGS} = (GeoN, GeoE, GeoA)$, where $GeoN$ represents the set of nodes and $GeoN \vee n$ indicates that there are n nodes in the UGS. $GeoE$ represents the set of edges, and the adjacency matrix $GeoA \in R^{n \times n}$ records the adjacency relationships between nodes, if there is an edge between the nodes, then $GeoA = 1$, otherwise $= 0$.

Node features encode geometric, visual, and functional attributes of each geo-object, while edge features describe the spatial distance and directional relationship between node pairs. The vector representation is as follows: $GeoF_i = [F_{i1}, F_{i2}, \dots, F_{ik}]$; $GeoR_{ij} = [R_{ij,1}, R_{ij,2}, \dots, R_{ij,k}]$, Where $GeoF_i$ is the feature vector of node $GeoN_i$, k is the dimension of the feature, F_{ik} is the k -th specific attribute. $GeoR_{ij}$ is the edge feature connecting $GeoN_i \wedge GeoN_j$, $R_{ij,k}$ represents the k -th specific information.

Fig. 3 shows the detailed description and equations of the node features of different geo-objects. Texture features and deep convolutional features are set as common attributes. They are the key to expressing different geo-objects and surface cover types. Especially when the targets have similar spectral features, they can provide richer spatial information, thereby enhancing the accuracy of classification and recognition (Kong et al., 2024; Lei et al., 2024).

Moreover, to more accurately characterize the fabric properties of urban geo-scenes, both node and edge features are refined under the guidance of geographic domain knowledge. Specifically, for individual buildings or special structures, we extract classical geometric descriptors, elevation data, and the Normalized Difference Built-up Index (NDBI). When multiple buildings are present, their attributes are aggregated to form a node feature vector. Functional semantics are derived by computing kernel density maps of Points of Interest (POIs), which are then processed through SE-DenseNet to extract diverse semantic patterns—providing functional descriptors that complement the visual and structural properties of the urban fabric. For vegetation, bare land, and water bodies, the Normalized Difference Vegetation Index (NDVI) is additionally calculated to enhance classification with environmental context (Eisenschink et al., 2025; Ji et al., 2024; Tucker, 1979).

For edge features matrix, we utilize both the Euclidean and cosine

distances between adjacent nodes (Lin et al., 2024), and introduced the Gaussian decay function F_{GD} to capture the strength of adjacent object interactions as auxiliary features (Foley and Dorsey, 1984; Izrailev and Castañeda-Mendoza, 2006). The calculation formula is as follows.

$$F_{GD} = e^{-\alpha(D_{ij})^2}, \alpha > 0 \quad (1)$$

where e is the natural logarithm, α is the decay factor, and D_{ij} is the Euclidean distance between two adjacent nodes.

The calculation of the Euclidean distance D_{ij} and Cosine distance cos_{ij} between two nodes $GeoN_i$ and $GeoN_j$ is given by the following formula:

$$D_{ij} = \sqrt{\sum_{c=1}^n (GeoN_{ic} - GeoN_{jc})^2} \quad (2)$$

$$cos_{ij} = \frac{\sum_{c=1}^n GeoN_{ic} * GeoN_{jc}}{\sqrt{\sum_{c=1}^n GeoN_{ic}^2} \sqrt{\sum_{c=1}^n GeoN_{jc}^2}} \quad (3)$$

where c is the c -th element in the node $GeoN$ vector and n is the dimension of the node vector.

4.1.2. Optimizing adjacency

In many prior studies, spatial descriptions of urban geo-scenes are often vague and overlook the influence of spatial distribution and inter-object relationships. To address this, we introduce spatial co-location patterns to optimize adjacency modeling within the UGS (Shekhar and Huang, 2001; Yu, 2016).

Spatial co-location patterns are a powerful approach in geographic analysis and spatial context inference. They reveal geo-objects or events that exhibit spatial proximity and frequent co-occurrence in the real world. This aligns with the first law of geography: “Everything is related to everything else, but near things are more related than distant things.”

Accordingly, incorporating spatial co-location patterns under the constraint of the street network enhances the reliability and contextual meaning of socio-functional inference (Dong et al., 2020).

As illustrated in Fig. 4, each urban geo-scene is abstracted in a Cartesian coordinate system, where the road network defines spatial boundaries, and geo-objects are represented as nodes. For each node, its nearest street segment is determined through a nearest-neighbor search and mapped via linear referencing. Each geo-object is encoded as a tuple $\langle s, pos \rangle$, where s denotes the nearest street segment and pos indicates the projected distance from the origin of s to the geo-object.

Thus, under the constraints of such a network, the spatial adjacency

Table 1
Details of the data.

Multimodal Data	Reference
Satellite Imagery	SuperView-1 satellite data, with a spatial resolution of 0.5 m and 4 multispectral bands (red, green, blue, near infrared) Beijing (24.06.2022): Dongcheng District, Xicheng District, Haidian District and Fengtai District; Shenzhen (05.08.2022): Futian District and Luohu District; Xi'an: (30.06.2021): Xincheng District, Lianhu District, and Beilin District; Urumqi: (11.07.2021): Shuimogou District, and Xinshi District.
Road Network (2022–2024)	Roads are the natural dividing lines of urban geo-scenes, and the irregular blocks formed by their interweaving are the basic units of urban planning and social functions. Road network data mainly includes urban main and secondary roads, tertiary roads, local roads and small roads. To ensure the quality of the blocks, the road network data is preprocessed (Fig. 1. (E)). (Extending independent road lines to connect with adjacent roads; removing unnecessary internal roads, etc.)
POI (2022)	POI is a product of human dynamic activities and contains rich socio-perceptual information. To ensure its precise and effective utilization, we filtered and reclassified the POI data to align with the UGS classification. Additionally, the POI data for the four cities corresponds in time with the satellite imagery.
Building Footprints Building Height	Building footprint and height are important spatial indicators of cities. The latest 3D-GloBFP data (2020) and OSM (2024) can provide accurate three-dimensional building space information, which plays an important role in understanding the spatial structure, functional distribution and planning layout of the city (Cai et al., 2023; Che et al., 2024). (Fig. 1. (F,G))
Urban Geo-Scenes	According to the <i>Urban Land Use Classification and Planning Construction Land Standards</i> and considering the types and attributes of urban geo-scenes, this study reclassifies UGS into 13 categories: Government and Administration (GA), Public Services (P), Residential Zones (R), Shantytown (ST), Underdeveloped (U), Commercial Zones (C), Business and Office Zones (BO), Industrial Zones (I), Education Services and Research (ER), Healthcare (H), Green Spaces and Parks (G), Transportation, Stations, Parking (T), and Places of Interest and Museums (PM).
EULUC-China (2018) EULUC-China 2.0 (2022) OSM Data (2022)	EULUC-China (2018), EULUC-China 2.0 (2022) and OSM (2022) are both existing mature spatial dataset products of land use/cover change, which are used to evaluate and monitor the accuracy and effectiveness of the results of the framework proposed in this paper.
Ground Truth (2021–2022)	The ground truth dataset was derived from high-resolution satellite imagery (2021–2022), historical Google Earth data (2021–2022), and POI data (2022). Preliminary urban geo-scene labels were generated through expert visual interpretation by specialists (3), followed by manual verification and refinement by independent reviewers (2) to ensure both semantic accuracy and temporal consistency with the imagery acquisition period.

relationship can be further defined as: For two urban geographic objects $GeoN_i \wedge GeoN_j$, if the shortest network distance from the mapping of $GeoN_i$ to the location $GeoN_j$ is less than or equal to the user-specified threshold, then they are adjacent. At the same time, considering that buildings are the main components of the geo-scene, all buildings will also be connected.

Hence, through the spatial co-location pattern, not only can the distribution and relationship of urban geo-objects be truly restored, but also unnecessary adjacency relationships can be optimized, reducing the complexity of network calculation.

The distance threshold is dynamically determined by computing the average network distance within the urban geo-scene.

4.2. Multimodal spatial semantic mining

4.2.1. DF-CNN: Deeper features catcher

To address the inefficiencies caused by feature redundancy in traditional deep convolutional neural networks (CNNs), we adopted the DenseNet architecture, which promotes feature reuse and captures multi-scale deep representations. Inspired by the self-attention mechanism, we further construct a custom deep densely connected convolutional network, SE-DenseNet, which enhances spatial encoding by capturing hierarchical image features while maximizing the utility of multi-scale semantics through dense connections.

Fig. 5 displays the network structure of SE-DenseNet. It is mainly composed of 4 Se-Dense Blocks, and 3 transition layers. The visual semantic features of the remote sensing image branch are generated as the output of the 4th layer. Se-Dense Blocks is a key module. Its main function is to achieve efficient feature reuse and information flow through dense connections. The core idea is that the input of each layer is the cascade of the outputs of all previous layers. The formula is as follows: $V_l = H_l([V_0, V_1, \dots, V_{l-1}])$, Where V_l represents the output of the feature vector of the l -th layer. H_l is the conversion function of the l -th layer, and $[V_0, V_1, \dots, V_{l-1}]$ means concatenating the outputs of all layers before the l -th layer as the input of the l th layer.

The transition layer serves as a bridge between SE-Dense blocks. Its primary function is to regulate network complexity and control the dimensionality of feature tensors—thereby optimizing feature flow, reducing computational overhead, and preserving critical semantic information. Its operation can be expressed as:

$$T_l = \text{AvgPool}(W \bullet T_i) \quad (4)$$

where T_i is the feature tensor output after the l -th SE-Dense Block, W is the weight matrix of the 1×1 convolution kernel, and AvgPool is a flat pooling operation to reduce the size of the feature tensor.

The SE module represents the channel attention mechanism module, which can adaptively assign weight to the channel, enhance the network's attention to important features, and suppress the influence of unimportant features. This module mainly consists of two parts: squeeze operation and excitation operation.

The squeeze operation represents the global response on the feature channel, and the spatial information of each channel is globally averaged and pooled.

The spatial information is compressed into a global description vector along the channel dimension to represent the importance of each channel.

The computation formula is as follows:

$$T_l = F_{sq}(X_c) \frac{1}{W^*H} \sum_{i=1}^W \sum_{j=1}^H V_{ij,c} \quad (5)$$

F_{sq} is the Squeeze operation function, $V \in R^{H \times W \times C}$ is the input feature map, W^*H is the size, C is the number of channels, and T_l is the compressed feature vector after squeezing.

The excitation operation rescales the weight coefficient s of each channel using the parameter W . The parameter W is used to learn the correlation between the simulated channels.

$$s = F_{ex}(T_l, W) = \sigma(W_2 \delta(W_1, T_l)) \quad (6)$$

F_{ex} is the excitation operation function and T_l is the result of the squeeze operation. W_1 and W_2 are the dimension reduction and expansion parameters, σ and δ are activation functions *sigmoid* and *ReLU* respectively.

The Reweight operation involves using the output weights s . These weights are then element-wisely applied through multiplication to the previous features, resulting in the recalibration of the original features along the feature-space dimension.

$$\tilde{V}_c = F_{scale}(V_c, s_c) = s_c \bullet V_c \quad (7)$$

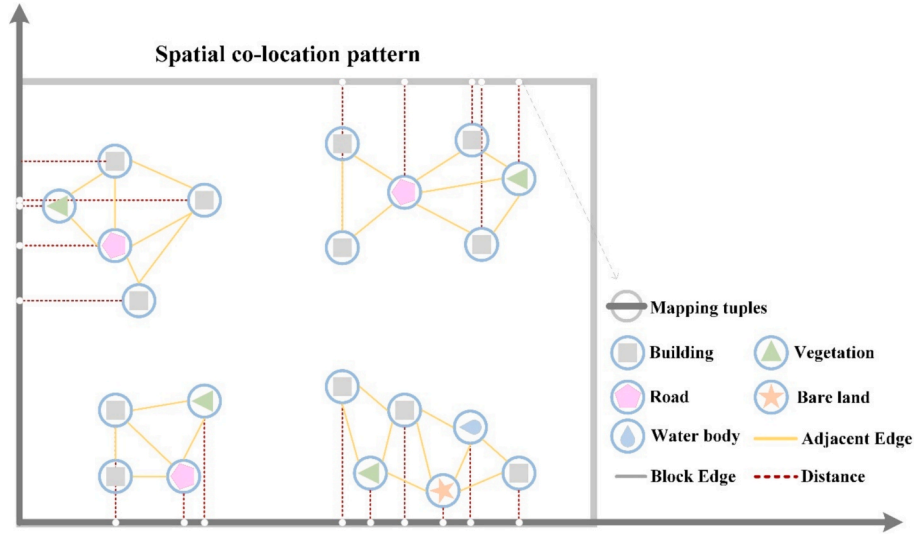


Fig. 4. Spatial co-location pattern optimizes adjacency.

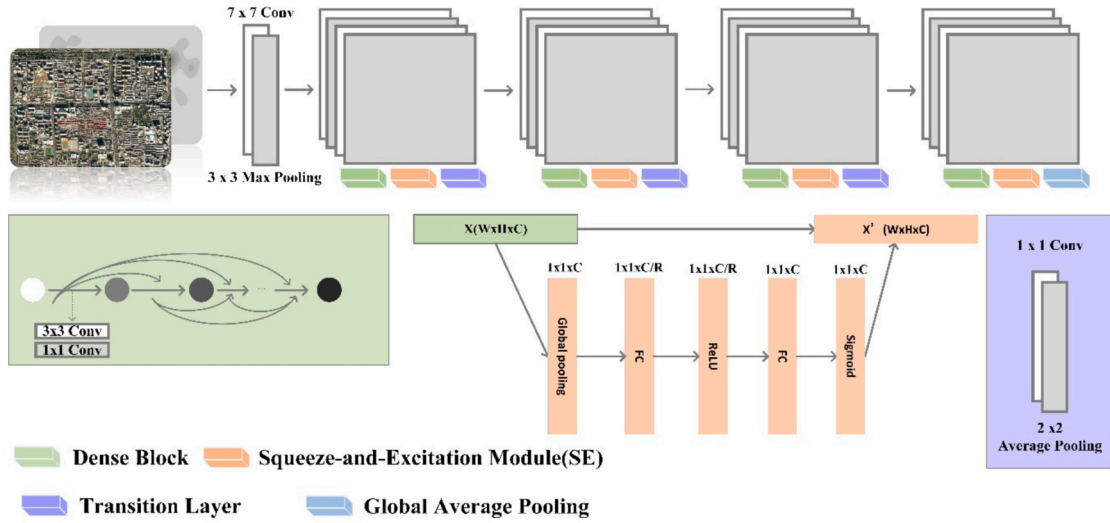


Fig. 5. Demonstration of Deeper Features Catcher (DF-CNN).

F_{scale} is the element-wise multiplication function. $V \in R^{H \times W \times C}$ is the input feature map, $W \times H$ is the size, C is the number of channels.

4.3. 4.2.2EA-GAT: Spatial semantics Detector

Geographic objects within urban geo-scenes are not merely adjacent entities but are interdependent and mutually influenced. To capture the spatial distribution and relational dependencies among urban geo-objects, we employ a Graph Attention Network with Edge Attention (EA-GAT). This model incorporates a self-attention mechanism and jointly utilizes both node and edge information to dynamically assign attention weights to neighboring nodes. By emphasizing the contributions of more relevant neighbors, EA-GAT enables a more accurate and context-aware representation of spatial structure, thereby better capturing the inherent characteristics of the urban fabric (Fig. 6).

Node feature aggregation: For each node v , its new feature tensor T_v can be expressed as:

$$\tilde{T}_v = \sigma\left(\sum_{u \in N(v)} \alpha_{vu} \bullet WT_u\right) \quad (8)$$

where W is a learnable weight matrix, T_u is the feature tensor of node u ,

and α_{vu} is the attention weight between node $GeoN_v$ and its neighboring node $GeoN_u$. σ is the activation function ReLU, which is used to perform nonlinear transformation on the aggregated features.

Attention mechanism: The neighboring feature weights are calculated using the characteristics of the edge. The formula is as follows:

$$\alpha_{vu} = \text{Softmax}\left(\frac{e_{vu}}{\sum_{u \in N(v)} \alpha_{vu}}\right) \quad (9)$$

$$e_{vu} = \text{LeakyReLU}(a^T [WT_v \vee WT_u \vee E_{vu}]) \quad (10)$$

where e_{vu} is the attention score of the edge (v, u) , E_{vu} is the edge feature matrix, a is a learnable parameter, and \vee represents the concatenation operation. The neural network activation function *LeakyReLU* is used.

Global Pooling: Aggregate all building nodes in each block to form an overall graph feature tensor, which is used to describe the spatial semantics of urban geographic scenes.

4.4. Multimodal semantic fusion

The diversity of fused features plays a critical role in enhancing the recognition of urban geo-scenes. The proposed multimodal semantic

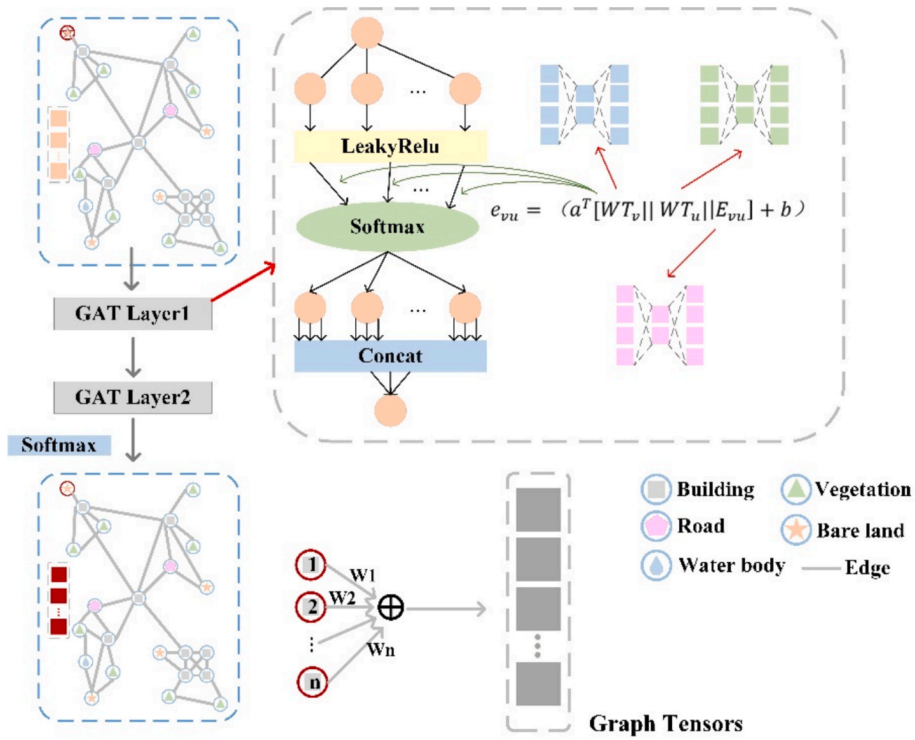


Fig. 6. Demonstration of Spatial Semantics Detector (EA-GAT).

fusion module comprises two key components: a multi-scale semantic module and a modality interaction module (Fig. 7).

Multi-scale semantic module: to reduce the number of parameters,

we use two deep convolution modules to replace the traditional convolution. Depthwise convolution is employed to capture localized semantic patterns, while applying convolution kernels of varying scales

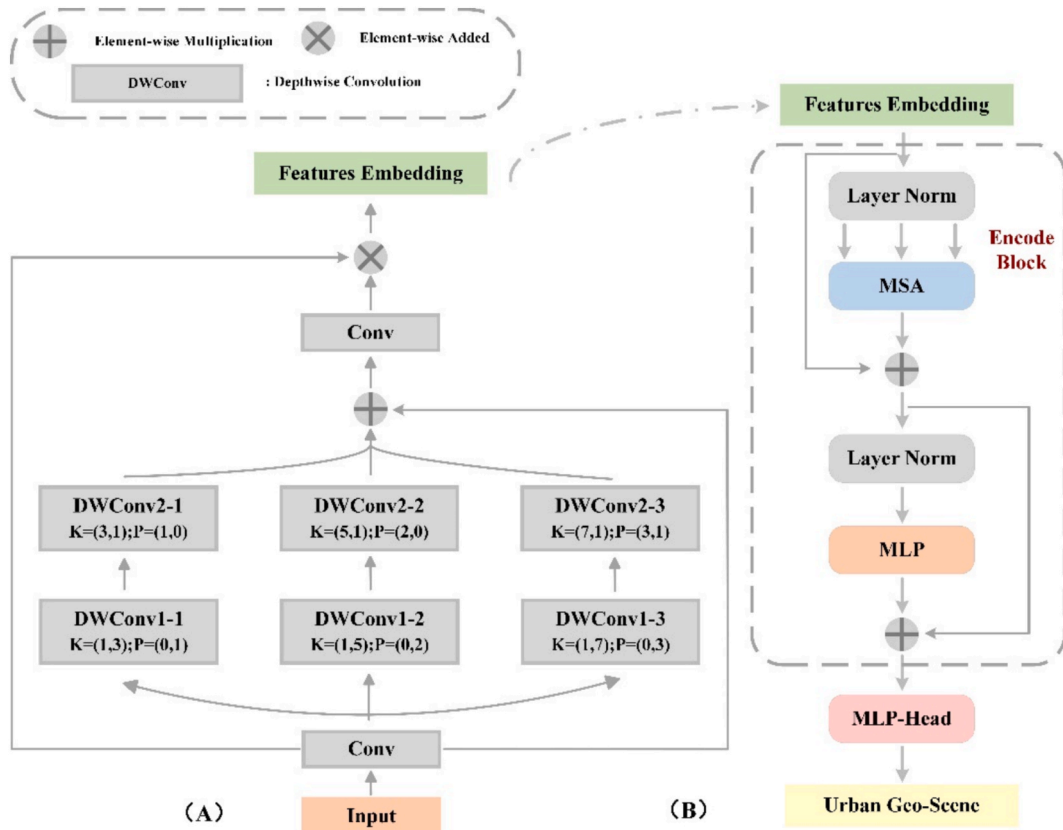


Fig. 7. Multimodal semantic fusion module: (a) multi-scale semantic module; (b) modal interaction module.

allows for the refinement of multi-scale contextual information. The resulting multi-scale semantic features are aggregated via element-wise summation, followed by pointwise convolution to adjust channel dimensionality and model inter-channel dependencies.

Modal interaction module: Inspired by the visual transformer, we introduce a modal interaction module to model multimodal fusion semantics. Each block consists of a multi-head self-attention module (MSA) and a two-layer multi-layer perceptron (MLP), with layer normalization (LN) applied between them to improve robustness and training stability. Finally, a classification head is appended to the output of the transformer encoder to predict the probability distribution over urban geo-scene categories.

The calculation of the ViT module is as follows:

$$V_l = \text{MSA}(\ln(V_{l-1})) + V_{l-1}, l = 1, \dots, L \quad (11)$$

$$V_l = \text{MLP}(\ln(V_{l-1})) + V_{l-1}, l = 1, \dots, L \quad (12)$$

We apply a Multi-Head Self-Attention (MSA) mechanism to capture cross-modal semantic interactions in the fusion module. Multiple heads are computed in parallel and concatenated:

$$\text{MSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (13)$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (14)$$

where W_i^Q, W_i^K, W_i^V are the learned projection matrices for queries, keys, and values in head_i , W^O is the output projection matrix applied after concatenating all heads. h is the number of attention heads.

The MLP is the process from input to output and is calculated as follows:

$$v^l = \sigma(W^l a^{l-1} + b^l) \quad (15)$$

where v^l is the output, σ is the activation function, W^l is the weight matrix of the l -th layer, a^{l-1} is the activation value of the $l-1$ layer, and b^l is the bias vector of the l -th layer.

4.5. Loss function

Cross entropy loss is suitable for category recognition (Ho and Wookey, 2020; Mao et al., 2023), and it is usually used as an optimization function of the target. To optimize the classification results, setting the weight of the category can solve the problem of category imbalance in the data set. Category balance weighting (Li et al., 2018) can reduce the dominant role of the majority class and increase the attention to the minority class samples. The calculation formula is as follows:

$$\mathcal{L}_{CB} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \left(\frac{1-\gamma}{1-\gamma^c} \right) \bullet y_{ic} \log(p_{ic}) \quad (16)$$

where N is the number of samples, C is the number of categories, y_{ic} is the true category label, p_{ic} is the predicted category label, and $\frac{1-\gamma}{1-\gamma^c}$ is the category balance factor.

4.6. Accuracy evaluation of test

To objectively and rigorously evaluate the model's performance, we use manually annotated urban geo-scene (UGS) ground truth data across four study areas. The ground truth was established through expert visual interpretation and annotation, using Amap POI data and street view imagery as references.

Considering the diversity and unevenness of urban geo-scene categories, we adopt Micro-F1 score (MIF1) and Macro-F1 score (MAF1) to evaluate the classification results. MIF1 represents the proportion of correctly classified images in all images, which can evaluate the overall

uneven distribution of categories. MAF1 is the harmonic mean of precision and recall, emphasizing the balanced performance of each category. The calculation methods of these indicators are as follows:

$$\text{Micro-Precision} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)} \quad (17)$$

$$\text{Micro-Recall} = \frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)} \quad (18)$$

$$\text{Micro-F1}_i = 2 \bullet \frac{\text{Micro-Precision} \bullet \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}} \quad (19)$$

where N is the types of classification, TP_i , TN_i , FP_i , and FN_i denote the numbers of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classification results, respectively.

Recall represents the proportion of correctly predicted samples in the total number of actual positive samples; precision indicates the rate of the predicted positive samples that are positive; Intersection over Union (IoU) is the ratio of the true positive predictions to the union of the predicted and actual positives.

$$\text{IoU}_i = \frac{TP_i}{TP_i + FP_i + FN_i} \quad (20)$$

$$\text{Macro-F1}_i = 2 \bullet \frac{\sum_i^n \frac{TN_i}{TN_i + FN_i} \bullet \sum_i^n \frac{TN_i}{TN_i + FN_i}}{n \left(\sum_i^n \frac{TN_i}{TN_i + FN_i} + \sum_i^n \frac{TN_i}{TN_i + FN_i} \right)} \quad (21)$$

The combination of the two not only considers the overall effect but also considers the performance of multiple categories and a few analogies, reflecting the test results more scientifically and objectively.

5. Experiments and results

The analysis was performed on a Windows 10 operating system using a CPU (3.4 GHz core i7-6700) and RAM (16 GB). Pytorch in Python 3.8 is the deep learning framework, which is accelerated by a GPU (NVIDIA RTX-A4000 16G). The calculation process memory and efficiency index refer to the Appendix (Table E).

5.1. Sample and training parameters

The construction of samples plays a crucial role in recognizing urban geo-scenes. In this experiment, we define samples based on the blocks as the fundamental units. To ensure the representativeness of the samples and maintain an appropriate and balanced number of samples for each category, we manually selected 950 block samples from Beijing as the initial dataset (Table 2). To ensure annotation quality and reduce labeling uncertainty, we adopted a multi-stage strategy involving expert visual interpretation, independent cross-validation, and semantic consistency checks using POI and land cover overlays.

Given the sample size and category differences, we apply data augmentation techniques—including rotation, inversion, and cropping—to enhance the model's generalization ability and mitigate overfitting (Table 2). To enhance training robustness, 80 % of the augmented samples are randomly assigned, with 80 % used for training and 20 % for validation. After model convergence and hyperparameter tuning, the validation set is incorporated into the training set for final model retraining.

We conducted independent testing on all blocks within the four study areas. For Beijing, the test dataset consisted of all remaining blocks after excluding the initial training and validation samples. For the other three cities, all available blocks were used as test sets, representing the data to be classified in our evaluation.

The patch size of the sample is set to 256 * 256. For the Deeper Features Catcher branch, we employed DenseNet-121 as the backbone

Table 2
Number of samples in UGS of the Beijing datasets. (Training: To train the model; Validation set: To validate the model; Test set (data to be classified): To classify.).

Category	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM
Original Samples	300	50	150	100	30	30	30	50	50	50	50	50	30
Augmentation Samples (Rotation, Flipping, Cropping)	900	200	500	300	100	100	100	200	250	100	200	100	100
Training	576	128	320	192	64	64	64	128	160	64	128	64	64
Validation	144	32	80	48	16	16	16	32	40	16	32	16	16
Test	2304	331	1034	516	113	79	92	239	646	131	376	282	100

for feature extraction. Training hyperparameters were uniformly configured with a learning rate of 0.0005, 400 training iterations, and a batch size of 32.

5.2. Performance of UGS understanding and recognition

5.2.1. Performance of USG model training (Beijing Datasets)

Fig. 8 shows the variations in the accuracy and loss values of the proposed classification model during the 400 training epochs. After training for 300 epochs, the curves of the training accuracy, validation accuracy, training loss, and validation loss exhibited no significant changes, indicating that the model gradually converged and reached a stable state. By the end of training (epoch 400), the model achieved a training accuracy of 94.73% and a validation accuracy of 82.94%. Subsequently, the test dataset was used to verify the generalizability of the trained model, obtaining 93.10% (Macro-F1) and 90.43% (Micro-F1) of classification accuracy (Table 3). These results confirm the effectiveness and strong generalizability of the proposed framework in urban geo-scene (UGS) functional classification.

5.2.2. Accuracy evaluation of model test results (Beijing Datasets)

Table 3 presents the confusion matrix and evaluation metrics for the classification results of the Beijing test dataset, covering 13 functional urban geo-scene (UGS) categories. The F1 scores for all functional types exceed 85 %, with Residential Zones, Shantytowns, and Green Spaces & Parks achieving F1 scores above 95 %, indicating relatively high classification accuracy. Notably, Green Spaces & Parks (GP) and Shantytowns (ST) exhibit the best classification performance, with F1 scores of 96.82 % and 96.30 %, respectively.

However, some misclassifications occurred between Residential, Commercial, and Business & Office categories, which exhibit strong inter-category confusion (Figure A: thicker lines (Appendix)). This is largely due to mixed land-use scenarios, such as commercial housing along streets or company-residential hybrid zones. Similarly, schools and kindergartens often reside within residential blocks, adding to classification ambiguity.

Intersection-over-Union (IoU) was further used to evaluate the semantic clarity and functional distinction of UGS categories. Consistent

with the F1 score results, Residential Zones, Shantytowns, and Green Spaces achieved high IoU scores. In contrast, Industrial and Healthcare areas show lower IoU values, which may be attributed to limited sample sizes and greater functional heterogeneity within these categories.

As illustrated in Fig. 1, the superior classification performance of Shantytowns stems from their consistent physical texture (typically small and rectangular), distinct spectral signatures, and spatial clustering. Likewise, Green Spaces & Parks are characterized by low building density and simple spatial structures, contributing to their distinguishability. Conversely, BO is more frequently misclassified as a commercial zone due to its high spectral and social functional homogeneity, as well as its relatively complex spatial relationships, resulting in lower classification accuracy.

As shown in Fig. 6, the UGS understanding and recognition results of the Beijing study area are shown. In addition, the figure also contains two zoomed-in scenes, presenting more detailed spatial information of urban functional distribution from different spatial perspectives. Various building characteristics, such as building density and height, can be observed in different functional UGSs, especially in the ER and ST categories.

5.2.3. Accuracy evaluation of test results on the other three areas

We analyzed the generalization ability of the proposed UGS recognition model by inputting three cities of different economies and scales (Xi'an, Shenzhen, and Urumqi) into the model trained on Beijing. We evaluated the recognition results by comparing the classification results with ground truth data (confusion matrix). In addition, to intuitively illustrate the recognition effect of the model (Fig. F (appendix)), Fig. 9 (C) visualizes the difference between the results and the corresponding ground truth.

As shown in Tables B, C, D and Fig. E (Appendix), the Urumqi area, located in the remote western region, has a poor urban scale and economy, and the number and distribution of urban geo-scene categories are uneven, so the recognition accuracy is the lowest among the three cities, at only 90.73 % (MIF1). In contrast, Shenzhen, whose economic status and urban scale are like Beijing, has a higher overall accuracy and is like Beijing.

The reason for this slight difference may be the uncertainty of the

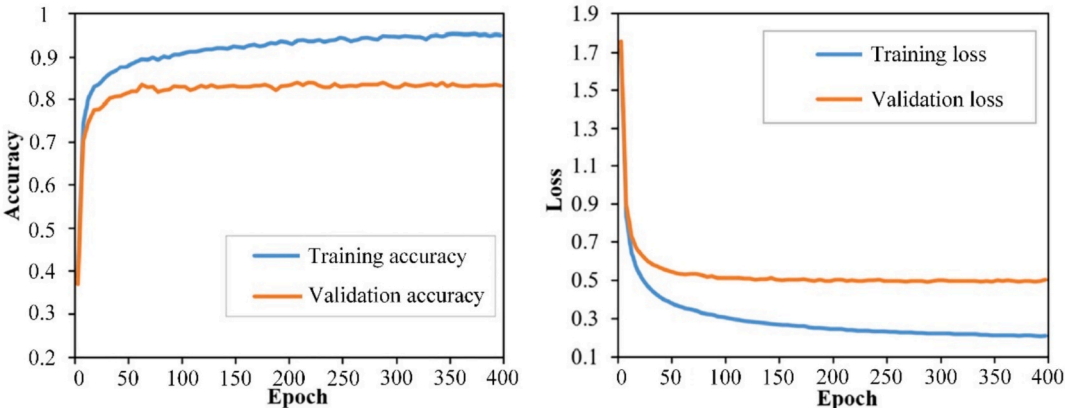


Fig. 8. Variation in training accuracy, validation accuracy, training loss, and validation loss of the proposed model in the training process.

Table 3
Quantitative results for the Beijing datasets.

Label	Number													P (%)	R (%)	F1 (%)	IoU (%)
	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM				
R	2214	2	25	17	1	1	0	5	13	3	12	1	1	96.09	96.47	96.28	92.83
ST	3	320	2	1	1	0	0	0	0	0	2	0	1	96.68	96.97	96.82	93.84
C	24	2	957	24	2	0	0	4	9	2	4	1	1	92.55	92.91	92.73	86.45
BO	17	1	28	455	3	0	0	5	5	1	4	0	1	88.18	87.50	87.84	78.31
I	2	0	4	2	99	3	3	0	0	0	5	0	0	87.61	83.90	85.71	75.00
U	4	2	0	0	3	70	1	0	1	0	0	2	1	88.61	83.33	85.89	75.27
T	0	2	0	1	2	1	83	0	0	0	4	0	0	90.22	89.25	89.73	81.37
GA	8	0	5	4	0	0	0	214	4	2	9	1	2	89.54	85.94	87.70	78.10
ER	12	0	3	2	0	0	0	3	589	6	3	2	0	92.46	95.00	93.72	88.17
H	5	0	2	0	0	0	0	1	6	113	3	1	0	86.26	86.26	86.26	75.84
P	12	1	7	8	2	1	5	4	6	4	328	0	3	87.23	86.09	86.66	76.46
GP	2	0	1	0	0	3	0	1	3	0	1	273	1	96.81	95.79	96.30	92.86
PM	1	1	0	2	0	0	0	2	1	0	1	1	89	89.00	90.82	89.90	81.65

sample. Our model is only trained based on the Beijing dataset, and the differences in data from different research areas (such as spatiotemporal differences and the number of categories) are not fully considered. However, the Micro-F1 scores across all three transfer cities remained close to 90 %, suggesting that the proposed framework exhibits strong cross-city adaptability and robustness. This demonstrates that dissecting urban fabric—by explicitly modeling spatial interactions and functional composition—can effectively compensate for the limitations of visual modality alone in UGS recognition, as further discussed in Section 6.2.

Moreover, the framework maintained high classification accuracy even in semantically complex categories such as Commercial Zones, Business & Office areas, and Public Service Facilities, where visual patterns are often ambiguous. For structurally distinct classes such as Shantytowns, Urban Green Spaces, and Parks, the model yielded even higher accuracy. These results further highlight the pivotal role of the urban fabric concept in enabling precise identification of complex urban geo-scene systems.

6. Discussion

Compared to traditional, idealized models of urban form (e.g., concentric, sectoral, or scattered patterns), the urban fabric offers a more refined scale-space paradigm. It effectively captures micro-level urban heterogeneity and overcomes the limitations of conventional scene classification that often relies solely on surface-level visual features. By integrating functional roles and spatial relationships (e.g., co-location, connectivity, distribution), this approach advances a deeper, context-aware understanding of urban systems.

Grounded in geographic principles and dissection-based thinking, the urban fabric concept draws on ideas such as Tobler’s First Law and location-based interaction. It enables a semantically enriched interpretation of urban geo-scenes (UGS), facilitating more accurate recognition and spatial reasoning.

6.1. Effectiveness of UGS recognition framework

6.1.1. Comparison with other methods

Tables 4 and 5 summarize the performance of various models and fusion strategies across four representative urban regions. Traditional machine learning approaches based on handcrafted visual features (e.g., Random Forests) performed poorly, whereas deep learning models demonstrated significantly better results. Among existing models, Vision Transformer (ViT) achieved a strong accuracy of 83.79 % in the Beijing (Micro-F1). SE-DenseNet, with its ability to capture local features, enabled a more fine-grained understanding of urban components, achieving an accuracy (Micro-F1) of 84.41 %.

For spatial modeling, GraphSAGE and GAT outperformed GCN but still underperformed RF. However, hybrid models such as RF + GAT and

DenseNet-121 + GAT produced more promising results. Notably, the combination of ViT and GAT further improved accuracy to 88.62 % in Beijing-MIF1. Our proposed recognition framework significantly outperformed all other approaches, achieving an accuracy of 93.10 % in the same research area. These results highlight that for spatially complex (UGS), the concept of urban fabric—integrating visual features, functional information, and spatial semantics—offers a more robust and accurate recognition capability.

6.1.2. Comparison with existing products

To evaluate the effectiveness of our proposed framework in capturing fine-scale urban functional structure from the perspective of urban fabric, we conducted comparative evaluations with three representative existing land-use products: OSM, EULUC-China and EULUC-China 2.0(Gong et al., 2020; Li et al., 2025).

As shown in Fig. 7, our results largely overlap with OSM, but exhibit better spatial coherence and semantic completeness. This is primarily because OSM relies heavily on crowd-sourced POI tags—urban blocks without social-function labels are often left blank. To ensure fairness, only blocks with verified OSM labels were included in the quantitative evaluation, and those with missing or ambiguous information were excluded, thus mitigating bias due to inconsistent data coverage (see Section 6.4 for further discussion on limitations).

Although EULUC-China is a widely used national-scale land-use dataset, it suffers from limited spatial resolution and insufficient categorical granularity. In contrast, our proposed framework offers advantages in fine-grained classification, benefiting from high-resolution input data that allows detailed characterization of the urban fabric at the intra-block level. To mitigate potential biases introduced by temporal misalignment, we verified the consistency of relatively stable functional categories (e.g., residential areas, hospitals, and parks).

Fig. 10. (A, B) presents visual comparisons in the Beijing area. While all products capture major land-use categories, EULUC-China 2.0 performs relatively well at finer scales. However, our proposed framework exhibits a higher degree of differentiation by additionally identifying complex and nuanced urban elements, including shantytowns, cultural landmarks (e.g., museums), and scenic zones (e.g., the Forbidden City). In addition, the refinement and optimization of the road network yield more coherent delineations and more consistent recognition outcomes that better align with the urban fabric concept. These findings highlight the model’s enhanced capability to capture both functional heterogeneity and morphological complexity, which is essential for fine-grained Urban Geo-Scene (UGS) classification and comprehensive urban fabric analysis.

Moreover, Table 6 reveals a consistent issue in baseline datasets: a negative correlation between classification accuracy and urban development level. For instance, EULUC-China shows significantly lower accuracy in less-developed cities such as Xi’an (65.23 %) and Urumqi

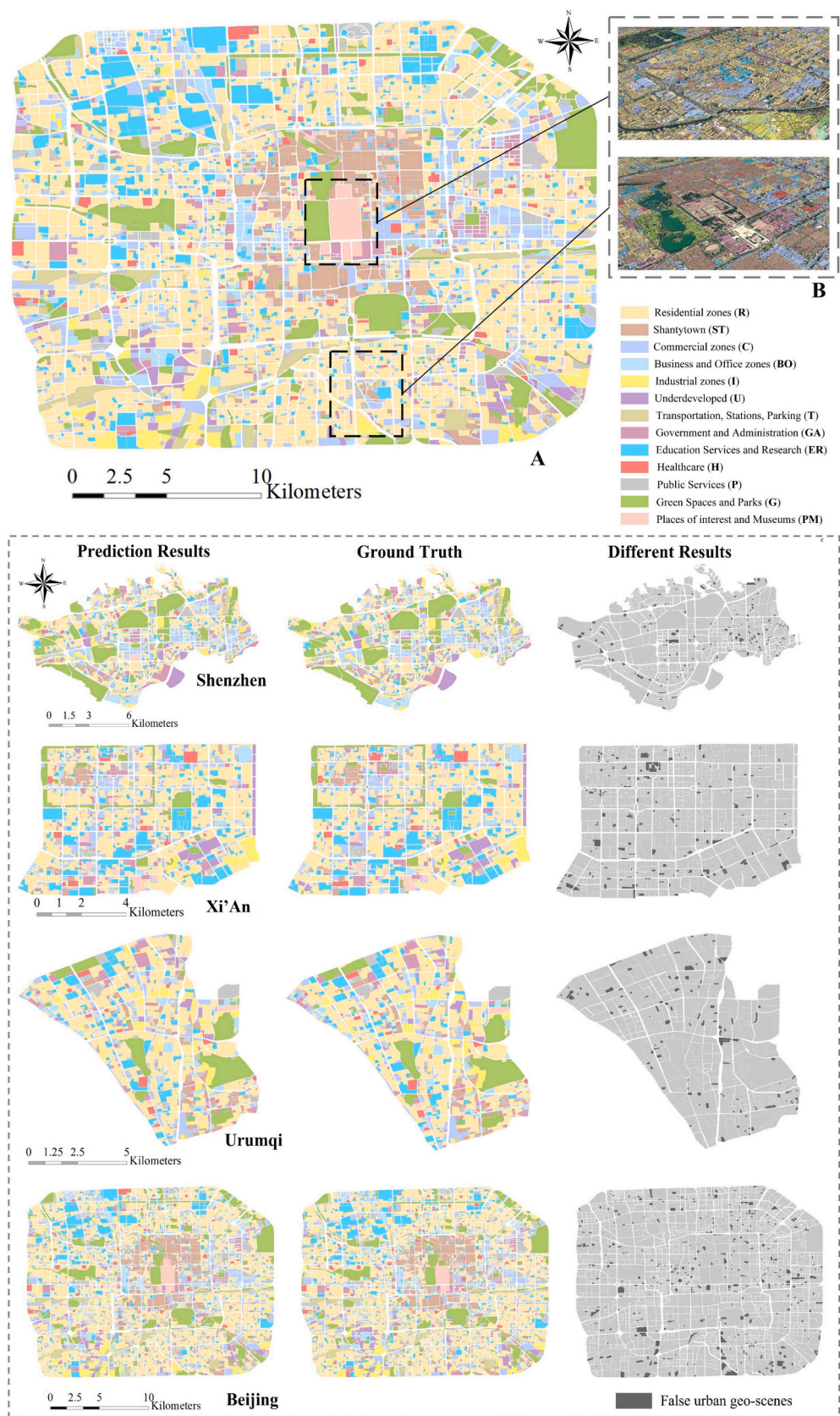


Fig. 9. Urban Geo-scene understanding and recognition mapping. (A: UGS results; B: 3D detail map; C: Urban Geo-scene recognition mapping with ground truth.)

Table 4
The MIF1 and MAF1 accuracy of the different single methods. **RF**: Random Forest Model (RSI + POI); **VGG-16**: Visual Geometry Group Model (RSI + POI); **ResNet**: Residual Neural Network (RSI + POI); **DenseNet-121**: Dense Convolutional Network (RSI + POI); **ViT**: Vision Transformer (RSI + POI); **GCN**: Graph Convolutional Network (Multimodal fusion); **GAT**: Graph Attention Network (Multimodal fusion); **GraphSAGE**: Graph Sample and aggregate Network (Multimodal fusion).

Study Area	Methods (MIF1 %and MAF1%)									
	RF	VGG-16	ResNet-50	DenseNet-121	ViT	SE- DenseNet	GCN	GraphSAGE	GAT	EA-GAT
Beijing	74.45	75.75	77.60	80.44	80.66	82.39	<u>71.21</u>	71.35	71.94	76.45
	76.87	79.12	81.31	83.36	83.79	85.41	<u>74.33</u>	74.19	75.31	79.49
Shenzhen	74.02	76.43	78.02	80.51	81.22	83.15	<u>71.28</u>	72.25	72.10	76.88
	76.88	78.72	80.89	83.45	84.30	86.20	<u>73.82</u>	75.26	75.66	79.81
Xi'an	72.13	74.55	76.75	80.05	79.87	81.33	<u>69.45</u>	70.52	71.43	75.28
	74.91	77.38	80.24	82.95	83.01	84.45	<u>72.22</u>	73.61	74.40	79.52
Urumqi	71.63	73.37	76.26	79.45	79.64	81.41	<u>69.39</u>	71.09	72.08	75.43
	73.12	77.02	79.73	82.63	83.06	84.27	<u>72.31</u>	73.93	74.89	79.56

Table 5
The MIF1 and MAF1 accuracy of the different multi-methods.

Study Area	Methods (MIF1 %and MAF1%)				
	RF + GAT	DenseNet + GCN	DenseNet + GAT	ViT + GAT	Our
Beijing	<u>79.03</u>	82.88	84.55	85.67	90.43
	<u>82.26</u>	85.44	87.59	88.62	93.10
Shenzhen	<u>79.41</u>	83.11	84.25	84.85	90.10
	<u>83.23</u>	86.23	88.52	89.02	92.74
Xi'an	<u>77.55</u>	82.32	84.23	84.19	88.21
	<u>81.38</u>	86.03	88.37	88.23	91.56
Urumqi	<u>76.96</u>	81.52	83.74	84.07	86.80
	<u>80.89</u>	85.34	87.62	87.92	90.73

(55.23 %), even falling behind OSM in some cases. This degradation is largely attributable to coarse road network delineation, which affects parcel segmentation quality in these areas.

By contrast, our multimodal framework maintains more consistent performance across cities of different sizes and development stages. Through improved structural modeling and semantic integration, it reduces inter-city performance disparities and enables urban fabric-level recognition with high generalizability, offering practical value for urban planning and spatial decision-making.

6.2. The effectiveness of the urban fabric

To further analyze the effectiveness and robustness of the urban fabric, we conducted a series of ablation experiments to assess the contribution of different components and design strategies. As shown in Table 7, the comparison between Experiments 1 to 4 highlights the advantages of the SE attention mechanism. Comparing Experiments 1 and 2 with Experiments 3 and 4 reveals that tuning parameters can yield marginal improvements in accuracy. However, due to the limited sample size, the gains are not substantial. The Experiment 11 once again emphasizes the importance of combining visual features with spatial semantics, and the core of the urban fabric is the weaving of geographic objects and spatial relationships. Experiments 5 to 10 show the effectiveness incorporating spatial co-location patterns and edge features into the model. Notably, Experiments 6, 7, 8, and 10 demonstrate that the edge attention mechanism—incorporating both distance and angular information—outperforms a simple linear distance weighted to a certain extent. Similarly, Experiments 11 to 14 perform minor operations on spatial relationships based on visual features to express the urban fabric as realistically as possible.

Therefore, at the heart of the urban fabric lies a complex tapestry of

geo-objects and spatial semantics that collectively define the urban geo-scene’s essence and rhythm. By dissecting the components and interconnections of the urban fabric, we can restore and digitize the urban pattern as much as possible and mine the latent spatial semantics to achieve refined UGS recognition and understanding.

To further enhance interpretability, we employ the Grad-CAM technique to visualize class-specific attention with feature maps. By back-propagating the gradients of a given prediction, Grad-CAM highlights spatial regions that most strongly influence the model’s decision, with higher responses shown as red clusters (Appendix Fig. H). Lower-level convolutional layers capture fine-grained texture and shape information, while deeper layers progressively encode abstract semantic structures.

When applied to the Beijing dataset, the proposed framework demonstrates superior delineation of functional boundaries compared to baseline models. In educational areas, attention is focused on playgrounds; in residential zones, the model captures uniform spatial patterns; and in transport hubs, road-related features are emphasized. Appendix H further demonstrates the distinct responses across model branches, illustrating the complementary roles of visual, semantic, and spatial cues in UGS classification.

6.3. Ablation experiments of different data sources

An ablation study was conducted to analyze the impact of different data sources on UGS recognition. The multimodal spatial semantic fusion model was trained using various data source combinations, and its performance on different and independent research areas is presented in Table 8.

From a single data point of view, RSI contributes the most in the four study areas, and the model of Beijing POI alone is the worst. This shows that RSI in different cities has similar visual features, which can realize the recognition of UGS to a certain extent. In contrast, the spatial distribution patterns of POI and building data vary significantly across cities and are highly sensitive to economic development levels, leading to inconsistent classification performance.

Integrating multiple data sources notably enhances the model’s capability to capture UGS characteristics. For example, experiments 1, 5, 8, and 11 demonstrate that multimodal combinations improve performance by jointly modeling physical structures, social-functional semantics, and spatial relationships. These results indirectly illustrates that urban fabric-oriented modeling reduces inter-city variability, compensating for disparities in urban scale and socioeconomic conditions.

Further analysis of data source contributions reveals that RSI plays a

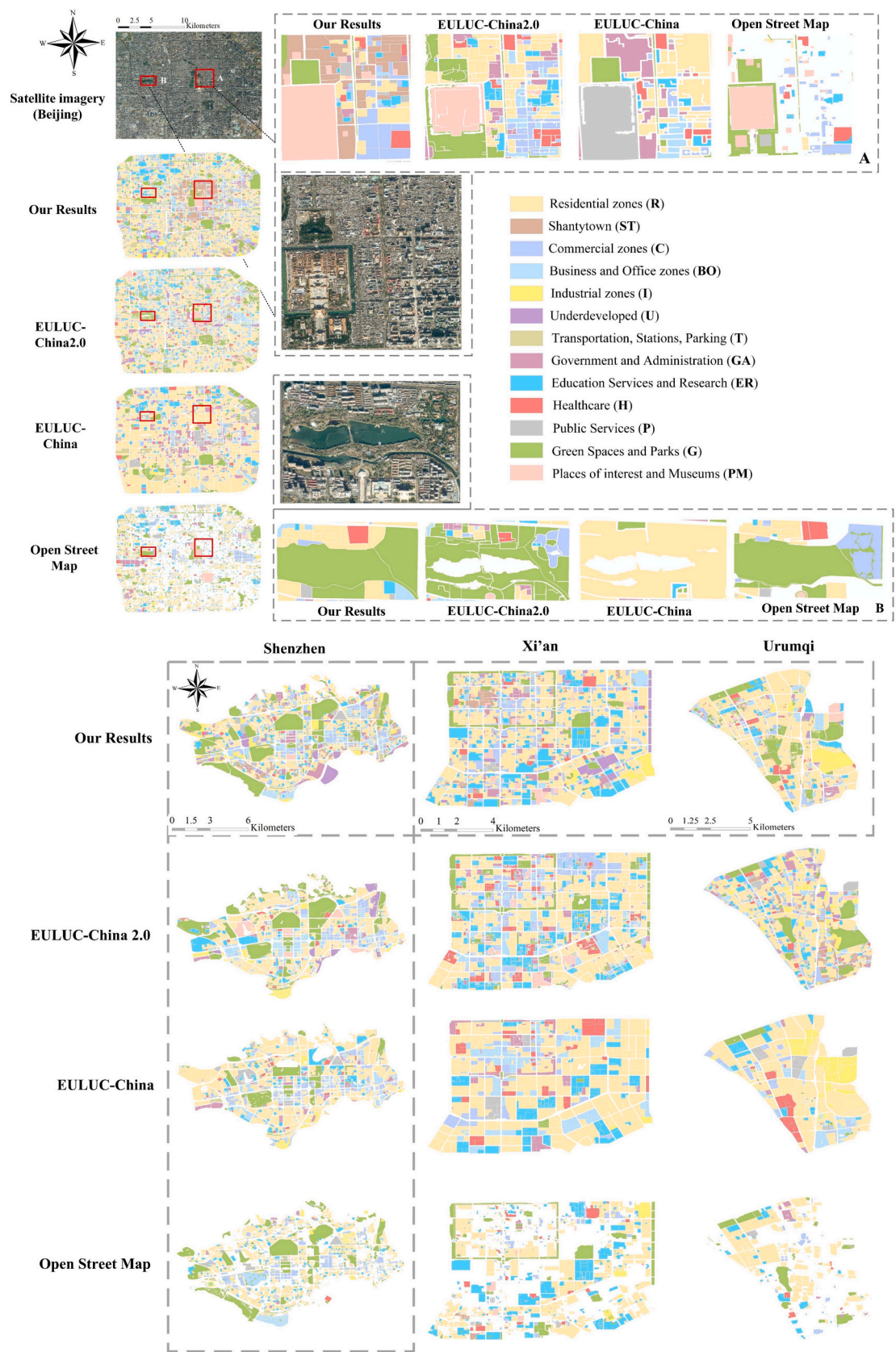


Fig. 10. UGS display and comparison of different products. (A and B: detailed maps of the study area Beijing).

Table 6
The MAF1 and MIF1 accuracy of the different products.

Products	MAF1(%)				MIF1(%)			
	Beijing		Shenzhen		Xi'an		Urumqi	
OSM(no blank)	<u>65.46</u>	<u>65.35</u>	<u>65.33</u>	<u>65.42</u>	<u>63.07</u>	<u>66.12</u>	62.34	62.23
EULUC-China 2.0	85.63	88.59	86.37	89.30	83.19	86.11	81.95	85.04
Our	90.44	93.11	90.10	92.74	88.21	91.56	86.80	90.73

Table 7
Urban Fabric Effectiveness Experiment (Beijing datasets). **DenseNet-121**: 121-layer Dense Convolutional Network; **DenseNet-169**: 169-layer Dense Convolutional Network; **SE**: Squeeze-and-Excitation module (r = Squeeze ratio); **GAT**: Graph Attention Network; **SCP**: Spatial co-location pattern; **DW**: Distance-weighted; **EA**: Edge Attention Model.

Exp.	Strategy								MAF1(%)	MIF1(%)
	DenseNet-121	DenseNet-169	SE ($r = 16$)	SE ($r = 8$)	GAT	SCP	DW	EA		
1	✓								81.44	84.36
2		✓							81.61	84.45
3	✓		✓						82.39	85.41
4	✓			✓					82.41	85.55
5					✓				<u>71.34</u>	<u>74.33</u>
6					✓		✓		72.05	75.26
7					✓			✓	72.78	76.09
8					✓	✓			74.12	77.59
9					✓	✓	✓		75.86	78.52
10					✓	✓		✓	76.45	79.49
11	✓		✓		✓				85.97	89.01
12	✓		✓		✓			✓	87.48	90.35
13	✓		✓		✓	✓			88.37	91.56
14	✓		✓		✓	✓		✓	90.43	93.10

Table 8
Demonstration of evaluation metrics for different data source combinations.

Exp.	Data Category	MAF1(%)				MIF1(%)			
		Beijing		Shenzhen		Xi'an		Urumqi	
1	Building	<u>78.90</u>	<u>78.08</u>	<u>78.23</u>	<u>77.43</u>	<u>78.32</u>	<u>78.5</u>	<u>77.95</u>	<u>78.13</u>
2	RSI	81.17	82.37	80.52	81.7	79.59	80.79	77.2	78.42
3	POI	66.44	66.45	66.78	66.8	64.85	64.87	64.49	64.5
4	Building + RSI	82.54	85.81	81.84	85.12	80.97	84.23	78.59	81.86
5	Building + High	79.89	83.76	79.25	83.09	78.31	82.18	75.92	79.81
6	Building + POI	80.55	83.78	79.9	83.1	78.97	82.2	76.6	79.83
7	RSI + POI	84.19	87.94	83.52	87.27	82.61	86.35	80.24	83.99
8	Building + RSI + High	86.55	89.68	85.89	89.02	84.97	88.1	82.6	85.71
9	Building + RSI + POI	87.39	89.67	86.72	89	85.79	88.09	83.44	85.68
10	Building + High + POI	81.67	85.74	81.03	85.09	80.09	84.17	77.69	81.79
11	Building + RSI + High + POI	90.44	93.11	90.1	92.47	88.21	91.56	86.8	90.73

dominant role in UGS recognition, particularly for more visually distinguishable urban types. For example, RSI had the highest score for MIF1 in a single classification (Exp.1, 2, 3). Similarly, in terms of these three indicators, the scores of Exp.5, 6 were lower than those of experiment 8, 9 using RSI; however, it is interesting that the classification performance of experiment 10 did not exceed that of experiment 4 or experiment 7 in all scores. In addition, the accuracy of Buildings' MIF1 is slightly lower than that of MAF1. This is because MAF1 considers relatively rare categories (shanty towns and green spaces, parks) and ensures the performance of small samples.

Therefore, the similarities of urban fabrics are all patterns that are displayed by the organic composition of the relationship between urban elements and space. Dissecting urban fabrics can weaken the limitations brought by different cities to a certain extent.

6.4. Diversity degree analysis

With the acceleration of urbanization, the emergence of mixed-use and functionally diverse urban spaces has become increasingly crucial for accommodating a wide range of human activities, enhancing spatial

efficiency, and improving urban resilience. Residential areas frequently adjoin schools and commercial centers, while business districts evolve in tandem with service-oriented infrastructures. These intricate land-use patterns reflect the interwoven nature of the urban fabric and underscore the need for robust analytical tools to quantify and interpret heterogeneity within urban geo-scenes (UGSs).

To address this challenge, we introduce a dual-perspective diversity assessment method, designed to quantify functional and structural complexity within each UGS. Leveraging multi-source geospatial data—including points of interest (POIs) and building attributes—we compute information entropy to characterize both functional semantic diversity and physical structural diversity. This diversity analysis is seamlessly integrated into our UGS classification framework and serves as an auxiliary metric to assess the model's capacity to recognize intra-block heterogeneity. Specifically, POIs reflect the spatial semantics of the distribution intensity (e.g., commerce, education, and healthcare), which are spatially mapped onto building footprints. Furthermore, a building density factor is introduced to account for morphological diversity among urban blocks. These two components form a multimodal diversity profile for each UGS, supporting a more comprehensive

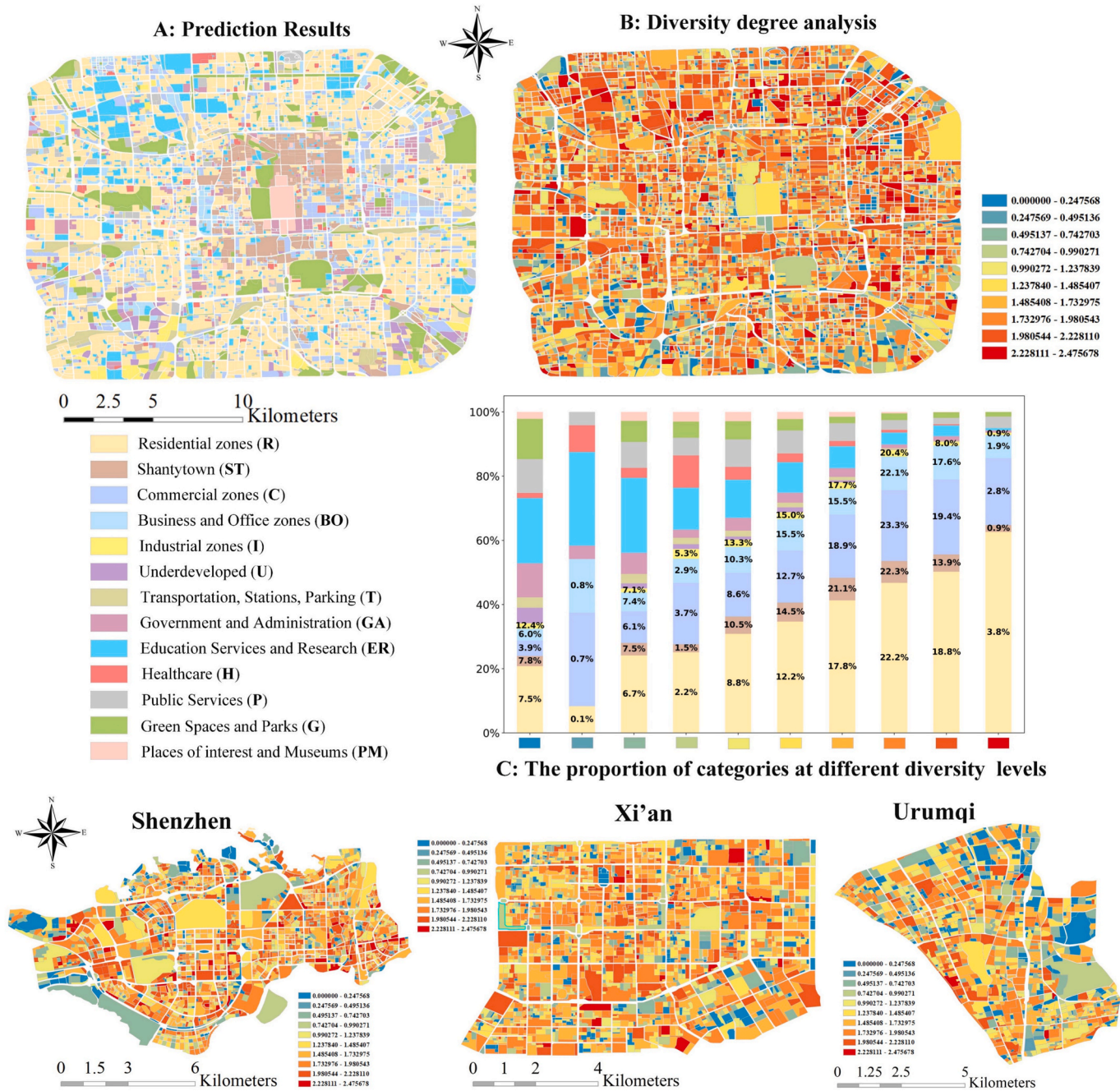


Fig. 11. Diversity degree results and analysis for Beijing.

understanding of urban complexity.

$$P_{ij} = n_{ij} / N_j \quad (22)$$

$$DD_j = -\omega_i \sum_{i=1}^k P_{ij} \log(P_{ij}) \quad (23)$$

where n_{ij} represents the number of the i -th category in the j -th UGS, N_j represents the total number in the j -th UGS, DD_j represents the urban diversity degree of the j -th UGS, $\omega_i = 1 + \lambda \cdot \text{Dens}_j$ is building density adjustment factor, and k represents the number of categories.

Fig. 11 visualizes the spatial distribution of UGS diversity across the Beijing study area, with a cold-to-warm color scale to represent entropy values. Low-entropy UGSs (cool tones) typically correspond to mono-functional zones, while high-entropy UGSs (warm tones) indicate complex, mixed-use environments. This visualization is empirically

supported by real-world observations. For example, historic Hutong neighborhoods—known for their dense and informal morphologies—exhibit high entropy due to the coexistence of residences, markets, and services. In contrast, the Forbidden City and large urban parks appear in cooler tones, reflecting their spatial uniformity and singular functions.

To further interpret these results, we analyzed the distribution of land-use types across different entropy intervals. Certain categories were found to be overrepresented in high-diversity UGSs: residential zones constitute 62.5 %, commercial zones 64.4 %, business/office districts 57.2 %, and shantytown areas 58.1 % of high-entropy areas. These functionally complex areas are traditionally difficult to identify using unimodal approaches due to their high internal variability. As further illustrated in Appendix Fig. D, categories with greater functional overlap are connected by thicker inter-category links, highlighting the “semantic

entanglement” that characterizes these urban blocks.

The strong performance of our multimodal in high-entropy areas (see Table 3) demonstrates its effectiveness in capturing this complexity. By integrating visual features, functional semantics, and spatial relationships, the model successfully encodes the structural and functional intricacies of urban environments. The diversity analysis thus provides a complementary validation mechanism, directly linked to the model’s classification output. It confirms that the framework can reliably differentiate between simple and complex urban blocks, enabling a fine-grained interpretation of urban form.

Ultimately, incorporating diversity assessment into UGS modeling not only strengthens the reliability of classification outcomes but also enhances their interpretability for urban planning applications. Understanding the functional heterogeneity of the urban fabric is essential for guiding resilient, integrated, and context-sensitive urban development strategies.

6.5. Limitations and future work

Despite the promising results, this study still has several limitations that warrant further exploration. First, there is continued debate over the appropriate unit and scale for Urban Geo-Scene (UGS) analysis, particularly due to the diversity of urban morphologies and development levels. Blocks, homogeneous grids, and superpixel-based objects have all been used as fundamental units in previous studies (Wen et al., 2020). However, the choice and definition of scale significantly affect the capture of urban fabric and may lead to ambiguity or inconsistency in UGS classification within units. Therefore, determining the optimal scale and unit of analysis for different cities remains a challenging task.

While this study focuses on block-level urban scene understanding, we acknowledge the value of incorporating finer-grained building-level data—such as footprints, floor area, and occupancy types—to more accurately capture intra-parcel heterogeneity (He et al., 2025). Future work will explore building-scale analysis to enable a more detailed interpretation of urban structure and function, particularly in compact cities and mixed-use developments.

Secondly, urban morphology and spatial patterns exhibit significant variability not only across cities, but also between broader geopolitical contexts—particularly between the Global East and Global West. These differences are shaped by a complex interplay of topographical conditions, climatic zones, policy regimes, economic structures, and cultural norms. To enhance the generalizability of our framework, future research will aim to expand the spatial coverage to include larger and more diverse regions. In parallel, the integration of additional environmental variables and socio-perceptual data sources (e.g., surveys, street-level imagery, or mobility patterns) is essential for deepening the interpretation and contextual understanding of Urban Geo-Scenes (UGSs).

Lastly, regarding the model, improvements are still needed in terms of transferability and generalizability. As shown in Figure G (Appendix Fig. G), limitations remain in sample quantity and quality (label noise, annotation uncertainty), as well as in feature extraction (Visual ambiguity and Sparse contextual signals) and data fusion methods (Mixed-use areas). In the future, we will explore training with multi-scale image pairs (high, medium, and low resolution) and incorporating GAN-generated data to simulate fine-grained urban features. Additionally, we consider few-shot learning frameworks and the integration of large vision-language models as promising avenues to reduce reliance on dense annotations and enhance semantic reasoning in heterogeneous

urban environments.

7. Conclusion

Dissecting the urban fabric enables a better understanding of urban geo-scenes from the perspective of the essence of the internal structure of the city and improve recognition performance.

In this study, we propose a multimodal UGS recognition framework grounded in urban fabric graph modeling. By integrating multi-source remote sensing and social sensing data, the framework captures the physical morphology, functional semantics, and spatial relationships of urban elements. Built upon enhanced SE-DenseNet and EA-GAT modules, the model enables deep feature mining and explicit spatial structure reasoning, while the incorporation of spatial co-location patterns ensures more realistic mapping of urban spatial interactions. Extensive experiments across four representative Chinese cities—each differing in scale, development, and morphology—demonstrated the effectiveness and generalizability of the framework, achieving Micro-F1 scores of approximately 90 %. Furthermore, two sets of ablation studies validated the contribution of each component, underscoring the robustness and interpretability of the urban fabric perspective. Diversity analysis further revealed that urban fabric modeling helps mitigate recognition uncertainty arising from spatiotemporal complexity and socioeconomic variability. Overall, this framework provides a scalable and interpretable solution for fine-grained urban analysis, offering theoretical and practical support for urban planning, spatial governance, and sustainable development. In future work, we aim to incorporate additional environmental and perceptual data sources to explore cross-regional and global differences—particularly between eastern and western urban systems.

CRedit authorship contribution statement

Hanqing Bao: Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Data curation, Conceptualization. **Lanyue Zhou:** Writing – review & editing, Validation, Supervision. **Lukas W. Lehnert:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Department of Geography, LMU Munich. We acknowledge the use of SuperView-1 imagery (used under licence), 3D-GloBFP data, OpenStreetMap data (© OpenStreetMap contributors), and EULUC-China (2018) / EULUC-China 2.0 (2022). Computational resources were provided by the Department of Geography, LMU Munich. We thank Wenkai Bao, Jishu Xin, Zhiqi Wang, and Songye Wei, as well as colleagues, for their assistance with sample annotation and for valuable comments on the manuscript. We also thank the anonymous reviewers for their constructive feedback. The views expressed are those of the authors and do not necessarily reflect those of the supporting institution.

Appendix

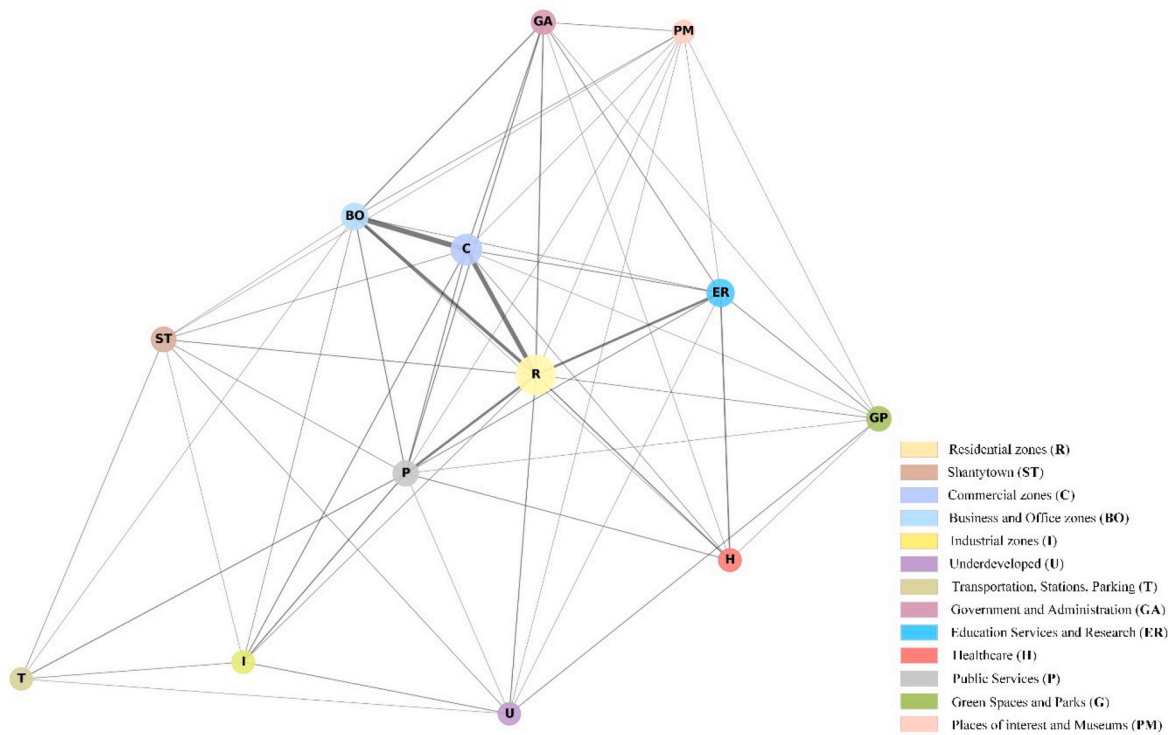


Fig. A. EA-GAT visualization (Beijing classification results).

Table B
Quantitative results for the Shenzhen datasets.

Label	Number														P(%)	R(%)	F1 (%)	IoU (%)
	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM					
R	661	1	3	3	0	1	0	2	2	1	2	1	0	95.66	97.64	96.64	93.49	
ST	2	84	1	1	0	0	1	0	0	0	1	0	0	95.45	93.33	94.38	89.36	
C	5	0	285	15	2	0	0	1	2	0	1	1	0	91.94	91.35	91.64	84.57	
BO	5	0	12	359	2	0	0	2	2	1	2	0	0	90.89	93.25	92.05	85.27	
I	2	0	1	2	68	2	1	0	0	0	1	0	0	89.47	88.31	88.89	80.00	
U	1	1	0	0	2	54	1	0	0	0	1	1	0	91.53	88.52	90.00	81.82	
T	0	1	0	1	0	1	38	0	0	0	2	0	0	88.37	88.37	88.37	79.17	
GA	3	0	2	2	0	0	0	63	2	0	2	1	0	88.73	84.00	86.30	75.90	
ER	3	0	2	2	0	0	0	1	174	1	2	2	0	93.05	93.05	93.05	87.00	
H	2	0	1	1	0	0	0	0	2	28	1	0	0	87.50	80.00	83.58	71.79	
P	5	1	3	8	2	1	2	1	2	1	117	0	0	88.64	81.82	85.09	74.05	
GP	2	0	0	0	0	0	0	1	1	0	0	95	1	93.14	95.00	94.06	88.79	
PM	0	0	0	1	0	0	0	0	0	0	0	1	14	93.33	87.50	90.32	82.35	

Table C
Quantitative results for the Xi'An datasets.

Label	Number														P(%)	R (%)	F1 (%)	IoU (%)
	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM					
R	843	1	15	3	0	2	0	2	5	1	2	1	1	93.56	96.23	94.88	90.26	
ST	3	49	1	1	1	1	0	0	0	0	1	0	0	94.23	85.96	89.91	81.67	
C	16	0	379	5	0	0	0	1	3	0	1	1	0	91.11	93.35	92.21	85.55	
BO	7	0	12	110	0	0	0	2	2	1	0	0	0	89.43	82.09	85.60	74.83	
I	0	0	0	0	15	2	0	0	0	0	1	0	0	83.33	83.33	83.33	71.43	
U	2	1	0	0	1	45	1	0	1	0	0	0	0	86.54	88.24	87.38	77.59	
T	0	0	0	0	0	1	14	0	0	0	1	0	0	87.50	87.50	87.50	77.78	
GA	7	0	2	0	0	0	0	66	2	0	1	0	2	88.00	82.50	85.16	74.16	

(continued on next page)

Table C (continued)

Label	Number													P(%)	R (%)	F1 (%)	IoU (%)
	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM				
ER	9	0	3	3	0	0	0	1	179	2	2	2	1	90.86	88.61	89.72	81.36
H	2	0	1	0	0	0	0	0	2	34	2	0	0	87.18	82.93	85.00	73.91
P	10	1	3	1	1	1	1	1	1	1	87	0	0	88.78	80.56	84.47	73.11
GP	2	0	0	0	0	0	0	2	1	0	0	44	1	89.80	88.00	88.89	80.00
PM	0	0	0	0	0	0	0	0	1	0	0	1	44	89.80	95.65	92.63	86.27

Table D
Quantitative results for the Urumqi datasets.

Label	Number														P(%)	R(%)	F1 (%)	IoU (%)
	R	ST	C	BO	I	U	T	GA	ER	H	P	GP	PM					
R	430	2	4	2	0	1	0	2	2	0	2	1	0	93.68	96.41	95.03	90.53	
ST	2	47	1	1	1	0	0	0	0	0	1	0	0	92.16	88.68	90.38	82.46	
C	5	0	180	4	3	0	0	1	2	1	1	1	0	90.91	90.91	90.91	83.33	
BO	2	0	6	76	2	0	0	2	1	0	0	0	0	89.41	85.39	87.36	77.55	
I	1	0	0	0	48	1	0	0	0	0	1	0	0	85.71	94.12	89.72	81.36	
U	1	1	0	0	1	23	0	0	1	0	0	0	0	88.46	85.19	86.79	76.67	
T	0	0	0	0	0	0	3	0	0	0	1	0	0	75.00	75.00	75.00	60.00	
GA	4	0	2	0	0	0	0	58	2	0	1	0	0	86.57	86.57	86.57	76.32	
ER	3	0	1	1	0	0	0	1	93	1	2	1	1	88.57	89.42	89.00	80.17	
H	2	0	1	0	0	0	0	0	1	23	1	0	0	88.46	82.14	85.19	74.19	
P	6	1	3	1	1	1	1	1	2	1	74	0	0	88.10	80.43	84.09	72.55	
GP	2	0	0	0	0	0	0	2	1	0	0	24	1	88.89	80.00	84.21	72.73	
PM	1	0	0	0	0	0	0	0	0	0	0	0	8	80.00	88.89	84.21	72.73	

Table E
Computational efficiency metrics.

Component	Parameters	FLOPs	Total Inference Time	Memory (Peak)	Notes
SE-DenseNet121 (120000 Samples)	~20 M	~3.5G	~57 min	~2.6 GB	Batch size ~ 64
EA-GAT (About 4000 graphs)	~0.3 M	~0.6G	~13 min	~1.0 GB	5–12 nodes per graph
Multimodal Fusion Module	~2.0 M	~1.3G	~11 min	~1.0 GB	Includes attention + MLP
Total (Full Inference)	22.3 M	~5.4G	~81 min	~5.0 GB	On RTX A4000 (16 GB)

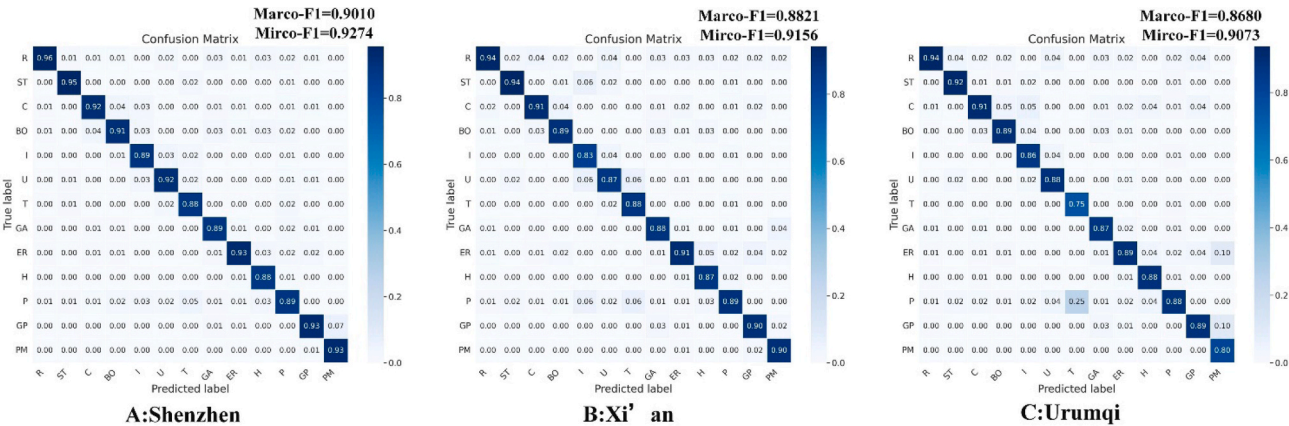


Fig. E. Confusion matrix of the three study areas. (A: Shenzhen; B: Xi'An; C: Urumqi)

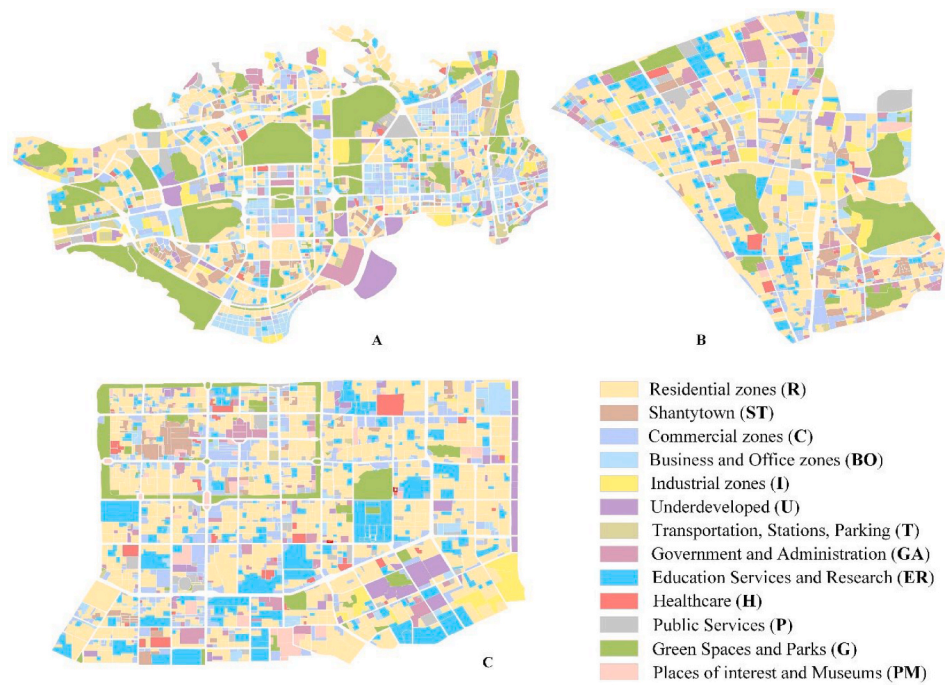


Fig. F. Urban Geo-scene understanding and recognition mapping. (A: Shenzhen; B: Urumqi; C: Xi'an.)

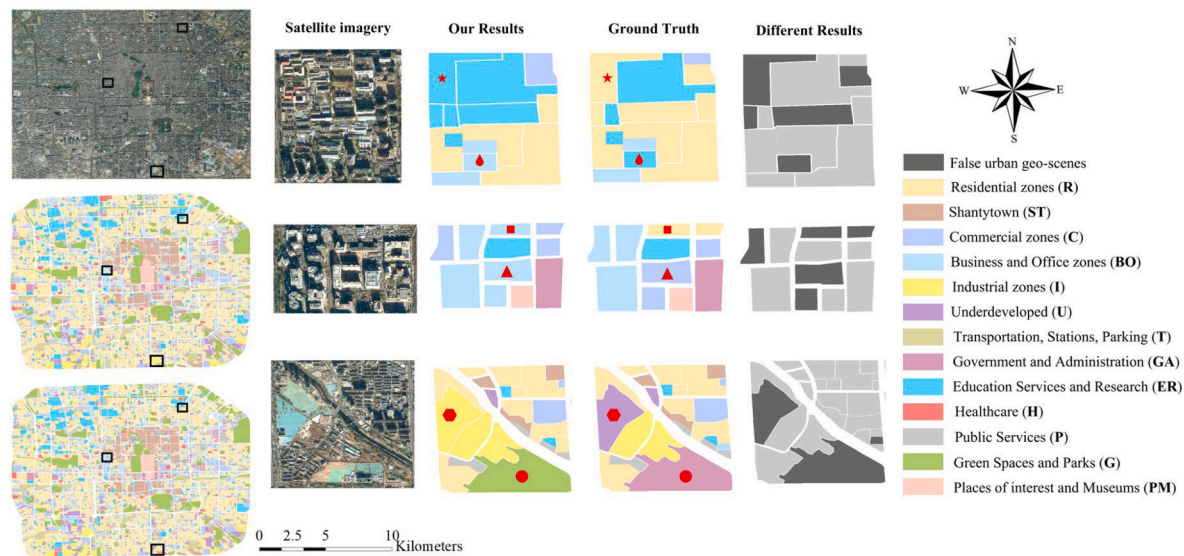


Fig. G. Examples of failure case (Beijing).

- These cases primarily fall into three categories:
1. Mixed-use areas: Parcels containing both residential and commercial elements often lead to misclassification due to blurred functional boundaries. (Square and Triangle).
 2. Visual ambiguity: Some parcels contain visually similar features (e.g., school vs. institutional buildings), making it difficult to distinguish based solely on imagery. (Hexagon and Circle).
 3. Sparse contextual signals: In some low-density or peripheral areas, the lack of strong spatial neighbors reduces the effectiveness of graph-based reasoning. (Five-pointed star and Water drop).

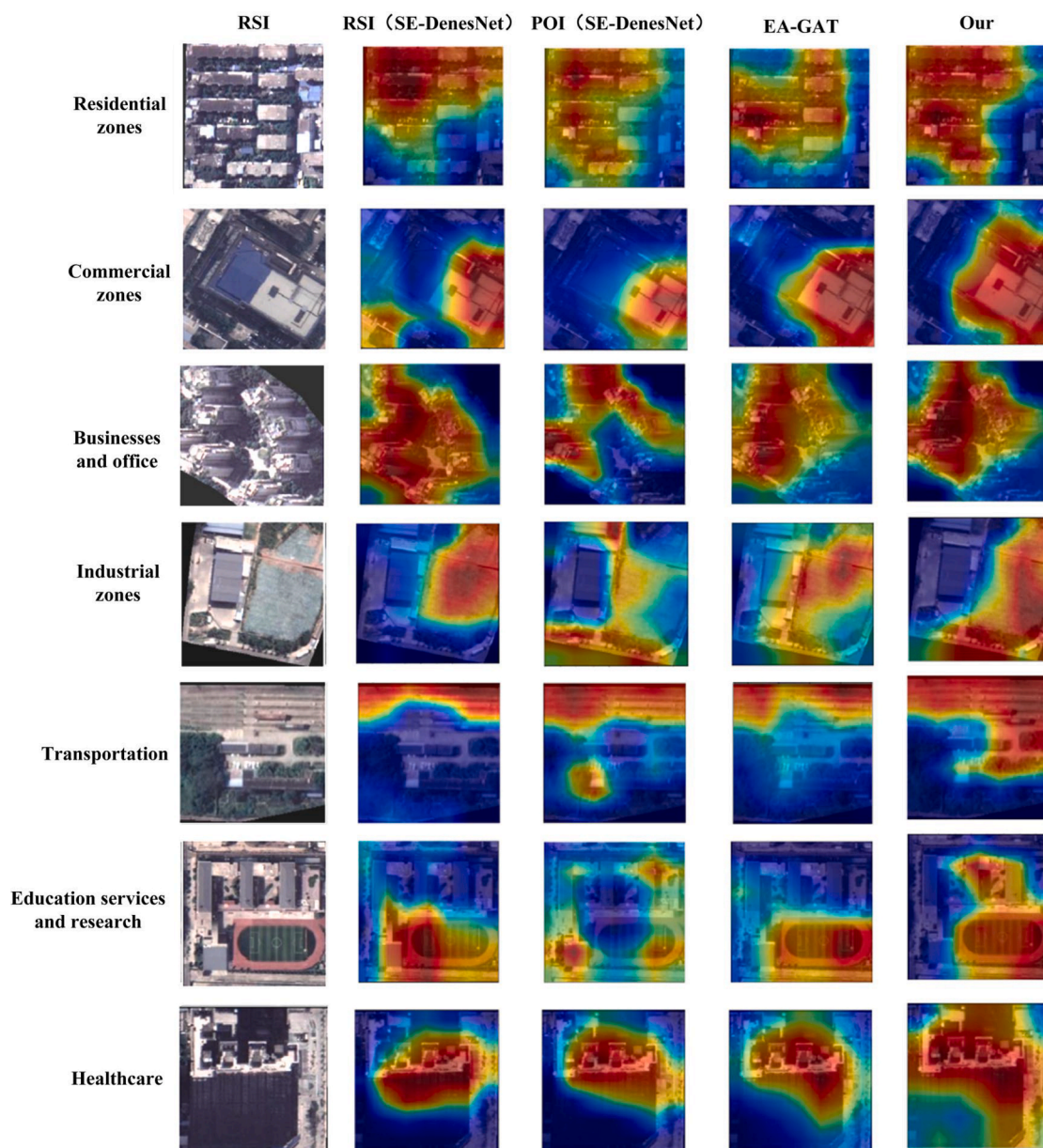


Fig. H. Visualization of class activation maps.

References

- Bai, L., et al., 2024. AP-Semi: improving the semi-supervised semantic segmentation for VHR images through adaptive data augmentation and prototypical sample guidance. *IEEE Trans. Geosci. Remote Sens.* 62, 1–13.
- Bai, L., et al., 2025. Integrating remote sensing with OpenStreetMap data for comprehensive scene understanding through multi-modal self-supervised learning. *Remote Sens. Environ.* 318, 114573.
- Bao, H., et al., 2020. DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. *Remote Sens. (Basel)* 12.
- Bao, H., et al., 2024. Deep siamese network for annual change detection in Beijing using landsat satellite data. *Int. J. Appl. Earth Obs. Geoinf.* 130, 103897.
- Cai, B., et al., 2023. Deep learning-based building height mapping using Sentinel-1 and Sentinel-2 data. *Int. J. Appl. Earth Obs. Geoinf.* 122.
- Çalışkan, O., et al., 2022. Morphological indicators of the building fabric: towards a Metric Typomorphology. *J. Urbanis.: Int. Res. Placemaking Urban Sust.* 1–30.
- Che, Y., et al., 2024. Building height of Asia in 3D-GloBFP.
- Chen, B., et al., 2022a. Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network. *Int. J. Appl. Earth Obs. Geoinf.* 109, 102794.
- Chen, D., et al., 2025. Interpreting core forms of urban morphology linked to urban functions with explainable graph neural network. *Comput. Environ. Urban Syst.* 118, 102267.
- Chen, G., et al., 2024. Remote sensing of diverse urban environments: from the single city to multiple cities. *Remote Sens. Environ.* 305, 114108.
- Chen, S.B., et al., 2022b. Remote sensing scene classification via multi-branch local attention network. *IEEE Trans. Image Process.* 31, 99–109.
- Chen, W., et al., 2021. Classification of urban morphology with deep learning: application on urban vitality. *Comput. Environ. Urban Syst.* 90, 101706.
- Chen, Z., Huang, B., 2024. Achieving urban vibrancy through effective city planning: a spatial and temporal perspective. *Cities* 152.
- Cheng, Y., et al., 2023. Multi-scale Feature Fusion and Transformer Network for urban green space segmentation from high-resolution remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* 124, 103514.
- Dong, X., et al., 2020. Exploring Impact of Spatial Unit on Urban Land Use Mapping with Multisource Data. In: *Remote Sensing*.
- Du, S., et al., 2021. 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.* 261, 112480.
- Du, S., et al., 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. *Gisci. Rem. Sens.* 57, 411–430.

- Du, S., et al., 2024. Mapping urban functional zones with remote sensing and geospatial big data: a systematic review. *Gisci. Rem. Sens.* 61.
- Eisenschink, P.M., et al., 2025. Forest variables from LiDAR: drone flight parameters impact the detection of tree stems and diameter estimates. *Eco. Inform.* 88, 103127.
- Fan, R., et al., 2022. Urban informal settlements classification via a transformer-based spatial-temporal fusion network using multimodal remote sensing and time-series human activity data. *Int. J. Appl. Earth Obs. Geoinf.* 111, 102831.
- Fan, Z., et al., 2025. Coverage and bias of street view imagery in mapping the urban environment. *Comput. Environ. Urban Syst.* 117, 102253.
- Foley, J.P., Dorsey, J.G., 1984. A review of the Exponentially Modified Gaussian (EMG) function: evaluation and subsequent calculation of universal data. *J. Chromatogr. Sci.* 22, 40–46.
- Gong, P., et al., 2020. Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018. *Science Bulletin* 65, 182–187.
- Gong, Z., et al., 2025. Multi-spatial urban function modeling: a multi-modal deep network approach for transfer and multi-task learning. *Int. J. Appl. Earth Obs. Geoinf.* 136, 104397.
- Guo, X., et al., 2024a. Contrastive learning-based knowledge distillation for RGB-thermal urban scene semantic segmentation. *Knowl.-Based Syst.* 292, 111588.
- Guo, Y., et al., 2024b. Identifying up-to-date urban land-use patterns with visual and semantic features based on multisource geospatial data. *Sust. Cities Soc.* 101, 105184.
- He, D., et al., 2025. Visual-language reasoning segmentation (LARSE) of function-level building footprint across Yangtze River Economic Belt of China. *Sustain. Cities Soc.* 127, 106439.
- Ho, Y., Wookey, S., 2020. The real-world-weight cross-entropy loss function: modeling the costs of mislabeling. *IEEE Access* 8, 4806–4813.
- Hong, D., et al., 2023. Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* 299, 113856.
- Hu, J., et al., 2023. Recognizing mixed urban functions from human activities using representation learning methods. *Int. J. Digital Earth* 16, 289–307.
- Hu, S., et al., 2021. Urban function classification at road segment level using taxi trajectory data: a graph convolutional neural network approach. *Comput. Environ. Urban Syst.* 87, 101619.
- Huang, W., et al., 2024. Zero-shot urban function inference with street view images through prompting a pretrained vision-language model. *Int. J. Geogr. Inf. Sci.* 38, 1414–1442.
- Huang, W., et al., 2022. Estimating urban functional distributions with semantics preserved POI embedding. *Int. J. Geogr. Inf. Sci.* 36, 1905–1930.
- Huang, W., et al., 2023. Learning urban region representations with POIs and hierarchical graph infomax. *ISPRS J. Photogramm. Remote Sens.* 196, 134–145.
- Izrailev, F.M., Castañeda-Mendoza, A., 2006. Return probability: exponential versus Gaussian decay. *Phys. Lett. A* 350, 355–362.
- Ji, S., et al., 2024. Above-ground biomass retrieval with multi-source data: prediction and applicability analysis in Eastern Mongolia. *Land Degrad. Dev.* 35, 2982–2992.
- Kong, B., et al., 2024. A graph-based neural network approach to integrate multi-source data for urban building function classification. *Computers, Environment and Urban Systems*, 110.
- Lei, B., et al., 2024. Predicting building characteristics at urban scale using graph neural networks and street-level context. *Comput. Environ. Urban Syst.* 111, 102129.
- Levy, A., 1999. Urban morphology and the problem of the modern urban fabric: some questions for research. *Urban Morphol.* 3, 79–85.
- Li, F., et al., 2018. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* 113, 390–400.
- Li, Z., et al., 2025. Enhanced mapping of essential urban land use categories in China (EULUC-China 2.0): integrating multimodal deep learning with multisource geospatial data. *Sci. Bull.*
- Li, Z., et al., 2024a. Deep learning for urban land use category classification: A review and experimental assessment. *Remote Sensing of Environment*, 311.
- Li, Z., et al., 2024. Deep learning for urban land use category classification: a review and experimental assessment. *Remote Sens. Environ.* 311, 114290.
- Li, Z., et al., 2023. SinoLC-1: the first 1 m resolution national-scale land-cover map of China created with a deep learning framework and open-access data. *Earth Syst. Sci. Data* 15, 4749–4780.
- Liang, X., et al., 2023. Revealing spatio-temporal evolution of urban visual environments with street view imagery. *Landsc. Urban Plan.* 237, 104802.
- Lin, A., et al., 2024. An MIU-based deep embedded clustering model for urban functional zoning from remote sensing images and VGI data. *Int. J. Appl. Earth Obs. Geoinf.* 128, 103689.
- Lin, X., et al., 2025. From points to patterns: an explorative POI network study on urban functional distribution. *Comput. Environ. Urban Syst.* 117, 102246.
- Liu, R., et al., 2024a. SoftFormer: SAR-optical fusion transformer for urban land use and land cover classification. *ISPRS J. Photogramm. Remote Sens.* 218, 277–293.
- Liu, R., et al., 2022. Hybrid transformer networks for urban land use classification from optical and SAR images. In: *IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 707–710.
- Liu, S., et al., 2024b. Lighting characteristics of public space in urban functional areas based on SDGSAT-1 glimmer imagery: a case study in Beijing, China. *Rem. Sens. Environ.* 306, 114137.
- Lore, M., et al., 2024. A hybrid deep learning method for identifying topics in large-scale urban text data: Benefits and trade-offs. *Comput. Environ. Urban Syst.* 111, 102131.
- Lu, W., et al., 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. *Remote Sens. Environ.* 270, 112830.
- Lv, X., et al., 2021. Improved object-based convolutional neural network (IOCNN) to classify very high-resolution remote sensing images. *Int. J. Remote Sens.* 42, 8318–8344.
- Ma, H., et al., 2024. How does spatial structure affect psychological restoration? a method based on graph neural networks and street view imagery. *Landsc. Urban Plan.* 251, 105171.
- Mao, A., et al., 2023. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In K. Andreas, B. Emma, C. Kyunghyun, E. Barbara, S. Sivan, & S. Jonathan (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 23803–23828). *Proceedings of Machine Learning Research: PMLR*.
- Moudon, A.V., 1997. Urban Morphology as an emerging interdisciplinary field. *Urban Morphol.* 1, 3–10.
- Ouyang, S., et al., 2023. MDFF: a method for fine-grained UFZ mapping with multimodal geographic data and deep network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 16, 9951–9966.
- Ouyang, S., et al., 2025. Object-based urban land-use change detection with siamese network and hierarchical clustering. *IEEE Trans. Geosci. Remote Sens.* 63, 1–15.
- Shekhar, S., Huang, Y., 2001. Discovering spatial co-location patterns: a summary of results. In: Jensen, C.S., Schneider, M., Seeger, B., Tsotras, V.J. (Eds.), *Advances in Spatial and Temporal Databases*. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 236–256.
- Su, C., et al., 2024. A multimodal fusion framework for urban scene understanding and functional identification using geospatial data. *Int. J. Appl. Earth Obs. Geoinf.* 127.
- Sun, R., et al., 2025. Urban region function classification via fusing optical imagery and social media data: a spatio-temporal Transformer interaction approach. *Inf. Fusion* 121, 103140.
- Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* 8, 127–150.
- Wang, Z., et al., 2025. Combining feature compensation and GCN-based reconstruction for multimodal remote sensing image semantic segmentation. *Inf. Fusion*.
- Wen, Z., et al., 2020. SO-CNN based urban functional zone fine division with VHR remote sensing image. *Remote Sens. Environ.* 236, 111458.
- Xu, H., et al., 2024. Segmenting Urban Scene Imagery in Real Time using an Efficient UNet-like Transformer. *Appl. Sci.* 14.
- Yu, W., 2016. Spatial co-location pattern mining for location-based services in road networks. *Expert Syst. Appl.* 46, 324–335.
- Yuan, J., et al., 2022. Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: a case study of Wuhan. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 15, 3972–3991.
- Zhang, K., et al., 2022. Distance weight-graph attention model-based high-resolution remote sensing urban functional zone identification. *IEEE Trans. Geosci. Remote Sens.* 60, 1–18.
- Zhang, X., et al., 2018. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. *Remote Sens. Environ.* 212, 231–248.
- Zhang, X., et al., 2020. Heuristic sample learning for complex urban scenes: application to urban functional-zone mapping with VHR images and POI data. *ISPRS J. Photogramm. Remote Sens.* 161, 1–12.
- Zhang, X., et al., 2023. Synergistic Classification of Multilevel Land Patches (SC-MLPs): reducing conflicts and improving mapping results for land uses and functional spaces with very-high-resolution satellite imagery. *IEEE Trans. Geosci. Remote Sens.* 61, 1–17.
- Zhang, Y., et al., 2024. Multi-level urban street representation with street-view imagery and hybrid semantic graph. *ISPRS J. Photogramm. Remote Sens.* 218, 19–32.
- Zheng, Z., et al., 2024. Global perspectives on sand dune patterns: Scale-adaptable classification using Landsat imagery and deep learning strategies. *ISPRS J. Photogramm. Remote Sens.* 218, 781–801.
- Zhong, Y., et al., 2023. Global urban high-resolution land-use mapping: from benchmarks to multi-megacity applications. *Remote Sens. Environ.* 298, 113758.
- Zhou, W., et al., 2023. Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information. *Remote Sens. Environ.* 297.
- Zhou, W., et al., 2024. Hierarchical building use classification from multiple modalities with a multi-label multimodal transformer network. *Int. J. Appl. Earth Obs. Geoinf.* 132.
- Zhu, X.X., et al., 2022. The urban morphology on our planet – Global perspectives from space. *Remote Sens. Environ.* 269, 112794.