ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on Feature extraction of City Block Style and Guidance of Architectural Form Planning Based on Deep Learning

Yu Han[1,*]

[1] Henan Polytechnic Institute, Nanyang City, Henan Province, 473000, China

SUMMARY: *This paper proposes a deep learning framework for feature extraction and architectural form planning guidance of city blocks. Street view images, remote sensing Windows, building contours, and facade labels from 126 sampling blocks are registered, and a data set consisting of 12,840 street view images, 126 groups of remote sensing Windows, 38 building form indicators, and 6 types of style labels is constructed. The model uses a dual-vision branch encoder and a morphological representation module to extract facade texture, roof contour, height rhythm and interface continuity information, and then maps the style representation into planning control parameters. Experimental results show that the Accuracy of the model in the landscape feature extraction task reaches 94.1%, and Macro-F1 reaches 0.912. In the architectural form planning guidance experiment, the guidance consistency of the proposed method reaches 91.6%, the boundary satisfaction reaches 93.3%, the block coordination reaches 90.9%, and the form deviation is 18.7%. The results show that the framework can transform the visual features of blocks into computable planning basis, and provide stable technical support for planning and design.*

KEYWORDS: *Deep learning; City block style; Feature extraction; Architectural form planning guidance*

1 Introduction

Urban block style carries information such as building scale, interface order, facade details and spatial relationships, and is also an important object in digital city modeling, building form recognition and planning decision calculation. In the face of large-scale street view images, remote sensing images and building contour data, manual interpretation methods are difficult to support continuous processing and unified expression of samples. Therefore, the introduction of deep learning into block style feature extraction and architectural form planning guidance can transform visual perception objects into trainable and computable morphological representations. For urban renewal, renovation of historic districts and control of newly built areas, if a stable style recognition model can be constructed on the basis of multi-source data, and the recognition results can be further mapped into building form control parameters, the planning guidance will have a clearer quantitative basis and stronger spatial adaptation ability.

In the related research of city block style recognition, streetscape visual perception, building geometry recovery and multi-source spatial representation constitute a group of technical paths that have entered the analysis framework of deep learning earlier. Wang L et al. studied the relationship between street perception and space syntax, and showed that deep learning can extract visual information with planning significance from street view images [1]. Yan Y et al.

*Hanyu2008045@163.com
https://doi.org/10.65102/is2026271
Published: 30 April, 2026

proposed a building height estimation method based on a single street view image, and verified the feasibility of deep network to recover building geometric attributes [2]. Pang H E et al. studied the 3D building reconstruction under the condition of single view street scene, which provided a new realization path for the expression of block form [3]. Cai J et al. proposed the generalization method of urban form based on unsupervised deep learning to enhance the abstract representation ability of urban form [4]. Najmi A et al. studied the joint mapping method of remote sensing images and street view images, indicating that multi-source data fusion is helpful to enhance urban space recognition [5]. Wu J et al. proposed an urban form analysis framework based on big data and machine learning, and demonstrated the computational value of form indicators in the analysis of historic districts [6]. Xu Y et al. studied the task of building function classification, and introduced deep learning into the process of building semantic recognition, which expanded the inference range of building objects from form to use [7]. Chen F C et al. proposed a building attribute estimation method based on street view images and feature fusion, which enhanced the correlation expression between building appearance and attribute variables [8]. The research at this stage makes the block style gradually shift from the empirical description to the quantifiable, trainable and transferable computational expression, and also lays the data and method foundation for the subsequent morphological feature modeling.

With the expansion of street view image data scale and the enhancement of visual learning ability, the research focus has shifted from single geometric recognition to more detailed style perception, spatial evaluation and facade analysis. Zhang L et al. studied the relationship between streetscape visual elements and urban green structures, indicating that local visual cues in streetscape images can support the quantification of spatial features [9]. He J et al. proposed a human perception extraction method based on street view images, which provides a basis for visual evaluation calculation in urban renewal scenes [10]. Ki D et al. studied the construction method of walkability index driven by street view images, which reflects the adaptability of deep learning in street environment quantification [11]. Kang Y et al. proposed a neighborhood safety perception analysis method combining GeoAI and survey data, which promoted the mapping from subjective visual cognition to computational model [12]. Patel P et al. studied the deep learning morphological method for urban scale environmental modeling, which strengthened the connection between urban morphological structure and environmental simulation [13]. Lu Y et al. proposed an improved SOLOv2 building facade analysis method to improve the accuracy of facade component segmentation [14]. Xu H et al. studied urban architectural style recognition and data set construction methods under the condition of street view images, so as to provide a stable data foundation for block style classification [15]. This kind of research shows that block style is not just a static collection of building appearance, but can be decomposed into multi-level visual information such as style, interface, component and environmental relationship through deep feature learning.

In higher-level research on urban form computing, related work has begun to focus on the cohesive mechanism between visual perception results and spatial structure, functional semantics, and planning context. Sun H et al. proposed a street spatial analysis method based on deep learning to quantify the relationship between element perception and overall perception [16]. Zhai Y et al. studied the relationship between color distribution, harmony and street functions of building facades, so that the visual features of facades have interpretable computational forms [17]. Ogawa Y et al. proposed a subjective perception evaluation method of street scenes based on street view images, indicating that there is a stable mapping between visual representation and spatial cognition [18]. Wu C et al. studied the machine learning representation method of urban form based on street network, which enhanced the structure of

block pattern calculation [19]. Fleischmann M et al. proposed the decoding method of urban form and function based on spatial explicit deep learning, which expanded the technical boundary of the translation from form features to planning semantics [20]. These results promote the research of urban form from "recognizing buildings" to "understanding blocks", and also show that there is a clear computing foundation and method extension space for further using the feature extraction results for architectural form planning guidance.

Existing research has formed a technical chain from street view perception, building attribute estimation to urban form modeling, but the integrated computing framework for city block style recognition and building form planning guidance still needs to be improved. Based on this, this paper constructs a unified representation mechanism for multi-source data, designs a block style feature extraction model, and establishes a planning guidance method driven by morphological representation learning, so that building height, facade rhythm, interface continuity, roof contour and block style vector form a mapping. This study establishes a data-model-guidance link between the visual feature analysis of urban blocks and the support of architectural form planning, which provides a new calculation path for form control and planning assistance.

2 Methods and materials

2.1 Multi-source data construction and sample representation for city block style recognition

In order to ensure that the recognition results of city block style have visual integrity, geometric consistency and planning interpretability, we do not use a single street view image as input at the data level, but combine street view images, orthogonal remote sensing images, building contours, facade annotations, plot boundaries and road centerlines to construct multi-source samples. Street view images are used to preserve the texture, color and interface continuous information of building facades, remote sensing slices are used to supplement roof contours, volume distribution and the overall pattern of blocks, building contours and plot boundaries are used to provide morphological constraints, and road centerlines are used to describe street openness and visual organization. All kinds of data are uniformly projected to the same coordinate system, and then cut and numbered according to block units, so that the samples can correspond to clear spatial locations and morphological contexts.

In order to ensure the consistency of multi-source block samples in space, time and semantic level, this paper uses the unified registration function to complete the basic alignment processing and calculation mapping:

$$\mathcal{A}_i = \arg\min_{\Delta x,\Delta y,\theta} \sum_{p\in\Omega_i} \left\| I_i(p) - R_i\left(T_{\Delta x,\Delta y,\theta}(p)\right)\right\|_2^2 + \lambda_1\Phi_i + \lambda_2\Psi_i \quad (1)$$

Here, $\mathcal{A}_i$ represents the optimal registration result of the i block unit, I_i represents the street view view, R_i represents the remote sensing slice, $T_{\Delta x,\Delta y,\theta}$ represents the spatial transformation composed of translation and rotation, Ω_i represents the effective pixel area, Φ_i represents the overlap constraint of building contour, Ψi represents the semantic label consistency constraint, λ_1 and λ_2 are the weight coefficients. This formula is used to weaken the interference of shooting deviation and spatial dislocation on subsequent style representation.

As shown in Fig. 1, the original data first goes through coordinate correction, time screening and duplicate sample elimination, and then enters the three stages of block slicing, building matching and semantic verification. Street view images are reorganized according to the

acquisition orientation and shooting distance, remote sensing images are generated by fixed-scale Windows according to block boundaries, and building contours are mapped to facade labels through topological relationships. After registration, each block unit simultaneously retains the image view, planar morphology and attribute label, and then forms a unified sample for network training.

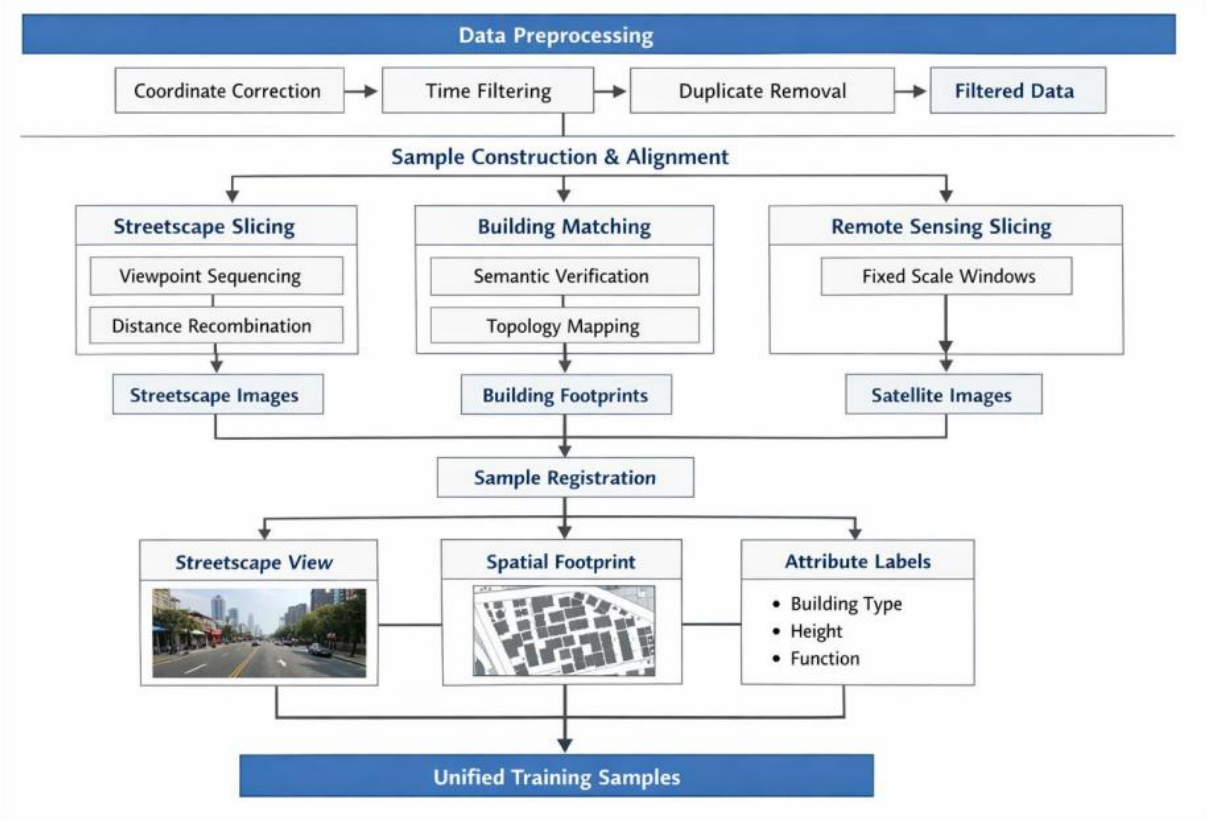

Figure 1: Flowchart of multi-source sample construction and registration for city blocks

In the stage of sample organization, this paper takes block as the basic unit of analysis, while preserving building level details and street level context. Each sample contains four types of core inputs, namely, street view main view, remote sensing window, building form matrix, and attribute label sequence. The main street view is uniformly scaled to 512×512, and the remote sensing window is uniformly cut to 256×256. The building form matrix is composed of indicators such as building height, layers, depth, roof form, base area and interface retreat, and the attribute label sequence records the style category, function type, age interval and street level. This design enables the sample to not only express the local details of the building, but also retain the overall pattern of the block.

In order to make street view images, remote sensing slices and building contours enter the same feature space, this paper defines the sample representation as the following joint coding form for expression:

$$X_i = [f_i^{sv} \,\|\, f_i^{rs} \,\|\, M_i \,\|\, L_i], \quad f_i^{sv} = E_{sv}(I_i), \quad f_i^{rs} = E_{rs}(R_i) \tag{2}$$

where X_i represents the block comprehensive sample vector, f_i^{sv} and f_i^{rs} represent the visual features output by the street view encoder and remote sensing encoder respectively, M_i represents the building form matrix, L_i represents the attribute label embedding, $\|$ represents the feature splicing operation. This representation compreszes multi-source input into a unified

sample interface, which is convenient for subsequent joint learning of deep networks.

As shown in Fig. 2, block samples are not simply superimposed before entering the network, but view grouping, building matching, indicator calculation and label review are completed first. The street view branch outputs the facade texture, color and window rhythm, the remote sensing branch outputs the roof relationship and volume distribution, the morphology branch outputs the height gradient, density level and boundary regression characteristics, and the label branch is responsible for establishing the correspondence between the style categories and the building form control items. After this process, the samples are transformed from the original image set to the training object with structural hierarchy.

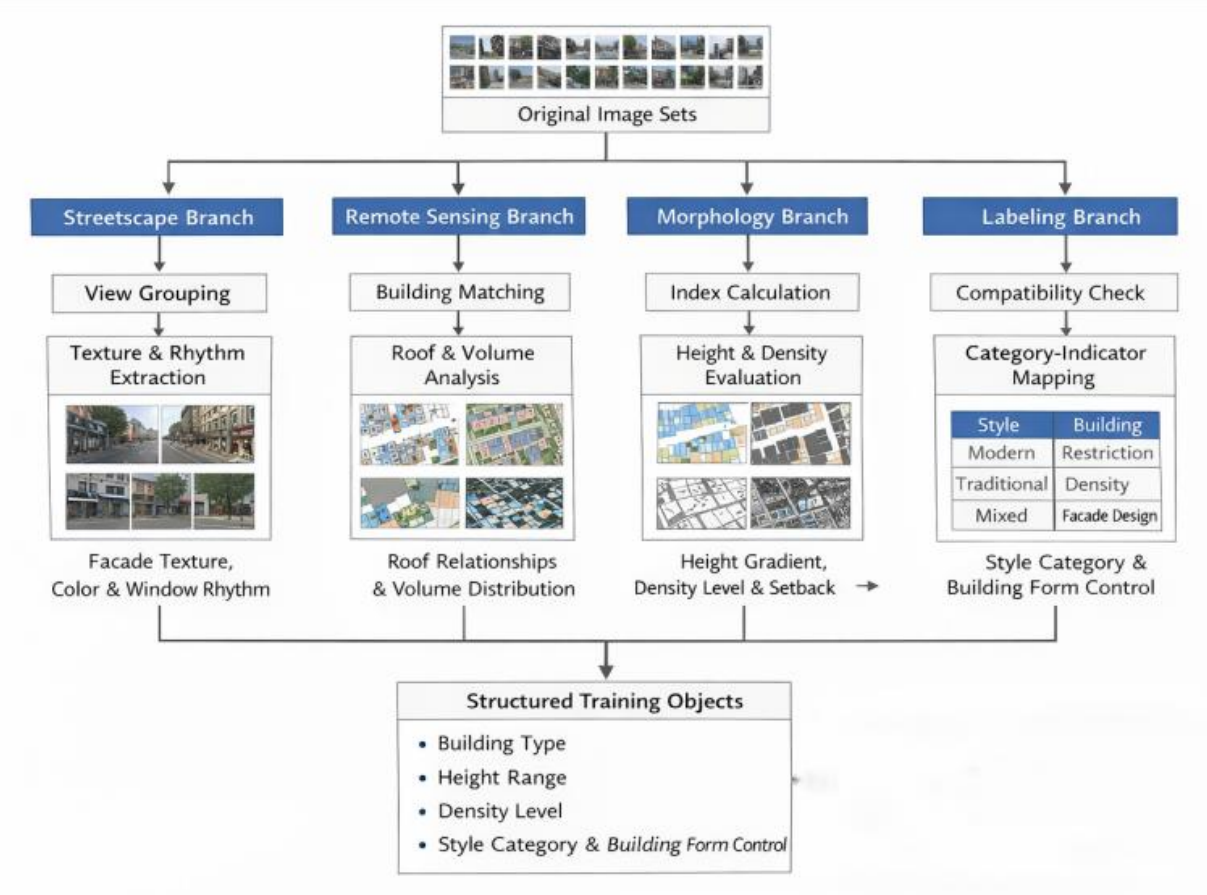

Figure 2: Schematic diagram of block sample structured representation and label generation

In order to quantify the differences of height fluctuation, interface continuity and volume density in the block style, this paper constructs the morphological statistical vector and gives the complete expression of the calculation method as follows:

$$m_i = [\mu_h, \sigma_h, \rho_c, \eta_d, \kappa_r], \quad \rho_c = \frac{1}{n_i - 1} \sum_{j=1}^{n_i - 1} \exp\left(-\frac{|s_{j+1} - s_j|}{\tau}\right) \tag{3}$$

Here, m_i represents the shape statistical vector of the i block, μ_h and σ_h represent the mean and standard deviation of building height, ρ_c represents the continuity of street interface, s_j represents the starting and ending position of adjacent building facades, η_d represents the building density index, κ_r represents the dispersion of roof contour, and τ represents the attenuation parameter. This formula can convert the order of the block perceived by the naked

eye into a comparable numerical index.

In order to suppress the offset effect of samples from different sources on the scale distribution, this paper introduces a normalized weight matrix to recalibrate and constrain the multi-modal features:

$$\tilde{X}_i = W_n(X_i - \mu)\Sigma^{-1/2}, \quad W_n = \mathrm{diag}(w_{sv}, w_{rs}, w_m, w_l) \tag{4}$$

where $\tilde{X}_i$ represents the normalized sample representation, μ and Σ represent the mean vector and covariance matrix of the training set respectively, W_n represents the modal weight matrix, w_{sv}, w_{rs}, w_m and w_l correspond to the weights of street view, remote sensing, morphology and label components respectively. The contribution of different modalities in joint training can remain stable through recali-fication.

In order to establish the adjacency correlation between block units and the style propagation path, this paper further uses the spatial proximity and attribute similarity to construct the graph structure representation as follows:

$$G = (V, E), \quad a_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma_d^2}\right) \cdot \exp\left(-\frac{\|m_i - m_j\|_2^2}{\sigma_m^2}\right) \tag{5}$$

Here, G represents the block graph structure, V represents the set of block nodes, E represents the set of adjacent edges, a_{ij} represents the edge weight between node i and node j, d_{ij} represents the spatial distance, and σ_d and σ_m represent the scale parameters. The introduction of graph structure makes the block style no longer limited to single sample expression, but can form a continuous propagation relationship between adjacent areas.

After the above processing, this paper forms a structured sample set with one-to-one correspondence of street view, remote sensing window and building level attributes. The set not only retains the visual differences of city block styles, but also retains the geometric constraints and semantic labels required for architectural form planning guidance, which provides a unified, stable and computable data basis and support for subsequent style feature extraction models and form planning guidance methods.

2.2 Feature extraction model of city block style based on deep learning

After the construction and unified representation of multi-source samples, this paper further constructs a deep learning feature extraction model for city block style recognition. The model does not separate street view images, remote sensing slices and morphological indicators, but completes visual coding, structure aggregation and style discrimination in a shared semantic space. The street view branch is responsible for capturing the facade material, window rhythm and color distribution, the remote sensing branch is responsible for characterizing the roof contour, volume combination and block texture, and the shape branch introduces the height gradient and boundary regression relationship. The three types of information are fused into the discrimination head, and the style category and shape response are output.

As shown in Fig. 3, the overall model consists of a visual encoding layer, a cross-modal fusion layer, a spatial relationship propagation layer, and a classification output layer. Street View images are first entered into a convolution-transformer hybrid encoder to preserve texture and facade relationships. The remote sensing image enters the hierarchical encoder, which is used to highlight the block contour and roof distribution. The morphological matrix is transformed into a low-dimensional morphological token by a multi-layer perceptual mapping. After the three-way coding is completed, the model performs the cross-attention operation in

the fusion layer, and then completes the adjacency propagation according to the block graph structure, so that the sample representation absorbs the spatial continuous information of adjacent blocks.

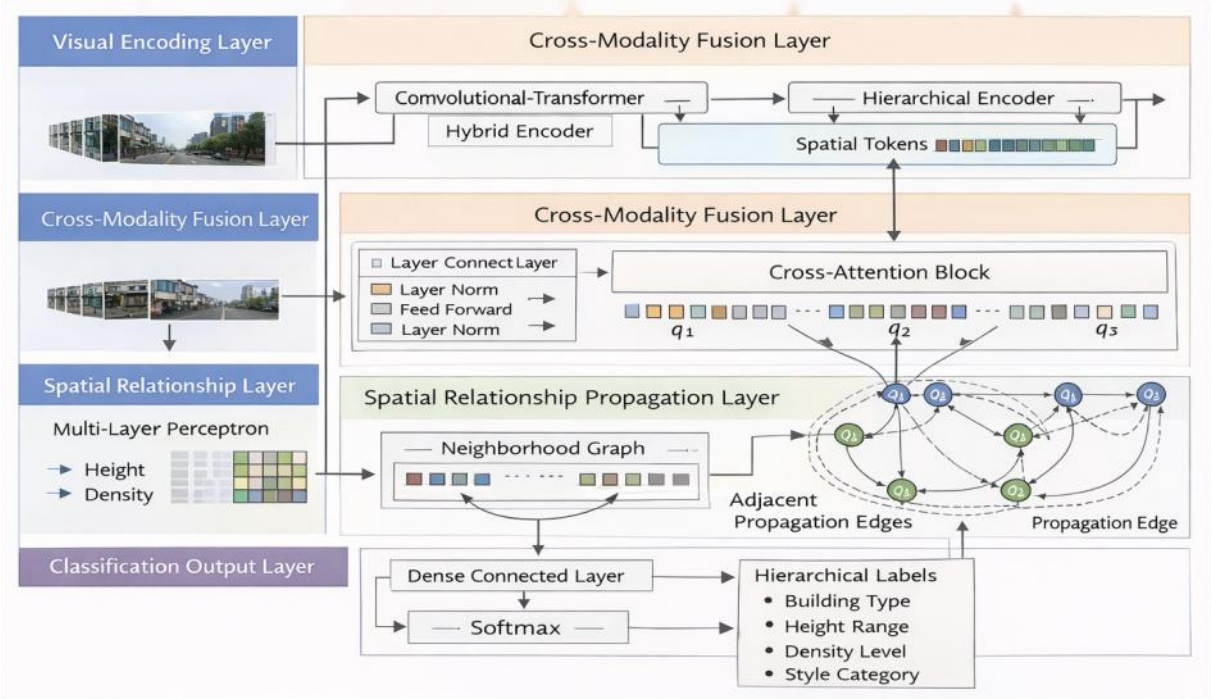

Figure 3: General structure diagram of the city block style feature extraction model

In order to unify the initial coding scales of the street view branch and the remote sensing branch, this paper first maps the two types of visual inputs into a shared feature space, which is calculated as follows:

$$z_i^{sv} = P_{sv}\,\phi_{sv}(I_i) + b_{sv}, \quad z_i^{rs} = P_{rs}\,\phi_{rs}(R_i) + b_{rs} \tag{6}$$

Here, z_i^{sv} and z_i^{rs} represent the street view feature vector and remote sensing feature vector of the i block, $\phi_{sv}(\cdot)$ and $\phi_{rs}(\cdot)$ represent the two-channel base coding function, P_{sv} and P_{rs} represent the projection matrix, and b_{sv} and b_{rs} represent the bias term. This formula is used to transform visual representations from different sources to the same dimension and provide an interface for subsequent joint learning.

In the local encoding process of the visual backbone, this paper uses window convolution and self-attention unit alternately stacked to take into account the texture details and the overall semantics of the block. The street view branch retains the facade edge and window density in the shallow layer, and extracts the style composition and interface rhythm in the deep layer. The remote sensing branch extracts block contours, road enclosions, and roof aggregation patterns under a large receptive field. Since the two types of inputs have different emphasis on characterizing the landscape, this paper introduces a morphological gating mechanism to make different blocks automatically adjust the contribution ratio of street view information and remote sensing information.

In order to make building height, interface distance and density features participate in visual fusion, this paper constructs a morphological gating vector to dynamically adjust the backbone features. The form is as follows:

$$g_i = \sigma\left(W_g m_i + b_g\right), \quad \hat{z}_i^{sv} = g_i \odot z_i^{sv}, \quad \hat{z}_i^{rs} = (1 - g_i) \odot z_i^{rs} \tag{7}$$

Here, g_i represents the morphological gating vector of the i block, m_i represents the morphological features obtained from the above statistics, W_g and b_g are trainable parameters, $\sigma(\cdot)$ represents the Sigmoid function, $\odot$ represents the Hadamard product. This formula enables the model to dynamically emphasize the facade information or volume information according to the shape difference of the block, avoiding the single mode dominating the classification results.

As shown in Fig. 4, the street view features, remote sensing features and morphological tokens after gated adjustment are input into the cross-modal fusion module. The module firstly performs bidirectional attention matching, and then retains the original modal features through residual aggregation, and then uses block adjacent edges to complete semantic diffusion in the graph propagation unit. The nodes after propagation represent the entry into the classification head to output the style class probability, and generate the style sensitive index response synchronously.

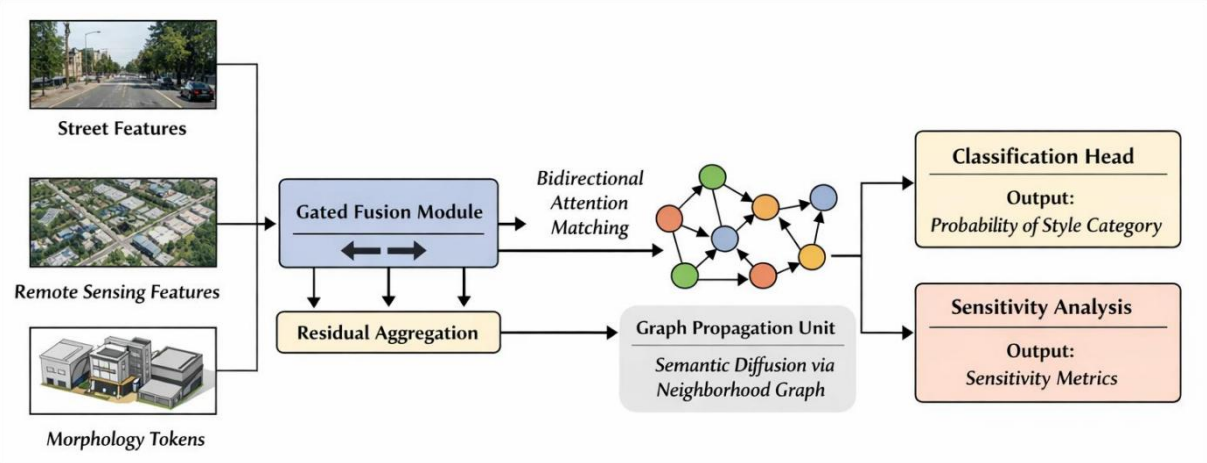

Figure 4: Cross-modal fusion and spatial relationship propagation module structure diagram

In order to enhance the expression of correspondence between different modalities, this paper uses bidirectional cross-attention mechanism in the fusion layer to complete the weight update, and the calculation process is expressed as follows:

$$B_{ij} = \mathrm{softmax}\left(\frac{(Q\,\hat{z}_i^{sv})(K\,\hat{z}_j^{rs})^{\top}}{\sqrt{d}}\right), \quad f_i = \sum_{j=1}^{N} B_{ij}\, V\, \hat{z}_j^{rs} + U t_i \tag{8}$$

Here, B_{ij} represents the attention weight of the i query vector of the street view branch to the j key vector of the remote sensing branch, Q, K, V, and U are linear mapping matrices, d represents the scaling dimension, t_i represents the morphological token, and f_i represents the fused intermediate representation. This formula establishes the explicit connection between facade information, roof information and morphological indicators by attention matching.

In block level style recognition, the visual features of a single sample are also affected by the organization of adjacent block boundaries. To this end, this paper puts the fused features into the graph propagation layer to perform semantic updates on the block adjacency graph. In the process of propagation, the neighboring weights are constrained by distance, morphological difference and road connectivity. This design makes the model maintain the stability of local

features while introducing neighborhood consistency, so that the style discrimination at the block boundary is smoother.

In order to describe the style transfer relationship between adjacent blocks, this paper performs a layer of normalized graph propagation operation on the adjacency graph, and its specific and complete expression is as follows:

$$h_i^{(l+1)} = \rho\left(\sum_{j\in\mathcal{N}(i)} \frac{a_{ij}}{\sqrt{\deg(i)\deg(j)}} W_h h_j^{(l)} + B_h h_i^{(l)}\right) \tag{9}$$

where $h_i^{(l)}$ represents the representation of node i at layer l, $\mathcal{N}(i)$ represents the adjacency set of node i, $\deg(i)$ represents the node degree, W_h and B_h are trainable matrices, and $\rho(\cdot)$ represents the nonlinear activation function. This formula enables the representation of block style to absorb the continuous organizational characteristics of adjacent Spaces.

At the classification output, we divide the propagated block representation into two branches, one for style category determination and the other for style sensitive attribute response estimation. The former outputs the probability of style classes such as continuous interface, diverse roof and traditional low floor, and the latter outputs indicators such as color harmony, interface integrity, volume balance and contour relief. The two branches share the intermediate representation, which is both discriminative and explanatory.

In order to map the fused block representation into a discriminative style class probability, this paper uses the following classification head to complete the final prediction and response calculation, which is expressed as follows:

$$p_i = \mathrm{softmax}\left(W_c h_i^{(L)} + b_c\right), \quad r_i = \tanh\left(W_r h_i^{(L)} + b_r\right) \tag{10}$$

where p_i represents the landscape category probability vector of the i block, r_i represents the landscape sensitive attribute response vector, W_c, W_r, b_c, b_r are the classification head and response head parameters, and $h_i^{(L)}$ represents the final layer representation. This formula provides two types of output for the subsequent planning guidance stage: category results and attribute strength.

In order to strengthen the constraint of consistent relationship between style categories and morphological attributes, this paper further sets the cross-task consistency regularization term, which is calculated as follows:

$$\mathcal{L}_{con} = \frac{1}{N}\sum_{i=1}^{N} \left\| Tp_i - r_i \right\|_2^2 + \beta \sum_{i=1}^{N} \sum_{j\in\mathcal{N}(i)} a_{ij} \left\| r_i - r_j \right\|_2^2 \tag{11}$$

Here $\mathcal{L}_{con}$ represents the consistency loss, T represents the mapping matrix from categories to attributes, and β represents the neighborhood smoothing weight. On the one hand, this formula constrains the semantic consistency between category prediction and attribute response, on the other hand, it constrains the attribute output of adjacent blocks not to be over-discrete.

In order to ensure the synchronous improvement of style classification accuracy and attribute response stability, this paper adopts multi-task joint objective organization of the overall training process, which can be specifically expressed as follows:

$$\mathcal{L} = \lambda_c \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{reg} + \lambda_{con} \mathcal{L}_{con} + \lambda_w \|\Theta\|_2^2 \quad (12)$$

Here, $\mathcal{L}$ represents the total model loss, $\mathcal{L}_{cls}$ represents the cross-entropy loss of style categories, $\mathcal{L}_{reg}$ represents the attribute response regression loss, $\mathcal{L}_{con}$ represents the consistency loss, Θ represents the set of model parameters, and λ_c, λ_r, λ_{con}, and λ_w are the weights of each item. Joint target training can make the model maintain stable convergence at three levels of category recognition, attribute expression and spatial continuity.

Through the above design, the landscape feature extraction model constructed in this paper no longer stays in the simple image classification, but integrates the street view appearance, remote sensing pattern and building form information into a unified computing framework. The model output can not only complete the recognition of city block style, but also generate sensitive attribute response and stable output with morphological meaning, which provides direct input for the guidance method of architectural form planning in the next section.

2.3 Architectural form planning guidance method based on morphological representation learning

The feature extraction model of city block style can output class probability and attribute response, but the planning guidance does not directly stop at the recognition result level, and it also needs to convert the style representation into executable building form control quantities. To this end, this paper proposes a guidance method for building form planning based on morphological representation learning, which maps the style vector, block map structure and plot constraints to the control results of height classification, interface retreat, roof organization, volume combination and facade rhythm. The method does not use a fixed rule table to match item by item, but learns a stable morphological correspondence between the historical district samples, the new area samples and the renovation samples, so that the output results not only maintain the continuity of the block style, but also meet the geometric boundary of the building layout and the spatial constraints of the road.

As shown in Fig. 5, the planning guidance module consists of a morphological prototype generation unit, a plot constraint encoding unit, a candidate scheme reasoning unit, and a guidance result output unit. Firstly, the block representation output by the style feature extraction model enters the prototype generation unit to form the corresponding morphological prototype vector in the shared embedding space. Subsequently, the plot boundary, back-off control, plot ratio interval, road interface and adjacent building contour are encoded as constraint vectors. The two types of vectors were coupled in the candidate scheme reasoning unit to obtain several candidate morphologies. Finally, the system output the planning guidance results according to the consistency score, boundary satisfaction degree and block coordination degree. This process enables planning suggestions to be derived from computable landscape mapping instead of static experience. Compared with the way of giving single-valued suggestions directly, this method emphasizes more on the intermediate reasoning process from block representation to plot form, so it can preserve the source basis and hierarchical structure of planning guidance. The prototype generation stage emphasizes style abstraction, the constraint encoding stage emphasizes plot boundary, the candidate reasoning stage emphasizes morphological combination, and the result output stage emphasizes screening logic. All of them together form a closed guiding link.

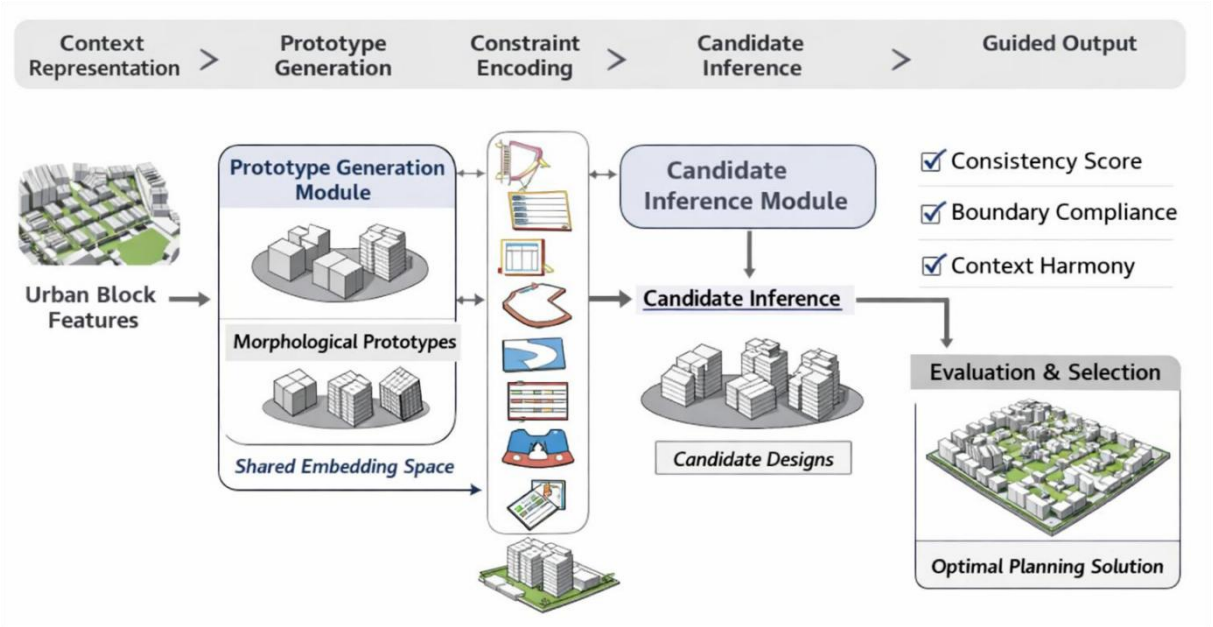

Figure 5: The overall flow chart of the guiding method for architectural form planning

In order to make the feature recognition results stably enter the planning and reasoning stage, and maintain the mapping relationship between category semantics and attribute strength, this paper defines the neighborhood shape prototype generation function as follows:

$$q_i = \tanh\left(W_q h_i^{(L)} + U_q r_i + b_q\right) \tag{13}$$

Here, q_i represents the morphological prototype vector of the i block, $h_i^{(L)}$ represents the final block representation of the model output in the previous section, r_i represents the attribute response vector, W_q and U_q are projection matrices, and $b_q$is the bias term. The formula is used to compress the landscape category information and attribute strength into a unified prototype representation, which provides the basic input for the subsequent inference of land-level control items.

In the plot level planning guidance, it is not enough to rely on the block prototype alone to express the actual control boundary. Geometric constraints such as red line of construction land, concession distance, road width, daylighting spacing and adjacent interface should be introduced simultaneously. In this paper, a constraint encoder is used to jointly map these discrete and continuous variables to constraint vectors, and jointly combine them with morphological prototypes before inference of candidate alternatives. After this process, the planning guidance output can not only retain the style trend of the original block, but also adapt to the shape difference and development intensity of different parcels.

In order to compress the control conditions such as plot boundary, road interface, backtrack distance and development intensity into a trainable representation, this paper constructs a constraint encoding function and adopts the following expression:

$$c_i = \sigma\left(W_c^{(1)} g_i^{pl} + W_c^{(2)} s_i^{rd} + W_c^{(3)} u_i^{bd} + b_c\right) \tag{14}$$

Here, c_i represents the constraint vector of the i plot, g_i^{pl} represents the geometric description of the plot, s_i^{rd} represents the road space constraint, u_i^{bd} represents the boundary

and backoff control, $W_c^{(1)}$ to $W_c^{(3)}$ is the mapping matrix, and $\sigma(\cdot)$ represents the activation function. This formula transforms the plot control condition into a continuous embedding, which enables it to participate in subsequent candidate morphological inference. Since different plot constraints come from heterogeneous data, this encoding step also assumes the role of scale unification and variable compression, so as to avoid representation fluctuations caused by high-dimensional control variables directly entering the generation process.

As shown in Fig. 6, the candidate morphological reasoning module adopts a two-branch structure. The first branch generates discrete control items such as height hierarchy, roof form and interface rhythm according to the shape prototype vector, and the second branch generates continuous control items such as volume concession, coverage and open proportion according to the constraint vector. The two branches are combined into a candidate scheme matrix in the fusion layer, and then fed into the result screening unit. The screening unit calculates the style consistency, geometric feasibility and block coordination at the same time, eliminates the schemes that do not meet the red line, back-off distance and adjacent interface relationship, and retains the guidance results that take into account both style inheritance and space adaptation. This structure ensures that planning proposals have a clear basis for generation and filtering logic. For the same plot, the system can output multiple candidate sequences and arrange them from high to low according to the comprehensive score for subsequent manual review and comparison of alternatives.

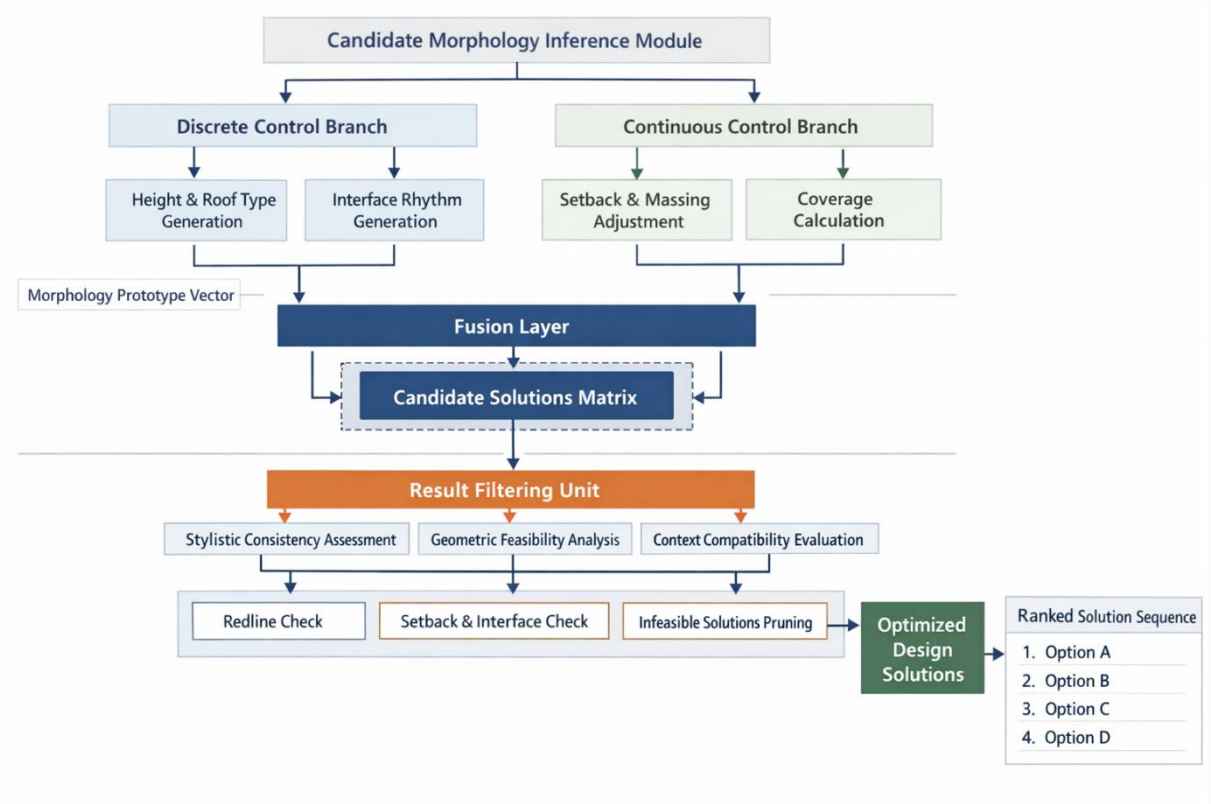

Figure 6: Structure diagram of candidate morphological reasoning and planning results screening

In order to describe the candidate scheme generation mechanism under the joint action of morphological prototype and plot constraint, this paper uses the joint inference function to complete the expression calculation of multiple groups of candidate results as follows:

$$y_i^k = W_y[\gamma_k q_i \parallel (1-\gamma_k)c_i \parallel e_k] + b_y, \quad k = 1, \ldots, K \tag{15}$$

Here, y_i^k represents the k candidate shape vector generated by the i plot, γ_k represents the style bias coefficient of the k type of scheme, e_k represents the candidate template embedding, W_y and b_y are inference parameters. This formula generates multiple groups of candidate results through different combinations of templates and weights, so that the system can cover a variety of volume organization methods. The candidate vectors can be backcomputed to obtain the building height interval, basement control line, roof type proportion and interface continuous level, so as to form a scheme expression with planning implications.

In order to ensure that the candidate morphology meets the requirements of style continuation boundary control and adjacent coordination at the same time, this paper sets up a consistent screening score function to complete the comprehensive score evaluation calculation as follows:

$$s_i^k = \alpha \cos(y_i^k, q_i) + \beta \exp(-\|B(y_i^k) - u_i^{bd}\|_2^2) + \delta \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \cos(y_i^k, y_j^*) \tag{16}$$

Here, s_i^k represents the score of the candidate solution, $\cos(\cdot,\cdot)$ represents the cosine similarity, $B(y_i^k)$ represents the boundary parameter obtained by the inverse calculation of the candidate vector, y_j^* represents the optimal solution of the adjacent plot, and α, β, δ are the weight coefficients. This formula evaluates style consistency, boundary satisfaction and neighborhood coordination at the same time, and is used to filter the final guidance results. The three items in the score function correspond to block inheritance, plot adaptation and neighboring coordination, so the output will not only favor a certain type of local index, but maintain a comprehensive balance.

In order to make the plan-guided network learn to generate quality ranking stability and neighborhood continuity relations synchronously in the training phase, this paper uses a joint objective function to organize the total training as follows.

$$\mathcal{L}_{pg} = \lambda_g \mathcal{L}_{gen} + \lambda_s \mathcal{L}_{sel} + \lambda_b \mathcal{L}_{bd} + \lambda_n \mathcal{L}_{nbr} \tag{17}$$

Here, $\mathcal{L}_{pg}$ represents the total planning guidance loss, $\mathcal{L}_{gen}$ represents the candidate morphological generation loss, $\mathcal{L}_{sel}$ represents the scheme ranking loss, $\mathcal{L}_{bd}$ represents the boundary constraint loss, $\mathcal{L}_{nbr}$ represents the neighborhood coordination loss, and λ_g, λ_s, λ_b, and λ_n are the weight parameters. This objective function makes the model pay attention to the morphological rationality of the generated results and the block continuity of the output results at the same time in the training stage. After training, the model can stably transform the landscape recognition output into land-level shape guidance suggestions, and provide a unified computing basis for consistency verification and planning response analysis in subsequent experiments.

Through the above design, this paper further transforms the block style recognition results into executable architectural form planning guidance results. While preserving the original spatial temperament of the block, this method can give control suggestions with calculation basis for the plot boundary, volume relationship and interface order, and provide a complete model support for the experimental analysis of planning guidance in the next chapter. At the same time, the three parts of morphological prototype, constraint coding and candidate screening also form a clear module boundary, which is convenient for subsequent transfer

training and parameter updating in different city samples.

3 Results

3.1 Experimental analysis of feature extraction of city block style

This section focuses on the recognition performance and training stability of the city block style feature extraction model. The experimental platform is configured with Intel Core i9-13900K processor, NVIDIA RTX 4090 graphics card, 24GB video memory, 64GB DDR5 memory and 2TB SSD, and the software environment is Ubuntu 22.04 LTS and PyTorch 2.1. In the training process, the batch size is set to 32, the initial learning rate is set to 0.0005, the optimizer uses AdamW, the weight decay coefficient is set to 0.01, the learning rate scheduling uses cosine annealing strategy, the minimum learning rate control is 1e-6, and the maximum number of training rounds is 120 epochs. The training was stopped when the validation set Macro-F1 did not continue to rise for 12 consecutive epochs. The experimental data consists of 126 sampling blocks, including 12840 street view images, 126 groups of remote sensing Windows, 38 building shape indicators and 6 types of landscape labels. The training set, validation set and test set are divided by the ratio of 7:1.

As shown in Fig. 7, the proposed method enters the stable convergence stage after the 18th epoch, and the decreasing trend of training loss and validation loss is synchronized without obvious separation. Compared with Baseline-CNN and Single-Scale Encoder, the convergence slope of the proposed model is larger in the first 30 epochs, indicating that the multi-source joint input and cross-modal fusion mechanism can shorten the parameter search phase. By the 52nd epoch, Macro-F1 of the validation set reached 0.901 and finally stabilized near 0.912. The corresponding validation loss is as low as 0.184, which is significantly lower than 0.263 and 0.241 of the two comparison models. This result shows that the joint training of street view branch, remote sensing branch and morphology branch does not introduce additional oscillation, but makes the style representation more concentrated.

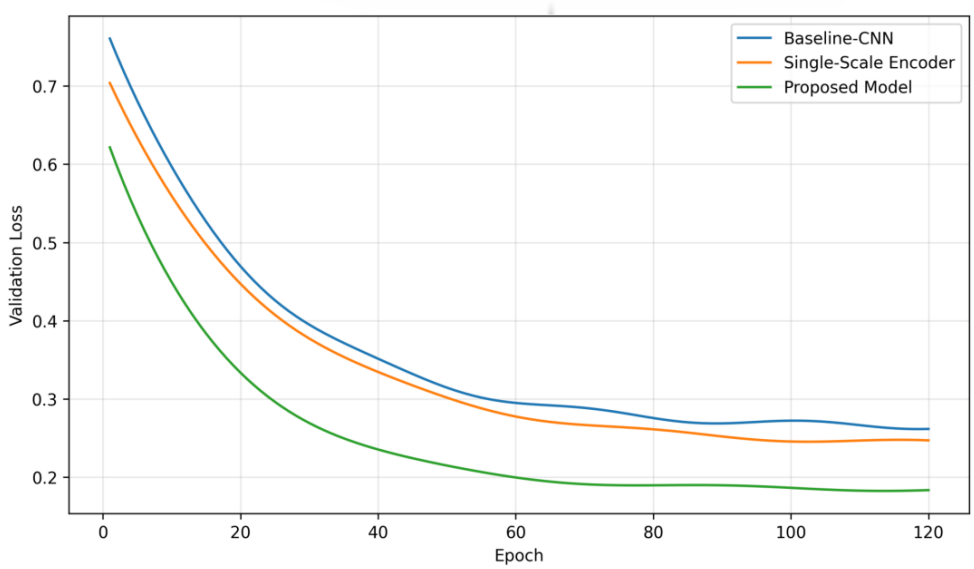

Figure 7: Convergence curve for training of the style feature extraction model

Table 1 shows that the overall recognition accuracy of the proposed model reaches 94.1%, which is 4.5%, 2.3% and 0.9% higher than that of Baseline-CNN, Single-Scale Encoder and

Dual-Branch Net, respectively. Macro-F1 increases by 0.048, 0.023 and 0.009, respectively. Although the average reasoning time is slightly increased, the single sample reasoning is still controlled within 30 ms, which can meet the computational requirements of batch analysis and planning auxiliary scenes of block landscape. The Kappa coefficient reaches 0.894, indicating that the model has high consistency on multi-category samples, and maintains stable discrimination ability between different types of features.

Table 1: Overall performance comparison of different models on the task of city block style recognition

Model	Accuracy / %	Macro-F1	Kappa	Average Inference Time / ms
Baseline-CNN	89.6	0.864	0.835	18.7
Single-Scale Encoder	91.8	0.889	0.861	21.4
Dual-Branch Net	93.2	0.903	0.881	24.9
Proposed Model	94.1	0.912	0.894	26.1

As shown in Fig. 8, after mapping the test set samples to the two-dimensional embedding space, traditional low-rise blocks, continuous interface blocks, modern mixed blocks, open high-rise blocks, industrial renewal blocks, and waterfront composite blocks form clear cluster boundaries. The embedding distribution output by the proposed model has few cross-bands, especially between the traditional low-rise blocks and the continuous interface blocks, which can form a stable distinction through the facade rhythm, roof contour and back boundary relationship. The comparison model overlaps between the modern mixed block and the waterfront complex block, indicating that it is difficult to rely on single-scale visual information to support complex landscape judgment. The embedding results also show that the morphological gating mechanism improves the division of labor efficiency between different modalities.

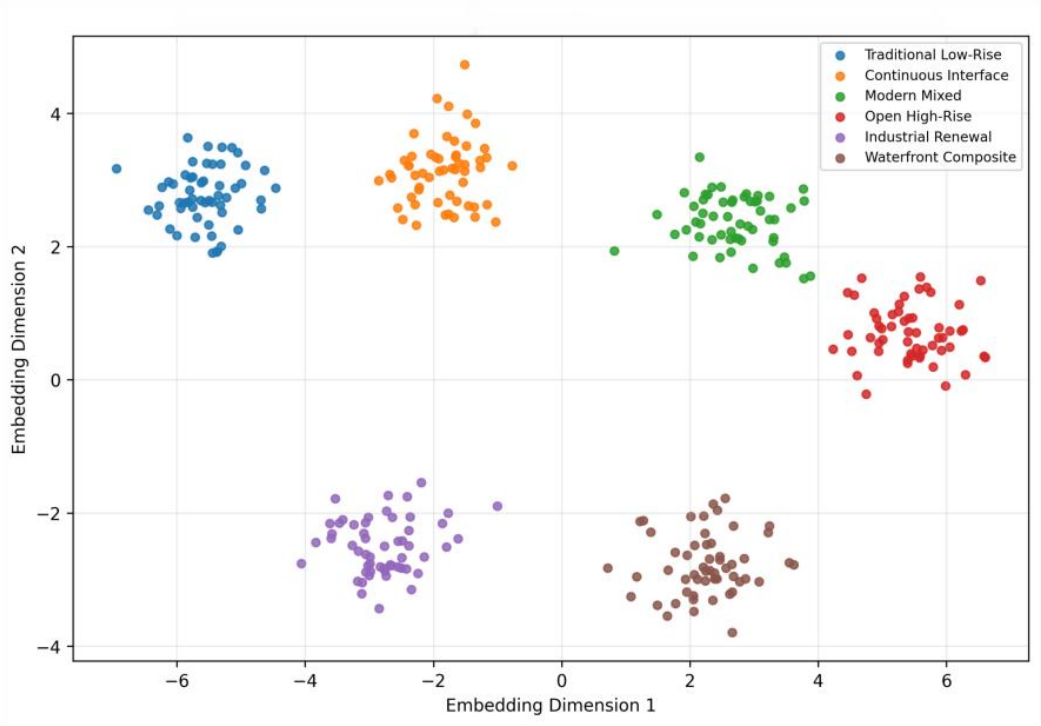

Figure 8: Visualization results of feature embedding for different style categories

Table 2 shows the migration identification results under different city grouping conditions. The overall accuracy of the central city sample is the highest, which indicates that regular road network and stable street interface are more beneficial to model learning. The recall rate of the historical area sample reaches 96.1% on the traditional low-rise, which reflects the strong sensitivity of the model to the characteristics of low-rise building density, roof continuity and interface details. The modern mixed recall rate of the newly expanded area sample reaches 94.9%, which indicates that the remote sensing pattern and volume relationship play a more direct role in the identification of new urban areas. The index of cross-region mixed samples decreased slightly, but the overall index remained above 0.906, indicating that the model in this paper has good robustness in cross-region landscape transfer.

Table 2: Model migration identification results under different city groupings

Test Group	Accuracy / %	Macro-F1	Traditional Low-Rise Recall / %	Modern Mixed-Style Recall / %
Central Urban Area Samples	94.6	0.918	95.2	92.7
Historical District Samples	93.8	0.910	96.1	89.6
Newly Developed Expansion Area Samples	94.3	0.914	92.8	94.9
Cross-Regional Mixed Samples	93.5	0.906	93.4	90.8

As shown in Fig. 9, in this paper, morphological gating, cross-attention fusion and graph propagation modules are removed in turn, five independent experiments are performed on the model and the average results are counted. After removing the morphological gating, the Accuracy decreases to 92.9%, and Macro-F1 decreases to 0.897, which indicates that the building height gradient and the retreating boundary difference have a direct impact on the discrimination of style. After removing the cross-attention fusion, Macro-F1 is further reduced to 0.891, indicating that the information alignment between street view and remote sensing is a key link to form stable category boundaries. After removing the graph propagation module, the misjudgment of the boundary samples in the central urban area increases significantly, and the overall Accuracy decreases to 93.1%. The complete model always maintains the highest index in the three groups of comparisons, and the standard deviation of each experiment is controlled within 0.4, indicating that the model structure has good repeatability and training stability.

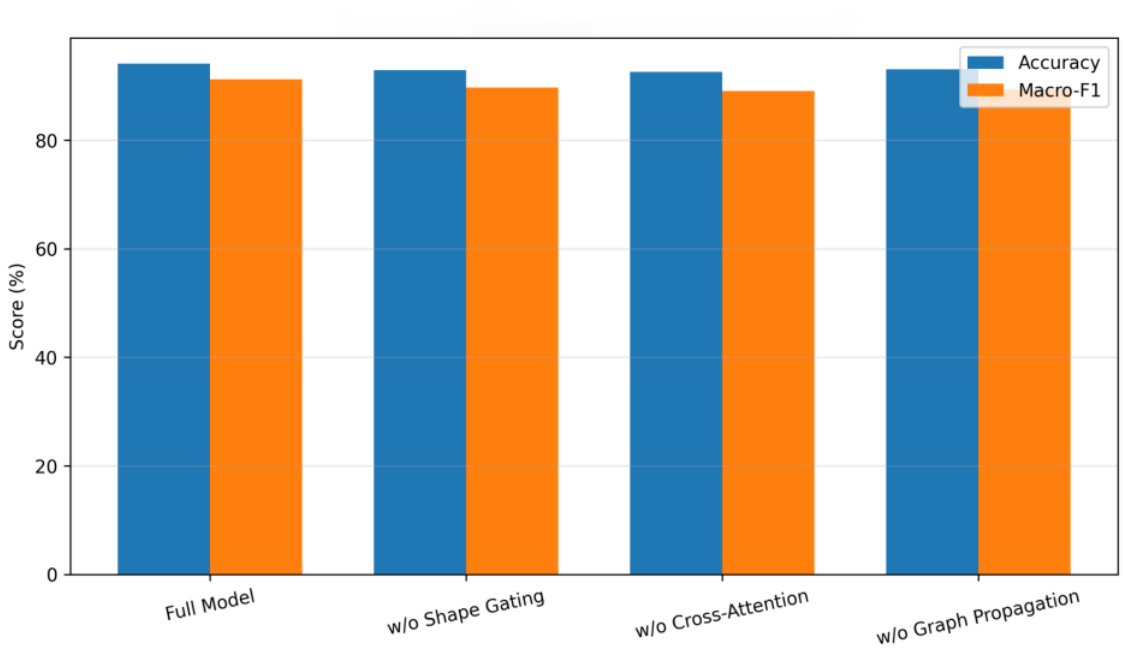

Figure 9: Comparison of ablation experiment results for different modules

Based on the above results, it can be seen that the feature extraction model of city block style constructed in this paper shows strong advantages in multiple dimensions such as convergence speed, overall accuracy, embedding distribution and cross-area migration. The multi-source sample representation ensures the geometric consistency of the input, the cross-modal fusion strengthens the correspondence between the facade information and the volume information, and the graph propagation mechanism improves the recognition stability of the block boundary samples.

3.2 Experimental analysis guided by architectural form planning

After the feature extraction of city block style, the planning guidance stage further maps the block style vector, shape prototype and parcel constraint into executable building form control results. The output of this part no longer stops at the recognition of style categories, but generates guidance information such as height hierarchy, interface retreat, roof organization, volume combination, and street continuity for plot scale. In order to test the adaptation ability of the method in different plot scenarios, this section analyzes from four dimensions of guidance consistency, boundary satisfaction, block coordination and shape deviation. The proposed method is compared with Rule-Mapping, MLP-Guidance, Graph-Constraint and Prototype-Search.

In this paper, the 126 sampled blocks are further mapped into 412 computable plot units, each of which contains block style representation, building form prototypes, boundary control information, road interface conditions, and adjacent building relationships. The planning guidance module completes candidate scheme reasoning and result screening under a unified training framework, and outputs land-level morphological suggestions. The overall performance of the different models in the multi-class plot scenario is shown in Fig. 10.

As shown in Fig. 10, the horizontal axis of the heat map shows four types of typical plot scenarios, including historical street and lane plots, main road interface plots, waterfront open plots, and updated mixed plots, and the vertical axis shows five guidance models. Darker colors indicate higher plan guidance consistency. The consistency of the proposed method in four types of scenes reaches 92.4%, 91.8%, 90.7% and 91.6%, respectively. The overall fluctuation

range is controlled within 1.7 percentage points, and the color distribution is more concentrated. Rule-Mapping performs reasonably well in historical street and lane plots, but decreases significantly in updated mixed plots, which indicates that fixed rules are difficult to maintain stable output under complex boundary conditions. The consistency of MLP-Guidance in the main road interface plot and the waterfront open plot is quite different, reflecting that the model's response to the interface continuity and open proportion is not balanced after the lack of spatial relationship propagation. Graph-Constraint is able to maintain high scores on road interface control, but the expression of roof organization and volume combination is not detailed enough. Prototype-Search outperforms the regular method in all four types of scenarios, but the boundary satisfaction is still affected when the candidate selection is insufficient. The proposed method maintains a stable color gradient in different scenes, which indicates a good coupling relationship between morphological prototypes, constraint encoding and candidate screening.

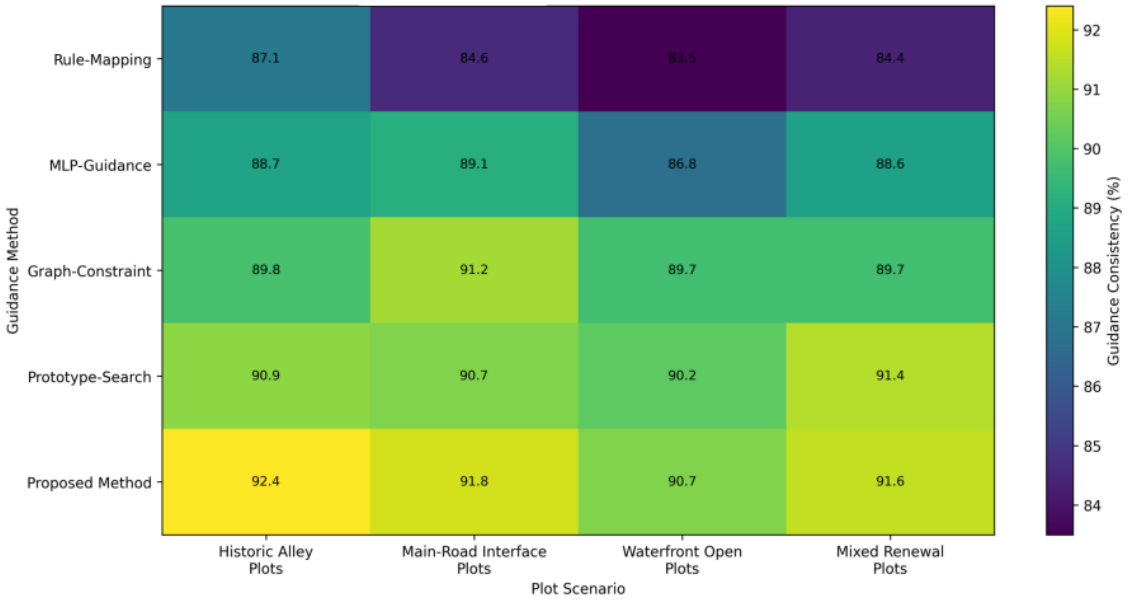

Figure 10: Heatmaps of planning guidance consistency for different plot scenarios

Table 3 shows that the proposed method achieves optimal results on the four core indicators. Compared with Rule-Mapping, the guidance consistency is increased by 6.7 percentage points, the boundary satisfaction is increased by 7.1 percentage points, the block coordination is increased by 8.2 percentage points, and the shape deviation is decreased by 4.9 percentage points. Compared with Graph-Constraint, the improvement of our method is mainly reflected in the two indicators of boundary satisfaction and morphological deviation, which indicates that although Graph Constraint alone can improve neighborhood coordination, it is still difficult to give more refined guidance results at the plot level without prototype reasoning and candidate screening. Prototype-search performs well in Prototype matching, but local volume shifts still occur on complex parcels due to insufficient integration of boundary and road constraints. The proposed method can control the building height gradient, roof relationship and basement retreat while maintaining the continuity of block style, so the overall output is closer to the planning guidance goal.

Table 3: Overall performance comparison of different planning guidance methods

Method	Guidance Consistency / %	Boundary Satisfaction / %	Block Coordination / %	Morphological Deviation / %
Rule-Mapping	84.9	86.2	82.7	23.6
MLP-Guidance	88.3	89.1	86.5	20.8
Graph-Constraint	90.1	91.4	88.7	19.6
Prototype-Search	90.8	90.2	89.5	19.1
Proposed Method	91.6	93.3	90.9	18.7

Table 4 shows that the comprehensive guidance effect of historical street and lane plots is the highest, especially the interface continuity reaches 95.4%, indicating that the model is highly sensitive to the low-level continuous street wall, roof relief and street and lane scale. The boundary satisfaction of the main road interface plot reaches 93.6%, which reflects that the constraint coding module can well absorb the conditions such as road red line, open interface and building retreat. The interface continuity of the waterfront open plots is slightly lower, which is related to the fact that the waterfront scene itself puts more emphasis on openness and spacing changes, but it still remains above 89.8% overall. The indicators of the updated mixed plots are relatively balanced, which indicates that the model can maintain a stable balance between style continuation and plot adaptation when dealing with the areas where new and old buildings coexist and the volume difference is large.

Table 4: Comparison of the guided results of the proposed method under different plot types

Parcel Type	Guidance Consistency / %	Height Matching Degree / %	Setback Satisfaction / %	Interface Continuity / %
Historic Street-Block Parcels	92.4	93.1	94.2	95.4
Main Road Frontage Parcels	91.8	92.7	93.6	94.1
Waterfront Open Parcels	90.7	91.9	92.4	89.8
Renewal Mixed Parcels	91.6	92.3	93.0	91.2

Table 5 shows the ablation results of different modules. After removing the morphological prototype generation, the guided consistency decreases to 88.9%, indicating that if the block style vector cannot be compressed into a stable prototype, the continuous style reference will be lost in the plot-level output. After removing the constraint coding module, the boundary satisfaction decreases most significantly, which is only 88.7%, and the shape deviation increases to 22.1%, which indicates that the constraints such as the red line of the plot, the road boundary back and the distance between adjacent buildings have a direct impact on the guidance results. After removing the candidate screening module, although the indicators are still higher than some comparison models, the overall accuracy is significantly reduced, which indicates that the consistency and feasibility screening after generating multiple candidate schemes is a

necessary step to control the volume relationship and interface order stably. The full model remains optimal in all four indicators, which also further verifies the synergy between the three-part modules.

Table 5: Comparison of ablation experiment results for different modules

Model Configuration	Guidance Consistency / %	Boundary Satisfaction / %	Block Coordination / %	Morphological Deviation / %
Full Model	91.6	93.3	90.9	18.7
Without Morphological Prototype Generation	88.9	90.5	87.8	21.4
Without Constraint Encoding Module	89.4	88.7	88.5	22.1
Without Candidate Screening Module	90.2	90.8	89.1	20.3

Table 6 shows that although the number of parameters of the proposed method is higher than MLP-Guidance, the average inference time is controlled at 29 ms, which is still at a high response level and better than Graph-Constraint and Prototype-Search. Rule-Mapping is the fastest, but its output depends on preset rules, which cannot adapt to diverse plot forms. The proposed method maintains a good tradeoff among parameter scale, video memory footprint, and inference speed, indicating that the morphological prototype and constraint encoding module do not impose an unacceptable computational burden. For planning assistance scenarios, this level of response time is sufficient to support batch plot analysis and interactive scheme comparison.

Table 6: Comparison of computational efficiency versus response performance of different methods

Method	Number of Parameters	Average Inference Time / ms	Memory Usage / MB	Single-Parcel Response Level
Rule-Mapping	0	8	42	Very High
MLP-Guidance	186432	24	318	High
Graph-Constraint	254780	31	406	Medium-High
Prototype-Search	301556	37	455	Medium
Proposed Method	276914	29	392	High

Comprehensive experimental results show that the proposed architectural form planning guidance method can effectively convert the block style recognition results into block-level control suggestions, and achieve stable performance in the dimensions of guidance consistency, boundary satisfaction, block coordination and form deviation. The heat map results show that the model has good adaptation ability in different plot scenarios. The table comparison results further show that morphological prototype generation, constraint coding and candidate screening together constitute the core source of planning guidance performance improvement. It can be seen that the guidance mechanism based on morphological representation learning can not only improve the computational consistency of planning output, but also provide a reusable technical basis for subsequent urban design support and scheme generation.

4 Conclusion

Focusing on the tasks of feature extraction of city block style and architectural form planning guidance, this paper constructs an integrated computing framework consisting of multi-source sample representation, deep feature extraction and morphological representation learning. The experimental results show that the Accuracy of the model in the city block style recognition task reaches 94.1%, Macro-F1 reaches 0.912, and Kappa coefficient reaches 0.894. The overall recognition performance is better than Baselin-CNN, Single-Scale Encoder and Dual-Branch Net, and the validation loss remains stable in the later stage of training, which indicates that the joint modeling of street view images, remote sensing Windows and building shape indicators can form a more reliable style representation.

In the plot scale building form planning guidance, the guidance consistency of the proposed method reaches 91.6%, the boundary satisfaction reaches 93.3%, the block coordination reaches 90.9%, and the shape deviation is controlled at 18.7%. Compared with Rule-Mapping, MLP-Guidance, Graph-Constraint and Prototype-Search, the proposed method shows more stable output ability in complex plot boundaries and mixed block scenarios. The results of heat map and ablation experiments show that morphological prototype generation, constraint coding and candidate screening can jointly enhance the continuity and adaptability of the plot level control results, and form a clearer correspondence between building height, interface retreat, roof organization and volume combination.

However, this paper still has some limitations. The boundary division of block samples, the accuracy of style labels and the quality of multi-source data registration will directly affect the stability of morphological representation. Local misalignment or label coasing will weaken the ability of prototype vectors to describe the real characteristics of blocks. In the updated mixed area and high-density interleaved area, architectural style fracture, volume jump and road interface mutation are more concentrated, and the differences between candidate schemes are sometimes compressed to a small range, which affects the discrimination of the ranking results. Although the model maintains a good overall performance in the cross-zone mixed samples, the differences in road scale, development intensity, facade update frequency and control rules in different cities still bring disturbance to the migration results. The existing framework is mainly based on two-dimensional street view, remote sensing window and structural morphological indicators, and it is not enough to absorb three-dimensional spatial relationships, continuous changes in time and textual planning constraints, which also makes the model still have room to continue to expand in more complex urban design scenarios.

The self-supervised pre-training and cross-city adaptation mechanism can be introduced in the future research to enhance the generalization ability of the model to few-sample regions and heterogeneous scenes. Temporal street view, 3D point cloud and planning text constraints were incorporated to improve the dynamic response ability of morphological reasoning. The combination of lightweight deployment and interactive parameter update enables the system to serve more frequent urban design support and planning review processes.

References

[1] Wang L, Han X, He J, et al. Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 215-230.

[2] Yan Y, Huang B. Estimation of building height using a single street view image via deep neural networks[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 192:

83-98.

[3] Pang H E, Biljecki F. 3D building reconstruction from single street view images using deep learning[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 112: 102859.

[4] Cai J, Chen Y. A novel unsupervised deep learning method for the generalization of urban form[J]. Geo-Spatial Information Science, 2022, 25(4): 568-587.

[5] Najmi A, Gevaert C M, Kohli D, et al. Integrating remote sensing and street view imagery for mapping slums[J]. ISPRS International Journal of Geo-Information, 2022, 11(12): 631.

[6] Wu J, Lu Y, Gao H, et al. Cultivating historical heritage area vitality using urban morphology approach based on big data and machine learning[J]. Computers, environment and urban systems, 2022, 91: 101716.

[7] Xu Y, He Z, Xie X, et al. Building function classification in Nanjing, China, using deep learning[J]. Transactions in GIS, 2022, 26(5): 2145-2165.

[8] Chen F C, Subedi A, Jahanshahi M R, et al. Deep learning–based building attribute estimation from google street view images for flood risk assessment using feature fusion and task relation encoding[J]. Journal of Computing in Civil Engineering, 2022, 36(6): 04022031.

[9] Zhang L, Wang L, Wu J, et al. Decoding urban green spaces: Deep learning and google street view measure greening structures[J]. Urban Forestry & Urban Greening, 2023, 87: 128028.

[10] He J, Zhang J, Yao Y, et al. Extracting human perceptions from street view images for better assessing urban renewal potential[J]. Cities, 2023, 134: 104189.

[11] Ki D, Chen Z, Lee S, et al. A novel walkability index using google street view and deep learning[J]. Sustainable Cities and Society, 2023, 99: 104896.

[12] Kang Y, Abraham J, Ceccato V, et al. Assessing differences in safety perceptions using GeoAI and survey across neighbourhoods in Stockholm, Sweden[J]. Landscape and Urban Planning, 2023, 236: 104768.

[13] Patel P, Kalyanam R, He L, et al. Deep learning-based urban morphology for city-scale environmental modeling[J]. PNAS nexus, 2023, 2(3): pgad027.

[14] Lu Y, Wei W, Li P, et al. A deep learning method for building façade parsing utilizing improved SOLOv2 instance segmentation[J]. Energy and Buildings, 2023, 295: 113275.

[15] Xu H, Sun H, Wang L, et al. Urban architectural style recognition and dataset construction method under deep learning of street view images: a case study of Wuhan[J]. ISPRS International Journal of Geo-Information, 2023, 12(7): 264.

[16] Sun H, Xu H, He H, et al. A Spatial analysis of urban streets under deep learning based on street view imagery: quantifying perceptual and elemental perceptual relationships[J]. Sustainability, 2023, 15(20): 14798.

[17] Zhai Y, Gong R, Huo J, et al. Building façade color distribution, color harmony and diversity in relation to street functions: using street view images and deep learning[J]. ISPRS International Journal of Geo-Information, 2023, 12(6): 224.

[18] Ogawa Y, Oki T, Zhao C, et al. Evaluating the subjective perceptions of streetscapes using street-view images[J]. Landscape and Urban Planning, 2024, 247: 105073.

[19] Wu C, Wang J, Wang M, et al. Machine learning-based characterisation of urban morphology with the street pattern[J]. Computers, environment and urban systems, 2024, 109: 102078.

[20] Fleischmann M, Arribas-Bel D. Decoding (urban) form and function using spatially explicit deep learning[J]. Computers, Environment and Urban Systems, 2024, 112: 102147.