A Scene Graph Generation Method for Historical District Street-view Imagery: A Case Study in Beijing, China

Xian Guo[1], XuanDi Liu[1], Jie Jiang[1]

[1] School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing 102616, China (guoxian@bucea.edu.cn, lxd386469007@163.com, jiangjie@bucea.edu.cn)

Keywords: Scene Graph, Street-view, Historical Districts, Image Captioning.

Abstract

Using street-view imagery for interpreting diverse street-scale elements and their relationships within historical districts offers high efficiency and low cost for preservation and management. Scene graphs provide a structured representation of objects and their relationships within a scene. However, applying existing scene graph generation techniques directly to street-view imagery presents challenges due to the complexity of elements and narrow street spaces. This paper introduces HSSGG (Historical Street-view Scene Graph Generation), a predictive model that effectively identifies elements and their relationships. By incorporating an end-to-end Relation Transformer with the parameter-free attention and coordinate attention modules, HSSGG improves relationship prediction accuracy, even with limited samples, and enhances the precision of scene graph generation in complex environments. Test on 200 panoramic images from historical districts in Beijing shows that HSSGG outperforms existing single-stage relation prediction models (such as RelTR and FCSGG) in accuracy and stability. These results provide valuable insights for the preservation and management of historical districts.

1. Introduction

Comprehensive extraction and precise representation of landscape elements at a fine-grained district scale are essential for informed urban management (Ranzato, 2017)and resource optimization (Yi et al., 2022), thereby fostering the sustainable development of historical districts (Li et al., 2022). Comparing to modern urban environments, historical districts are characterized by unique landscape elements, including distinctive attributes (such as color, texture, and structure) (Yang et al., 2024), spatial patterns (Yang et al., 2023), and complex interrelationships. Predicting both the location and category of these elements, while also considering their semantic connections, presents significant challenges, particularly in the context of historical and cultural preservation.

Street-view imagery offers significant advantages, such as rich facade information, ease of data acquisition, and strong timeliness, making it a valuable resource for urban scene analysis (Gong et al., 2018). Though previous studies on street view datasets have made incredible progress in feature extraction (Xiong et al., 2021; Yu and Ji, 2022), scene segmentation (Jiang et al., 2023), and target detection (Wu et al., 2020; Xiong et al., 2021; Yu and Ji, 2022), they often fall short in providing deep semantic understanding of these images, particularly within historical context. Scene graph generation (SGG) is a key task in semantic scene understanding, closely linked to visual relationship detection. In SGG, each image is represented by a graph where nodes represent entities and edges denote the relationships between them. At present, SGG has been applied to urban quantification, space perception, and socioeconomic prediction, achieving progress in fine-grained classification of urban functional areas and multidimensional evaluation of the ‘physical + social’ urban environment.

However, historical districts present unique challenges due to the multitude of landscape elements from various periods and styles, which are irregularly distributed, leading to highly complex relationships. Additionally, the narrow street spaces in these districts lead to significant occlusions and varying lighting conditions in street-view imagery. The dense street layout, with tightly packed structures, complicates detection and recognition as elements often overlap or blend into the background. Shadows from buildings and trees reduce visibility, while direct sunlight causes overexposure, obscuring key features of landscape elements. These challenges make historical districts difficult for scene analysis, emphasizing the need for robust models. Eeffective analyzing historical districts not only supports preservation efforts but also highlights the robustness and versatility of the developed approach to a broad range of urban environments. More importantly, existing annotated datasets, such as Visual Genome (VG) (Krishna et al., 2017)and COCO(Lin et al., 2014), are designed for modern streets and lack the necessary descriptions for landscape elements and their relationships in historical districts. As a result, directly applying existing SGG methods to historical districts is challenging, as it involves accurately identifying and modeling the intricate relationships between landscape elements in highly variable environments, and overcoming the inadequacy of current datasets to capture the unique characteristics of historical landscapes.

In response to these unique challenges in scene relationship modeling within historical districts, this paper introduces a fully annotated historical street-view imagery dataset and the HSSGG (Historical Street-view Scene Graph Generation) model, which is designed to identify landscape elements and their relationships within historical districts with only visual appearance. Specifically, the model employs an end-to-end relationship prediction framework, the Relation Transformer(Cong et al., 2023), to extract deeper semantic and contextual information in historical districts. Additionally, Parameter-Free Attention module (SimAM) (Yang et al., 2021)and Coordinate Attention module (CoordAtt) (Hou et al., 2021)are integrated to enhance the applicability of scene modelling in historical street scenarios. The key contributions of this work include:

1. The construction of a fully annotated historical street-view imagery dataset (i.e., historical street dataset), presenting the unique categories and relationships of landscape elements.

2. The development of the HSSGG model, providing an efficient solution for scene interpretation and relationship modeling in historical districts by deeply mining semantic and contextual information, even with limited training data.

3. The challenges presented by historical districts provide a rigorous testing for scene generation model, where overcoming obstacles related to landscape elements, lighting, and object occlusion would indicate the model’s capacity for broader and more reliable application across various urban landscapes.

2. Study Area and Data

2.1 Study Area and Historical Street Dataset

Beijing, the capital of five imperial dynasties (Liao, Jin, Yuan, Ming, and Qing) and the current capital of China, is located north of the North China Plain, covering an area of 16,410.54 km^2. Its traditional residential areas, recognized as a World Cultural Heritage site, hold immense historical and cultural value. The local government has invested substantial efforts in preserving and managing these areas.

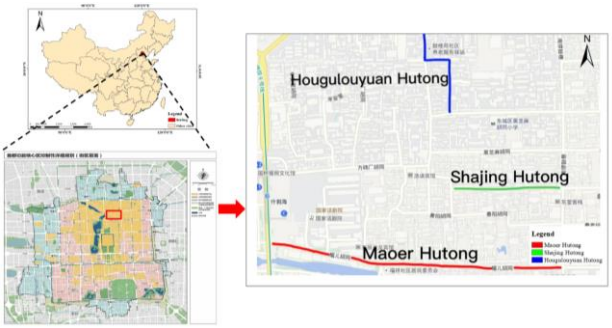

Figure 1. Study area

Existing public datasets lack specialized research focusing on historical districts. To bridge this gap, we collected and constructed a historical street dataset for Beijing's historical district, aimed at supporting the automated analysis of landscape elements and their relationships within complex traditional urban environments. Hougulouyuan Hutong, Shajing Hutong, and Mao'er Hutong, located in the northwest of Dongcheng District, are among the best-preserved ancient alleys in Beijing, characterized by traditional hutong features in both element types and spatial patterns. We collected 200 high-resolution panoramic street-view imageries (5760×2880) from these study areas using the Insta360 One X2 panoramic camera.

Figure 2. Label the categories of images using X-AnyLabeling and generate annotation files in COCO format.

The annotations are in the COCO dataset format, including the image names, bounding boxes, and category IDs, as shown in Figure 2-3. The rel.json file is the annotation file for relationships. The rel_categories field represents the relationship categories.

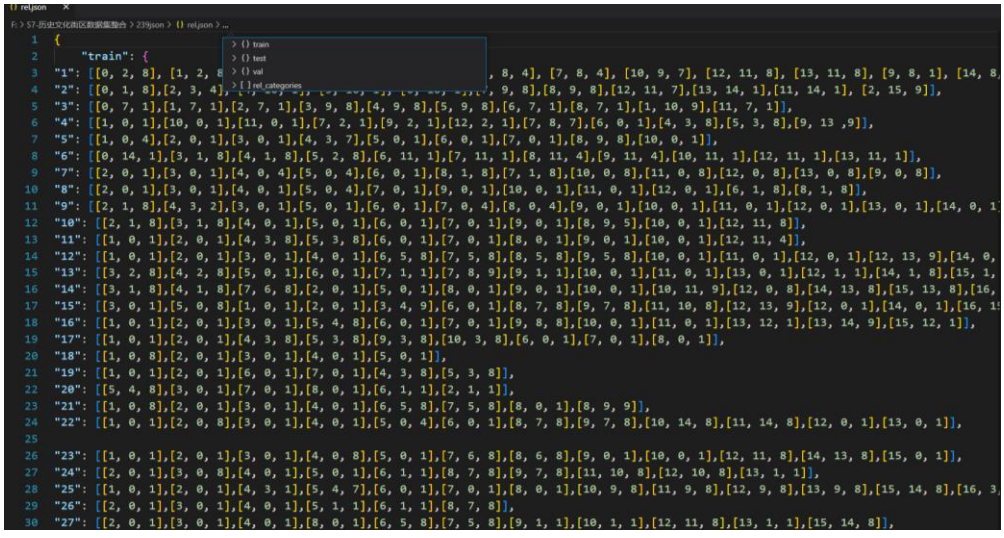

Figure 3. The `rel.json` file contains four fields: `train`, `val`, and `test`, which represent the relationship triplets for the training, validation, and test sets, respectively. The `rel_categories` field represents the relationship categories.

Notably, we further expanded the existing VG and COCO category system to accommodate the unique landscape elements in historical districts, adding 11 categories of historically and culturally significant elements in Hutongs (Table 1) and categorizing relationships into 8 types (Table 2). Our dataset includes both natural elements and artificial elements, such as manhole cover, plaque, and miscellaneous items. By presenting the unique categories and spatial distributions of landscape elements, the developed historical street dataset provides essential data for understanding the living environment of the historical districts. This dataset is the only dataset with annotated relationships in a historical street context, and will be made open source soon.

Class Name	Description
plaque	The plaque mounted on the building's façade
potted plant	The potted plant along the street
promotional signage	The promotional signage displayed outside
traditional window	The traditional window with intricate wooden carvings
traditional door	The traditional door leading into the historical building
electric bicycle	The electric bicycle parked near the sidewalk
monitor	The monitor mounted on a streetlight or wall
debris	Debris scattered or piled on the street
manhole cover	The manhole cover embedded in the street
air conditioning	The air conditioning unit mounted on the exterior wall
traditional decoration	Traditional items decorating the street

Table 1. Unique object categories in the historical street dataset compared to the COCO dataset and VG dataset.

Relationship categories	Examples
on the side of	traffic sign on the side of street
on the top of	plaque on the top of wall
near	debris near streetlight
walk on	people walk on street
ride	people ride electric bicycle
stuck on	promotional signage stuck on wall
hanging from	monitor hanging from streetlight
on	traditional window on wall

Table 2. The eight relationship categories in the historical street dataset

3. Methodology

As shown in Figure 4, the HSSGG model is built on the base of Relation Transformer for Scene Graph Generation (RelTR) architecture. As the representative of state-of-the-art SGG method, RelTR employs coupled subject and object encoding to learn queries, effectively capturing dependencies between relationships and contextual features. These features are then decoded into 'subject-predicate-object' triplets using feed-forward networks (FFNs) alongside attention heatmaps. To address the unique challenges of historical district analysis, the HSSGG model introduces SimAM and CoordAtt into the RelTR model.

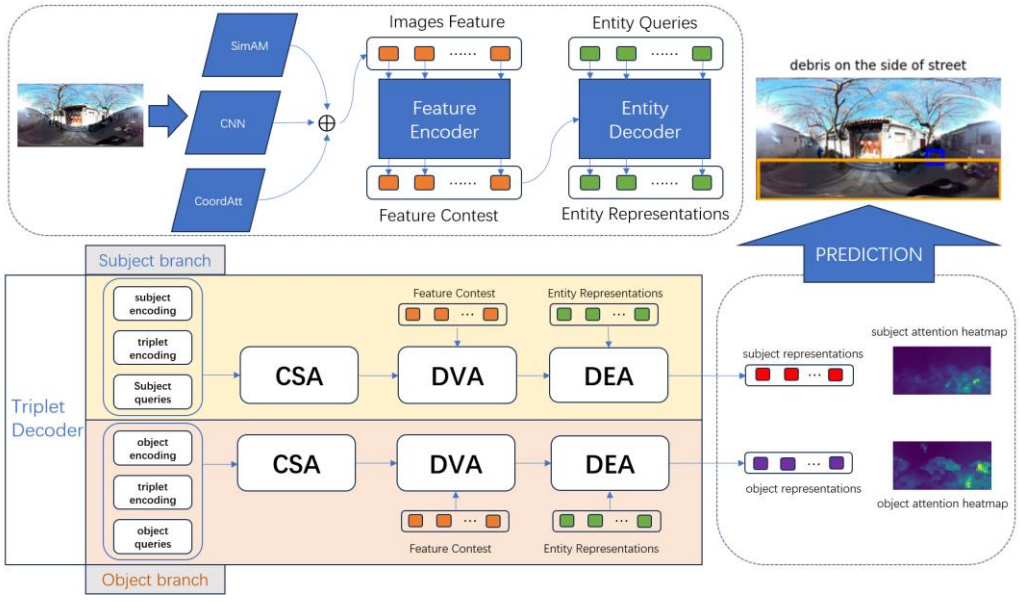

Figure 4. HSSGG Model Architecture Diagram. CSA, DVA and DEA stand for Coupled Self-Attention, Decoupled Visual Attention and Decoupled Entity Attention $\oplus$ indicates element-wise addition.

3.1 The SimAM module

SimAM module calculates attention weights by treating the entire feature map as a whole, evaluating the importance of each neuron through its energy function, thereby enabling the model to identify key local features of landscape elements in historical street scenes. SimAM assesses neuron importance by defining an energy function and deriving its closed-form solution. The simplified expression for the energy function is as follows:

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{\mu})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

Here, t represents the activation value of the target neuron; $\hat{\mu}$ is the mean activation value of all neurons in the current channel; $\hat{\sigma}^2$ represents the variance of the activation values of all neurons in the current channel; and λ is a regularization parameter used to control model complexity and prevent overfitting.

The importance of each neuron can be calculated as $\frac{1}{e_t^*}$, and the attention weights are adjusted through the Sigmoid function, resulting in the optimized feature representation:

$$\tilde{X} = sigmoid\left(\frac{1}{E}\right) \odot X \tag{2}$$

Where E represents the combined energy values across channel and spatial dimensions, and $\odot$ denotes element-wise multiplication, X the input feature.

3.2 The CoordAtt module

The CoordAtt module embeds spatial position information into channel attention to capture long-range dependencies after performing SimAM, thereby enhancing the model's sensitivity to spatial positioning within historical districts.

This module firstly decomposes global pooling operation into two 1-D feature encoding processes, aggregating features along horizontal and vertical directions, respectively. Given the input X, two spatial pooling kernels $(H, 1)$ and $(1, W)$, are used to encode each channel along the horizontal and vertical coordinates, respectively. The output of the c-th channel at height h and width w is formulated as:

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i) \tag{3}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w) \tag{4}$$

Where $z_c^h(h)$ and $z_c^w(w)$ represents the output at height h and width w for the c-th channel, respectively. The aggregated feature maps from both directions are then concatenated and passed through a shared 1×1 convolutional transformation function F_1 to generate direction-aware feature maps, expressed as:

$$f = \delta\left(F_1([z^h, z^w])\right) \tag{5}$$

where$[\cdot,\cdot]$denotes the concatenation operation along the spatial dimensions, δ is a non-linear activation function and f is the direction-aware feature map.

The feature map f is then split into two separate tensors, f^h and f^w. Attention weights g^h and g^w are generated through 1×1 convolution operations (F_h and F_w) with Sigmoid function σ:

$$g^h = \sigma\left(F_h(f^h)\right) \tag{6}$$

$$g^w = \sigma\left(F_w(f^w)\right) \tag{7}$$

The attention weights g^h and g^w are then applied to each channel of the input features, allowing the features at each spatial position to be adjusted according to global information. The final coordinate attention output is given by:

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \tag{8}$$

In summary, the SimAM module focuses on capturing spatial details, enabling the model to identify key local features in complex scenes, while the CoordAtt module complements this by considering both local and global information during relationship prediction. Additionally, this complementary fusion of information improves prediction accuracy and consistency without adding computational costs, making the model particularly well-suited for analysing the complex spatial relationships in historical districts.

4. Results

4.1 Implementation Details

Based on the developed historical street dataset, we trained the HSSGG model for 500 epochs on a single RTX 2080 Ti GPU. The batch size for the model was set to 2, with a weight decay of 10^{-5}, and clipping the gradient norm>0.05. The initial learning rates for both the Transformer and ResNet50 backbone networks were set to 10^{-4}, and the learning rates are dropped by 0.1 after 50 epochs. To mitigate overfitting on the small-sized training samples, the number of encoder and decoder layers in the Transformer was set to 3, with the triplet (decoder) and entity (decoder) layers configured to match.

To enhance the performance of the triplet decoder, the auxiliary loss was incorporated into the model. The multi-head attention module, consisting of 8 heads, was trained with a dropout rate of 0.1, and the model dimension d was set to 256. For all experiments, the number of entity queries N_e and coupled queries N_t was set to 50 and 100, respectively, while the IoU threshold in the triplet assignment was 0.7.

For fair comparison, the parameter settings of the RelTR model were kept consistent with those of the HSSGG model, while the FCSGG model was configured according to its recommended settings, optimized for the Visual Genome (VG) dataset. This experiment employed evaluation metrics commonly used in natural image scene graph generation, including Recall@K (R@K) and mean Recall@K (mR@K). To better estimate the model performance on the historical street dataset, we divided the historical district dataset into 145 images for training, 40 images for test, and 15 images for the validation.

4.2 Quantitative Results and Comparison

As shown in Table 3, the RelTR (first row) outperforms the FCSGG (second row) in the historical district scenario, with metrics R@20, R@50, and R@100 in scene graph detection exceeding those of FCSGG by approximately 0.61. The mR@20, mR@50, and mR@100 metrics also demonstrate an

improvement of about 0.14 to 0.15, suggesting RelTR has stronger generalization capability in the relationship prediction task. However, the proposed HSSGG (third row), an improved version of RelTR that incorporates SimAM and CoordAtt to better handle complex district scenes with limited training samples, outperforms both RelTR and FCSGG across all metrics. Specifically, HSSGG outperforms RelTR by 2.48 and 3.38 in R@50 and R@100, respectively, and shows improvements of 0.19 to 0.18 in mR@20, mR@50, and mR@100, achieving the best performance. These results highlight the superior ability of HSSGG to model scene relationships within historical districts while maintaining the lightweight advantages of RelTR..

Method	Scene Graph Detection (SGDET)					
	R@20	R@50	R@100	mR@20	mR@50	mR@100
RelTR	1.92	1.92	1.92	0.81	0.87	0.96
FCSGG	1.31	1.31	1.31	0.67	0.73	0.81
HSSGG	**1.92**	**4.40**	**5.30**	**1.00**	**1.06**	**1.14**

Table 3. Performance comparison of the HSSGG with RelTR and FCSGG in terms of R@K and mR@K

Layer Number		Scene Graph Detection (SGDET)		
Encoder	Triplet Decoder	R@20	R@50	R@100
3	3	1.92	**4.40**	**5.30**
4	4	2.54	3.49	3.49
5	5	1.92	1.92	1.92
6	6	2.54	3.21	4.12
3	6	1.58	2.21	2.21
9	9	**2.81**	2.81	2.81

Table 4. Impact of the number of encoder and decoder layers on the performance

Ablation Setting			SGDET			Params (M)
RelTR	Simam	Coordatt	R@20	R@50	R@100	
√	√	×	2.12	3.86	3.86	**36.35**
√	×	√	2.54	4.68	4.68	36.74
√	×	×	1.92	1.92	1.92	**36.35**
√	√	√	**1.92**	**4.40**	**5.30**	36.75

Table 5. The impact of the modules on model performance, where √ indicates the module is activated, and × indicates it is deactivated.

Query ID	181	199	50	31	117	33	28
Captioning	Streetlight on the side of street	Streetlight on the side of street	Debris on the side of street	Debris on the side of street	Monitor hanging from streetlight	Debris on the side of street	Monitor hanging from streetlight

Table 6. The relevant descriptive text corresponding to the Query ID, with correct text descriptions highlighted in red.

4.3 Ablation Studies

In the ablation studies, we examined the impact of various factors on the model's final performance. All the ablation studies were performed with proposed historical district dataset.

4.3.1 Number of Layers: The number of feature encoder layers and triplet decoder layers significantly influences the model's performance. As shown in Table 4, when the number of triplet decoder layers is set to 3, the model achieves better results, with R@20, R@50, and R@100 scores of 1.92, 4.40, and 5.30, respectively. However, increasing the number of decoder layers to 9 leads to a decline in performance, with R@20, R@50, and R@100 scores dropping to 2.81, which might be attributed to overfitting.

4.3.2 Module Effectiveness: To evaluate the contribution of each module, we conducted ablation studies by deactivating different modules, with the results summarized in Table 5. Initially, we assessed the baseline RelTR model on the historical street dataset. Despite its competitive performance in VG datasets, RelTR struggles in the historical district scenario, achieving suboptimal R@20, R@50, and R@100 scores of only 1.92, when trained on a small dataset.

We then explore the impact of the SimAM and CoordAtt modules individually. When the SimAM module was added to the RelTR, it did not increase the number of network parameters. Moreover,

this module enhances the model’s ability to capture spatial details by computing attention weights across the entire feature map, focusing on key local features. As shown in Table 5 (the first row), activating only the SimAM module significantly improves performance, with an R@20 score of 2.12 and a substantial increase in R@50 and R@100 scores to 3.86, approximately doubling the performance of the baseline RelTR.

Next, we assessed the CoordAtt’s effect. As shown in Table 5 (the second row), CoordAtt significantly enhances direction-aware and position-sensitive information, leading to improved scene graph quality. The CoordAtt alone improves R@20 by approximately 0.4 and R@50 and R@100 by around 0.8 compared to the SimAM-only model. Notably, this enhancement is achieved with only a minimal increase in parameters.

Finally, when both SimAM and CoordAtt are activated, the model’s complexity slightly increases. However, due to the lightweight design of both modules, the parameter increase remains minimal. The combined HSSGG model outperforms all comparison models in relationship prediction accuracy, particularly in the R@50 and R@100 metrics, achieving the highest recorded values to date, with scores of 4.40 and 5.30, respectively (see the fourth row of Table 5). These results underscore the practical value of the synergistic effect between the SimAM and CoordAtt modules.

4.4 Qualitative Results

Figure 5 visualized a selected scene from the historical street datasets. The model parameter settings are detailed in Section 4.1, where both the feature encoder and triplet decoder layers were set to 3. Due to space constraints, only high confidence samples are presented in Figure 6. Blue boxes are the subject boxes while orange boxes are the object boxes. We can generate a correct scene graph from the above seven triplets, as shown in Figure 7.

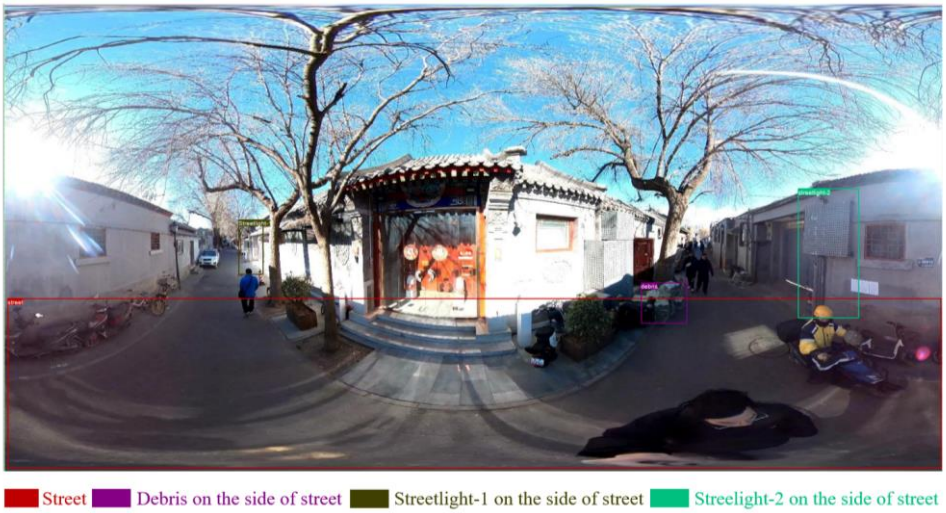

Figure 5. Ground truth annotations of a selected scene from the historical street datasets.

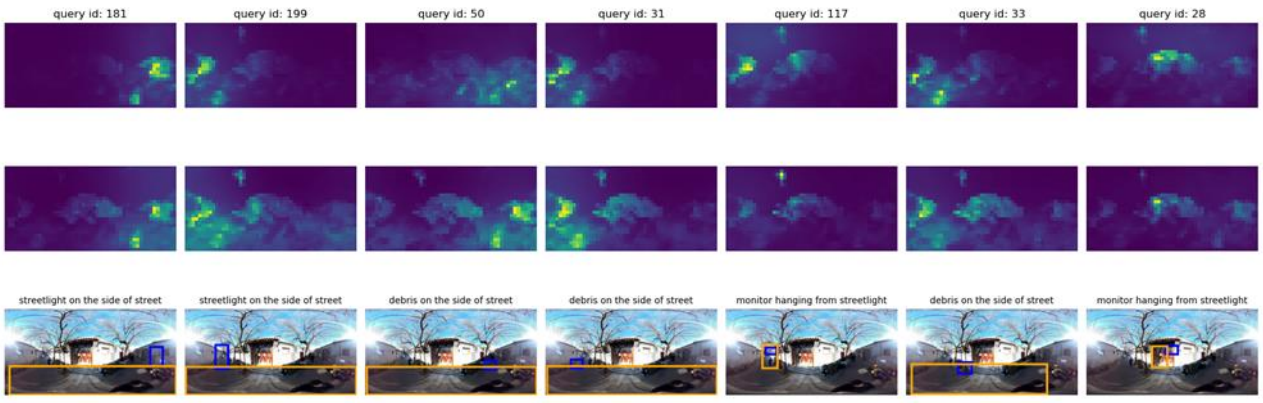

Figure 6. Prediction results of the relationships between landscape elements in Figure 5. Each column presents a specific query ID, with the top two rows displaying the heatmaps generated by HSSGG. These heatmaps illustrate the attention distribution during the queries, where brighter areas (highlighted in green) indicate regions most relevant to the query.

Each output image consists three components: the query ID, a heatmap, and the corresponding scene description. In Figure 6, each column presents a specific query ID, with the top two rows displaying the heatmaps generated by HSSGG. These heatmaps illustrate the attention distribution during the queries, where brighter areas (highlighted in green) indicate regions most relevant to the query. For example, in the case of query ID 181, the model correctly identified and focused on the "streetlight" as delineated by the bright spot in the heatmap.

In the scene description associated with the panoramic image, seven triplets were generated, four of which were correctly

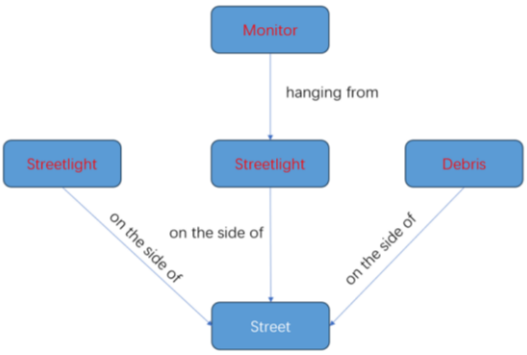

Figure 7. The scene graph constructed from the correct triplets in Figure 6

identified (highlighted in red in Table 6). These triplets include the subjects 'debris,' 'streetlight,' 'street,' and 'monitor'. As shown in Table 6, the HSSGG performs well under challenging historical scenes (e.g., poor lighting, panoramic image distortion, a wide variety of objects, and severe occlusion). The proposed model successfully generates sentences that recognize the location and category of multiple landscape elements in the scene, while describing relationships between them in the context of historical and cultural preservation. Besides commonly used triplets< streetlight on the side of street >, It should be noted that the prediction of the triplets, such as <monitor hanging from streetlight> indicates that the HSSGG model is capable of correctly understanding complex landscapes in the historical districts, despite 'hanging from' and 'monitor' being infrequent terms in the dataset. This indicates that our method not only effectively identifies common features but also demonstrates superior performance in recognizing certain unique features. From the visualization results, it is evident that the HSSGG model maintains a high level of recognition accuracy even with just over a hundred training samples. As the training samples in the proposed dataset expand, the HSSGG is expected to reach even higher accuracy levels in the future.

5. Conclusions and Future Work

This paper introduces a novel scene graph generation method specifically designed for historical districts. By integrating the SimAM and CoordAtt modules into the base RelTR mode, the proposed HSSGG effectively capturing the attributes and spatial distribution relationships of landscape elements at the street-level. Based on the historical street dataset, experimental results demonstrate that HSSGG can accurately predict objects and their interrelationships within scenes, even under limited samples. The model effectively describes conventional street-level element relationships while also capturing the unique relational characteristics in historical districts. Comparative experiments with state-of-the-art models reveal the effectiveness of the HSSGG in predicting element relationships within historical districts.

Future work will focus on expanding and refining the historical district dataset, as well as incorporating additional attention modules and prior knowledge to enhance the model's performance under limited training conditions.

Acknowledgements

The research was supported by the The National Key Research and Development Program of China (No. 2021YFB3900904).

References

Cong, Y., Yang, M.Y., Rosenhahn, B., 2023. RelTR: Relation Transformer for Scene Graph Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 11169-11183.

Gong, F.-Y., Zeng, Z.-C., Zhang, F., Li, X., Ng, E., Norford, L.K., 2018. Mapping sky, tree, and building view factors of street canyons in a high-density urban environment. *Building and Environment*, 134, 155–167.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate Attention for Efficient Mobile Network Design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13708-13717.

Jiang, B., An, X., Xu, S., Chen, Z., 2023. Intelligent Image Semantic Segmentation: A Review Through Deep Learning Techniques for Remote Sensing Image Analysis. *J Indian Soc Remote Sens*, 51(9), 1865–1878.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L., 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int J Comput Vis*, 123(1), 32–73.

Li, J., Nassauer, J.I., Webster, N.J., 2022. Landscape elements affect public perception of nature-based solutions managed by smart systems. *Landscape and Urban Planning*, 221, 104355.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context, *Computer Vision – ECCV 2014*, *Lecture Notes in Computer Science*. Springer International Publishing, Cham, 740–755.

Ranzato, M., 2017. Landscape elements as a basis for integrated water management. *Urban Water Journal*, 14(7), 694–703.

Wu, Y., Zhang, H., Li, Y., Yang, Y., Yuan, D., 2020. Video Object Detection Guided by Object Blur Evaluation. *IEEE Access*, 8, 208554–208565.

Xiong, S., Tan, Y., Li, Y., Wen, C., Yan, P., 2021. Subtask Attention Based Object Detection in Remote Sensing Images. *Remote Sensing*, 13(10), 1925.

Yang, H., Wu, R., Qiu, B., Zhang, Z., Hu, T., Zou, J., Wang, H., 2024. The next step in suburban rural revitalization: Integrated whole-process landscape management linking ecosystem services and landscape characteristics. *Ecological Indicators*, 162, 111999.

Yang, H., Xu, W., Yu, J., Xie, X., Xie, Z., Lei, X., Wu, Z., Ding, Z., 2023. Exploring the impact of changing landscape patterns on ecological quality in different cities: A comparative study among three megacities in eastern and western China. *Ecological Informatics*, 77, 102255.

Yang, L., Zhang, R.-Y., Li, L., Xie, X., 2021. SimAM: A Simple, Parameter-Free Attention Module for Convolutional Neural Networks, *Proceedings of the 38th International Conference on Machine Learning. Presented at the International Conference on Machine Learning*, PMLR, 11863–11874.

Yi, L., Ma, S., Tao, S., Zhang, J., Wang, J., 2022. Coastal

landscape pattern optimization based on the spatial distribution heterogeneity of ecological risk. *Front. Mar. Sci*., 9, 1003313.

Yu, D., Ji, S., 2022. A New Spatial-Oriented Object Detection Framework for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sensing*, 60, 1–16.