Journal of Engineering, Project, and Production Management
2026, 16(6), 2025-223

3D Extraction of Urban Heritage Elements Using Polarized Self-Attention and Enhanced Neural Radiance

Xiaofei Chen[1] and Liqun Guo[2]
[1] Instructor, School of Art and Design, Wuhan Institute of Technology, Wuhan, 430205, China, E-mail xiaofeichen0666@126.com (corresponding author).
[2] Professor, School of Art and Design, Wuhan Institute of Technology, Wuhan, 430205, China

Project Management
Received October 3, 2025; revised December 28, 2025; accepted June 9, 2026
Available online June 21, 2026

Abstract: With the growing need for precise extraction and digital documentation of urban historical landscape elements, such as ancient building brackets, city wall textures, traditional street layouts, and stone inscription patterns, this study proposes a novel extraction algorithm based on improved polarized self-attention and enhanced neural radiance fields. It tackles technical challenges by jointly optimizing 3D scene representation and fine-grained feature extraction, enabling accurate and robust element identification in complex environments. Tests on real-world urban historical landscapes show that the model outperforms existing methods. For 3D digital documentation of ancient building brackets, it achieves an average error of 0.9 to 1.8mm, detail completeness of 93.8% to 99.8%, and texture clarity above level 4.2. In wall texture recognition, accuracy exceeds 96.9%, with lesion localization errors of 1.2 to 2.8cm. The method significantly improves extraction precision and anti-interference capability. It not only extracts elements accurately but also generates 3D semantic models, offering core technical support for historical landscape restoration and urban heritage management. This provides a new technological pathway for digital preservation and sustainable urban development.

Keywords: Polarized self-attention mechanism, neural radiance fields, urban historical landscape element extraction, three-dimensional scene representation, multi-source data.

DOI 10.32738/JEPPM-2025-223

1. Introduction

Rapid urban expansion poses unprecedented challenges to the protection and inheritance of historical urban landscapes. The urban historical landscape system encompasses multi-dimensional indicators, including architectural styles, cultural heritage, and ecological environments, among which complex interactions and dynamic changes exist (Zhou et al., 2023). Therefore, the precise identification and extraction of urban historical landscape elements are of significant importance for preserving urban characteristics, promoting cultural heritage, and achieving sustainable development (Yarza Pérez and Verbakel, 2025). However, traditional image analysis methods and single deep learning models often encounter issues, like inaccurate feature extraction and limited recognition accuracy, when processing high-resolution, multi-source, heterogeneous urban historical landscape data (Hasanvand et al., 2023).

Neural Radiance Fields (NeRF) represent a method for 3D scene representation and rendering. It utilizes a deep neural network to learn the continuous volumetric density and color radiance of a scene, thereby finding widespread applications in 3D scene reconstruction (Wang et al., 2023; Zhu et al., 2025). For instance, Huang et al. (2024) addressed the issue of devices inability to effectively render high dynamic range scenes by proposing a NeRF-based method. Experimental validation demonstrated that the proposed method not only accurately synthesized views but also rendered them naturally. Sun et al. (2023) aimed to further enhance the practical value of NeRF and introduced a framework integrating a Transformer with NeRF. This framework achieved its goals by identifying target 3D bounding boxes using NeRF, and the results indicate that the proposed framework outperformed traditional methods. To address degraded quality in novel view synthesis caused by severe geometric constraints in NeRF, Chen et al. (2023) proposed a solution based on StructNeRF. Experimental results showed that the proposed solution outperformed other methods.

The Polarized Self-Attention (PSA) mechanism can adaptively focus on key information within input data and serves as an attention mechanism in deep learning for processing image and sequential data (Qi et al., 2024; Feng et al., 2023). Wang et al. (2024) addressed the challenge of ship visual detection in sea fog environments by proposing a detector based

on a PSA and a pyramid network, trained on datasets. Experiments demonstrated that the proposed detector exhibited superior performance metrics. Liu et al. (2023) targeted the complex computations in pose estimation networks by proposing a lightweight pose estimation network based on PSA, which reduced feature loss. Experiments indicated that the proposed pose estimation network achieved high accuracy.

In summary, NeRF exhibits issues of low efficiency and susceptibility to overfitting, while PSA, although capable of adaptively focusing on key information, overly relies on local information (Jing et al., 2023). Therefore, this study optimizes PSA by introducing Atrous Spatial Pyramid Pooling (ASPP) and then integrates it with Improved Neural Radiance Fields (INeRF), ultimately resulting in a method for extracting urban historical landscape elements based on INeRF-Improved Polarized Self-Attention (IPSA). This study aims to enhance the model's ability to capture the multi-dimensional features of urban historical landscapes, thereby enabling precise identification and extraction of landscape elements. The innovation of the research is that, for the first time, polarized self-attention is combined with NeRF to extract historical landscape elements, and the advantages of INeRF in three-dimensional scene reconstruction and the feature-focusing ability of IPSA are combined to achieve accurate identification of landscape elements. The integration of these two approaches not only improves the accuracy of feature extraction but also enhances the model's adaptability to complex changes in urban environments, providing scientific evidence and technical support for the protection and planning of urban historical landscapes.

2. Methods and Materials

2.1. Construction of a 3D Scene Representation Model for Urban Historical Landscapes Based on INeRF

The 3D representation of urban historical landscapes offers an intuitive perspective, holding significant importance for the protection and preservation of urban cultural heritage, as well as for enhancing the quality of urban planning and design. NeRF technology demonstrates notable advantages in constructing 3D scene representation models. With its ability to synthesize new viewpoint images from sparse views, NeRF enables high-quality 3D reconstruction without the need for complex geometric modeling, making it suitable for handling complex, detail-rich scenes such as urban historical landscapes (Zhang et al., 2024; Guo et al., 2024). The operation of NeRF is illustrated in Fig. 1.

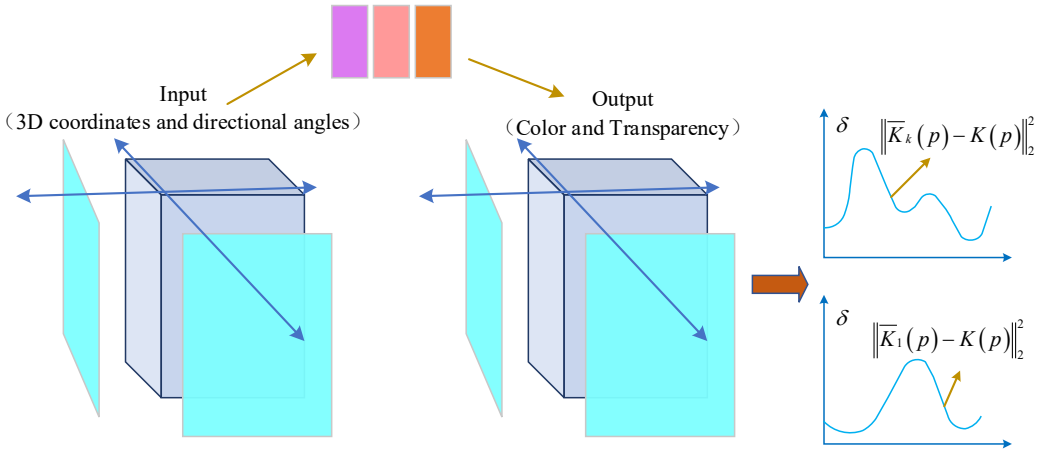


Fig. 1. NeRF's operational diagram

In Fig. 1, the input images are processed by a Multi-layer Perceptron (MLP), which outputs the volume density and color information of the scene. This information is utilized in the volume rendering process, where new viewpoint images are synthesized by calculating the interaction between light rays and the scene. Subsequently, by comparing the rendered and real images, the loss function is computed to guide the optimization of model parameters. This process enables the precise reconstruction of the 3D structure and appearance of urban historical landscapes, achieving high-quality 3D scene representation (Han et al., 2023; Naumann et al., 2024). Features are encoded using sine and cosine functions of different frequencies to capture their periodic and multi-scale characteristics, as shown in Eq. (1).

$$M(m)=\left(\sin\left(2^0\pi m\right),\cos\left(2^0\pi m\right),\ldots,\sin\left(2^{n-1}\pi m\right),\cos\left(2^{n-1}\pi m\right)\right) \tag{1}$$

In Eq. (1), $M(m)$ denotes a certain feature representation of the input parameter m (such as the position or orientation of a point in the scene). The n denotes the length of the feature vector, which determines the richness of features input into the neural network. The MLPs are interconnected through weights and introduce nonlinearity via activation functions, thereby aiding NeRF in processing data with complex spatial or temporal structures (such as sequential data or images) (Chen et al., 2023). The MLP structure receives multi-view image data through its input layer, extracts key features in the hidden layers, and generates precise 3D scene representations through the output layer. These representations encompass both geometric and appearance attributes of the scene, thus providing high-quality 3D reconstruction and visualization means for the digital preservation and display of cultural heritage (Yao et al., 2023; Zhou et al., 2023). This

feature representation, serving as input to the deep neural network, facilitates the learning of the scene's volume density and color radiance, enabling high-quality 3D scene rendering that synthesizes new views from sparse views. Subsequently, the color along the ray path from the camera to a given point in the scene is calculated by integrating the color contributions and density information at each point along the ray path to generate the final image, as shown in Eq. (2).

$$K(p)=\int_{x_0}^{x_1} X(x)\delta\left(p(x)\right)k\left(p(x),b\right)dx \tag{2}$$

In Eq. (2), $K(p)$ represents the true color along the ray p emanating from the camera, which is the color observed from the data. $\int_{x_0}^{x_1}$ denotes the integral of the ray from the near endpoint x_0 to the far endpoint x_1. $\delta\left(p(x)\right)$ is the volume density along the ray p at position x, serving as a measure of the material distribution in the scene. $X(x)$ is the transmittance from the starting point of the ray to position x, indicating the proportion of light that has not been absorbed from the starting point to position x. $k\left(p(x),b\right)$ represents the color radiance along the ray p at position x. b is the viewing direction. The transmittance $X(x)$ is used to adjust the color contribution along the ray path to reflect the attenuation of light in the medium, as shown in Eq. (3).

$$X(x)=\exp\left(-\int_{x_0}^{x}\delta\left(p(v)\right)dv\right) \tag{3}$$

In Eq. (3), $\delta\left(p(v)\right)$ represents the volume density at position v along ray p. It serves as a metric for the material distribution within the scene, influencing the propagation and scattering of light. dv is the infinitesimal distance element, indicating the infinitesimal path length along ray p. Subsequently, by integrating the color and density information along the ray path and calculating the cumulative final color along ray p through summation, realistic reconstruction and rendering of complex scenes are achieved, as shown in Eq. (4).

$$\overline{K}_1(p)=\sum_{s=1}^{S} X_s\beta_s k_s,\ \beta_s=1-e^{-\varepsilon_s\delta_s} \tag{4}$$

In Eq. (4), $\overline{K}_1(p)$ represents the foreground color calculated along ray p starting from the camera, which is the color predicted by the model. β_s is the color contribution coefficient of the s th sampling point. ε_s is the sampling interval of the s th sampling point, namely the distance from this point to the next sampling point. By minimizing the loss function, the predicted color is made to approximate the truly observed color, thereby enhancing the model's accuracy in representing the 3D scenes of urban historical landscapes and generating more realistic rendered images, as shown in Eq. (5).

$$J=\sum_{p\in Z}\left(\left\|\overline{K}_k(p)-K(p)\right\|_2^2+\left\|\overline{K}_1(p)-K(p)\right\|_2^2\right) \tag{5}$$

In Eq. (5), Z denotes the set of all rays. $\overline{K}_k(p)$ represents the background color calculated along ray p. $\|\ \|_2^2$ is the square of the Euclidean norm, which is used to compute the distance between vectors. NeRF is widely applied in the reconstruction of 3D scenes. However, when the scene is relatively complex and the amount of data is limited, NeRF may overfit the training data. Meanwhile, the rendering quality of NeRF is highly dependent on the sampling strategy (Qu et al., 2024; Zhou et al., 2023). Therefore, this study introduces the Stochastic Structural Similarity Loss (S3IM) to optimize NeRF, and the operation of INeRF is illustrated in Fig. 2.

In Fig. 2, minimizing the S3IM loss enables the model to learn a more realistic scene representation, thereby generating higher-quality rendered images of 3D scenes. The calculation formula for S3IM is shown in Eq. (6).

$$\mathrm{S3IM}\left(\overline{A},A\right)=\frac{1}{C}\sum_{c=1}^{C}\mathrm{SSIM}\left(g^{(c)}\left(\overline{K}\right),g^{(c)}\left(K\right)\right) \tag{6}$$

In Eq. (6), $\mathrm{S3IM}\left(\overline{A},A\right)$ represents the stochastic structural similarity metric between the image $\overline{A}$ rendered by the model and the real image A. $\frac{1}{C}$ denotes the average value obtained from A sampling results, which is used to reduce randomness and enhance the stability of evaluation. $g^{(c)}\left(\overline{K}\right)$ is the c th image patch randomly sampled from the image $\overline{K}$ rendered by the model. $g^{(c)}\left(K\right)$ is the c th image patch randomly sampled from the real image K.

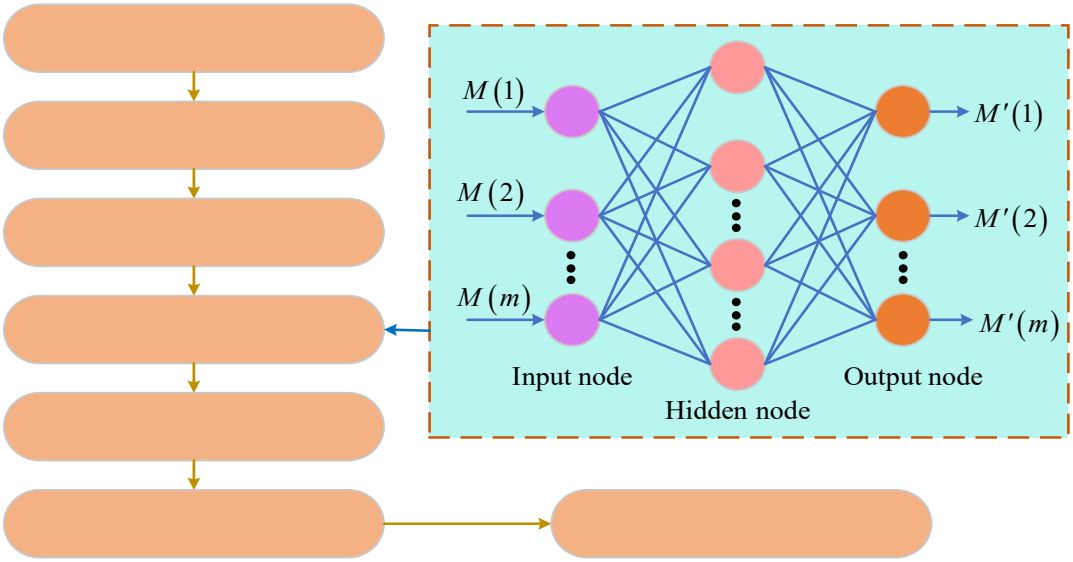


Fig. 2. Operation of IneRF

Finally, the model's performance is optimized by combining the traditional loss function J and the structural similarity loss $J_{S3IM}\left(\overline{A}\right)$, as shown in Eq. (7).

$$J_{all}\left(\overline{A}\right)=J+\eta J_{S3IM}\left(\overline{A}\right) \tag{7}$$

In Eq. (7), η serves as a weight parameter, which is used to balance the contributions between the traditional loss function J and the structural similarity loss $J_{S3IM}\left(\overline{A}\right)$. $J_{all}\left(\overline{A}\right)$ represents the overall loss function of the model, which combines the traditional loss function and the structural similarity loss for minimization during the training process. In summary, the 3D scene representation model for urban historical landscapes based on INeRF is shown in Fig. 3.

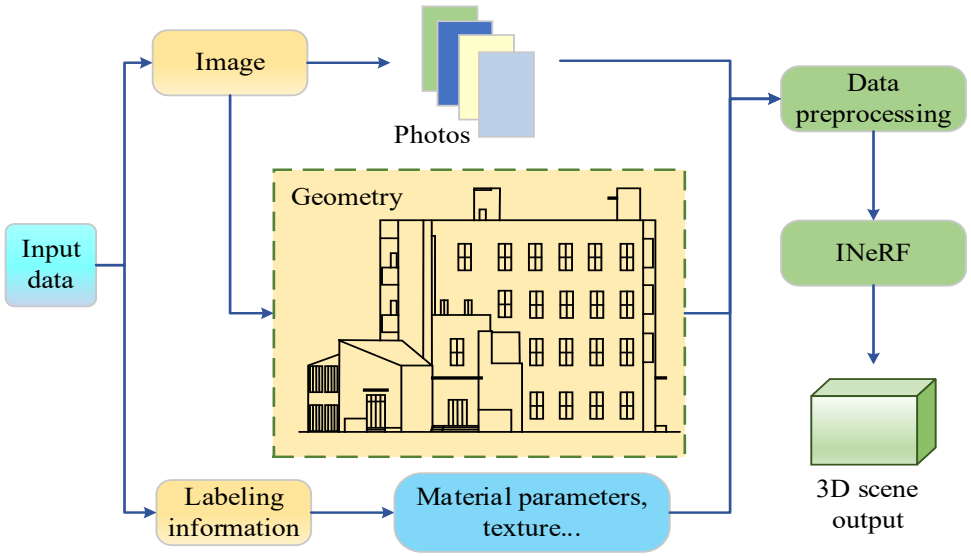


Fig. 3. Flowchart for the 3D scene representation model of urban historical landscapes

As can be observed from Fig. 3, multi-view image data is first collected and preprocessed. Subsequently, feature extraction techniques are applied to obtain key information. Next, the NeRF model is initialized, and the S3IM loss function is used to evaluate the structural similarity between the generated and actual images. Then, the model is trained using multi-view images to minimize the loss. Ultimately, an accurate 3D scene representation of urban historical landscapes is achieved.

2.2. Construction of an Urban Historical Landscape Element Extraction Method Based on INeRF-IPSA

The 3D scene representation model, built on INeRF, accurately captures the spatial structure and appearance details of urban historical landscapes, providing an intuitive, rich 3D data foundation for element extraction. However, when dealing

with fine-grained elements in historical landscapes, such as the wood texture of ancient buildings and the patterns on stone inscriptions, relying solely on the geometric and appearance representation of the scene makes it difficult to efficiently focus on and precisely capture these key features. In contrast, PSA can adaptively focus on information across both channel and spatial dimensions of feature maps, selectively strengthening key regional features while suppressing redundant information, thereby further addressing the shortcomings of INeRF in fine-grained feature extraction (Wu et al., 2023). The operation of PSA is shown in Fig. 4.

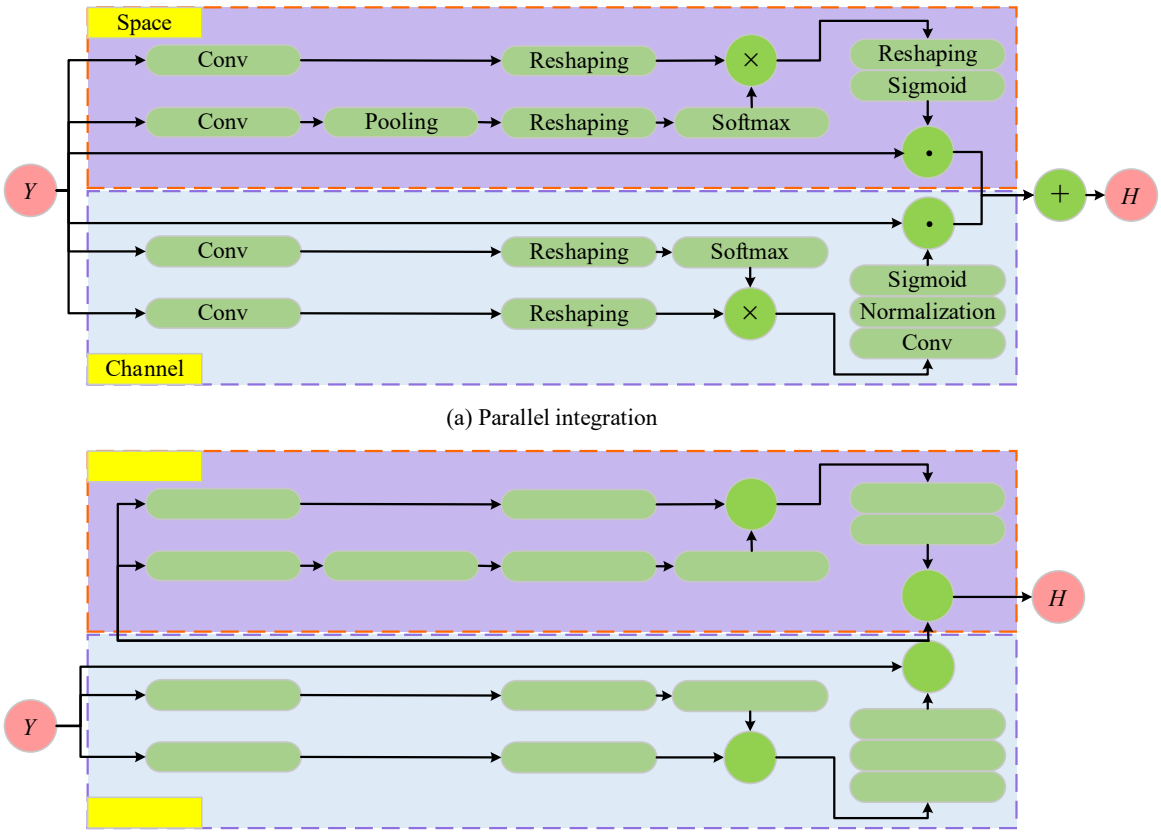


Fig. 4. PSA operating chart

In Fig. 4(a), PSA enhances feature representation capability by combining spatial attention and channel attention. Spatial attention and channel attention are typically computed in parallel, and then their outputs are combined to obtain richer feature representations (Jamshidi et al., 2023). This combination enables the model to consider both channel and spatial information of features simultaneously, thereby improving its performance and generalization (Xie et al., 2024). Fig. 4(b) illustrates series fusion, which helps the model optimize features through stepwise processing. Channel attention allows the model to adaptively focus on different channels in the input feature map, thereby enhancing key features while suppressing unimportant ones, as shown in Eq. (8).

$$R_{ch}(Y) = f_{\text{Sigmoid}}\left[Q_{H|\alpha_1}\left(t_1\left(Q_a(Y)\right) * \overline{f}_{\text{Softmax}}\left(t_2\left(Q_d(Y)\right)\right)\right)\right] \tag{8}$$

In Eq. (8), $R_{ch}(Y)$ represents the channel attention of the input feature map Y. f_{Sigmoid} is the Sigmoid activation function, which maps the input to the [0,1] interval and calculates the attention weights. $Q_{H|\alpha_1}$ is a weight matrix with parameter α_1, used to map the channel attention to the feature space. $t_1\left(Q_a(Y)\right)$ is obtained by performing a 1×1 convolution operation Q_a on the input feature map Y and then conducting tensor reconstruction using the reconstruction operator t_1 to adapt to subsequent operations. $\overline{f}_{\text{Softmax}}$ is the Softmax normalization function, which converts the input into a probability distribution in the attention mechanism. $t_2\left(Q_d(Y)\right)$ is also obtained by performing a 1×1 convolution operation Q_d on the input feature map Y and then conducting tensor reconstruction using the reconstruction operator t_2 to adapt to subsequent operations. Additionally, the purpose of spatial attention is to enable the model to adaptively focus on different spatial positions in the input feature map, thereby enhancing key features and suppressing unimportant ones, as shown in Eq. (9).

$$R_{sp}(Y) = f_{\text{Sigmoid}}\left[t_3\left(\overline{f}_{\text{Softmax}}\left(t_1\left(f_{pooling}\left(Q_d(Y)\right)\right)\right)\right) * t_2\left(Q_a(Y)\right)\right] \tag{9}$$

In Eq. (9), $R_{sp}(Y)$ represents the spatial attention output of the input feature map Y. $f_{pooling}$ denotes the global pooling operation, which is used to aggregate spatial information within the feature map. t_3 is the reconstruction function, employed to adjust the dimensions or format of the feature map so that it is suitable for subsequent multiplication operations. Subsequently, by fusing channel and spatial attention in parallel, the model can more effectively utilize the information within the feature map, enabling a more comprehensive understanding of the input feature map while capturing both global and local feature information, as shown in Eq. (10).

$$F_{Pa}(Y) = I_{ch} + I_{sp} = R_{ch}(Y) \underset{ch}{\odot} Y + R_{sp}(Y) \underset{sp}{\odot} Y \tag{10}$$

In Eq. (10), $F_{Pa}(Y)$ represents the feature map Y that has undergone processing through the PSA mechanism. I_{ch} and I_{sp} are the feature maps enhanced by channel and spatial attention, respectively. $\underset{ch}{\odot}$ denotes the element-wise multiplication along the channel dimension, used to apply channel attention weights to each channel of the feature map. $\underset{sp}{\odot}$ is the element-wise multiplication along the spatial dimension, used to apply spatial attention weights to each spatial location of the feature map. In series fusion, channel attention first adjusts the channel weights of the feature map, and then spatial attention further adjusts the result along the spatial dimension. This step-by-step processing aids the model in gradually optimizing feature representations and better capturing the hierarchical structure of features, as shown in Eq. (11).

$$F_{co}(Y) = I_{sp}(I_{ch}) = R_{sp}\left(R_{ch}(Y) \underset{ch}{\odot} Y\right) \underset{sp}{\odot} R_{ch}(Y) \underset{ch}{\odot} Y \tag{11}$$

In Eq. (11), $F_{co}(Y)$ represents the series fusion operation. In the feature enhancement module, the Sigmoid function is utilized to adjust the importance or weights of features, as shown in Eq. (12).

$$f_{\text{Sigmoid}}(Y) = \frac{1}{1+e^{-Y}} \tag{12}$$

In Eq. (12), the Sigmoid function enables better capture of crucial information within the input data and demonstrates greater flexibility and adaptability when managing complex tasks. Additionally, when the function values approach 0 or 1, its derivative approaches 0. The derivative of the Sigmoid function facilitates stable convergence of the network, as shown in Eq. (13).

$$f_{\text{Sigmoid}}'(Y) = \frac{e^{-Y}}{\left(1+e^{-Y}\right)^2} \tag{13}$$

In Eq. (13), the derivative of the Sigmoid function is not only used for gradient calculation but also enables self-adjustment by updating weights during the backpropagation process. PSA exhibits deficiencies in capturing features at different scales and may encounter issues such as information loss or inadequate feature representation when processing images with complex structures. These problems restrict the model's performance in recognition and classification tasks. Therefore, to address these issues, the study introduces ASPP to enhance the model's perception of multi-scale features. By employing pooling operations at different scales, ASPP captures richer contextual information, thereby improving the robustness and accuracy of feature representation (Matos et al., 2024). The structure of ASPP is shown in Fig. 5.

In Fig. 5, the ASPP module captures features of different scales and densities in an image by applying multiple dilated convolutional layers with different dilation rates in parallel, along with a global average pooling layer. This ability to extract multi-scale features is crucial for enhancing the performance of the feature enhancement module, especially when dealing with images that have complex structures and varying scales (Das et al., 2024). Through this approach, ASPP optimizes PSA, enabling it to process and understand input data more effectively, as shown in Eq. (14).

$$G_o = G_{o-1} + (j-1) * H_{o-1} \tag{14}$$

In Eq. (14), G_{o-1} represents the feature maps at the $o-1$ th layer before processing by the ASPP module, and these feature maps have undergone processing through the PSA mechanism. G_o denotes the feature maps at the o th layer after processing by the ASPP module, which enhances the model's ability to capture multi-scale information by combining features from different scales and sampling rates. j is the size of the convolution kernel. H_{o-1} refers to the feature maps at the $o-1$ th layer after specific processing, where information from different scales is captured through different branches and then combined to enhance the feature representation. By calculating the output dimensions of the feature maps after dilated convolution, the module can adapt to input feature maps of varied sizes and flexibly adjust the receptive field size, as shown in Eq. (15).

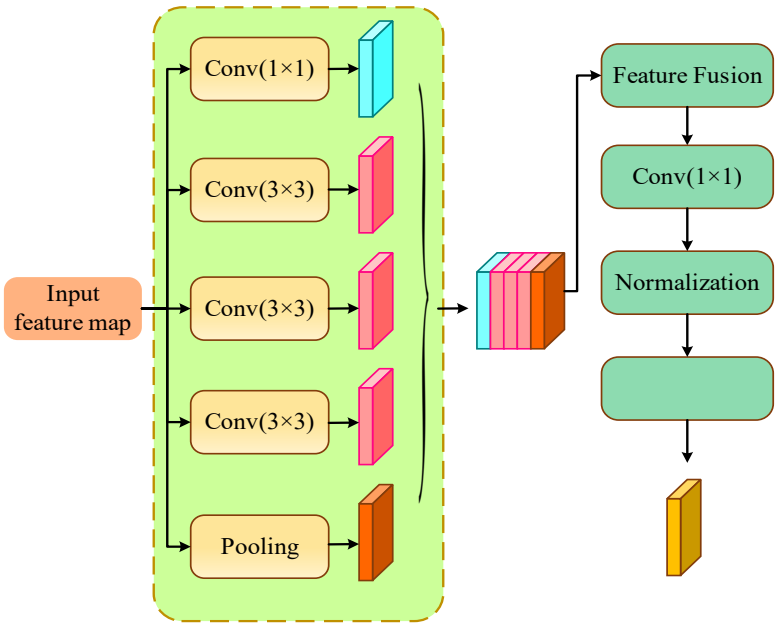


Fig. 5. ASPP structure diagram

$$G=(u-1)*(j-1)+j \tag{15}$$

In Eq. (15), u represents the dilation rate, which defines the spacing between elements in the convolution kernel. It enables the capture of a broader range of contextual information by increasing the receptive field of the convolutional kernel without adding to the number of parameters. Therefore, the operation of the IPSA-based feature enhancement module is illustrated in Fig. 6.

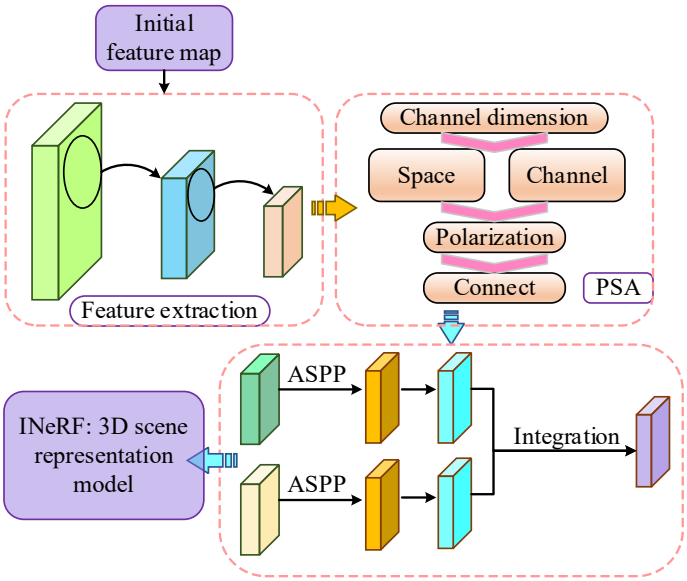


Fig. 6. Feature enhancement module flowchart

In Fig. 6, optimizing dilated convolution enhances the capability to capture multi-scale information. Meanwhile, during the feature fusion stage, the PSA module is introduced to perform weighted processing across both channel and spatial dimensions, further improving the expressive power of the features. Consequently, the study constructs a 3D scene representation model for urban historical landscapes using INeRF, clearly restoring the spatial structure and appearance details of the landscapes. Additionally, by optimizing PSA with ASPP, a feature enhancement module capable of accurately capturing fine-grained features is developed, addressing the deficiencies of traditional methods in feature focusing and multi-scale perception. To meet the requirements of urban historical landscape element extraction tasks in practical operations, the study proposes an urban historical landscape element extraction method based on INeRF-IPSA, as shown in Fig. 7.

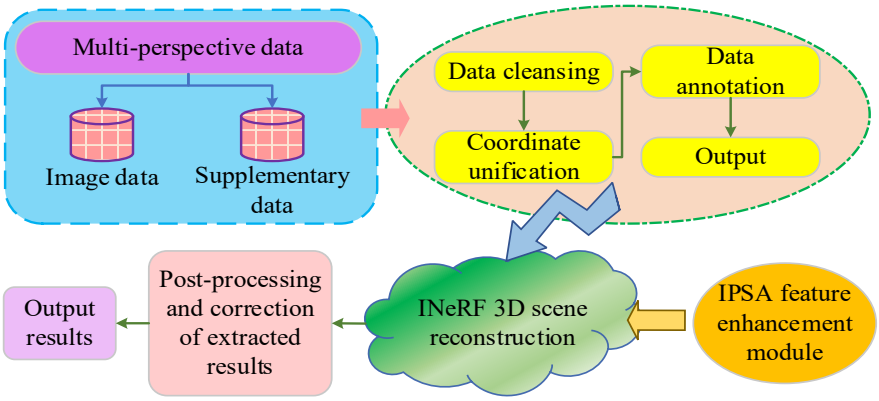


Fig. 7. Flowchart of the INeRF-IPSA-based method for extracting urban historical landscape features

In Fig. 7, through rational process design, the 3D scene data provides spatial support for feature extraction, while the enhanced feature information, in turn, facilitates the accurate identification and localization of elements within the scene, forming a complete technical closed loop that spans from scene modeling to feature optimization and then to element extraction. This method can provide a three-dimensional model with millimeter-level accuracy, clearly restore component details and disease locations, and accurately meet the data needs of heritage protectors during the restoration process. At the same time, the output three-dimensional semantic model captures the patterns of historical streets, lanes and architectural features and can be seamlessly imported into planning and design software, providing urban planners with an intuitive and efficient design reference. In addition, for inscriptions and ancient building components, this method can achieve high-precision texture restoration with high texture clarity. It also has text recognition capabilities that can quickly assist in creating digital collections and support online exhibitions and academic research in cultural museums. In practical applications, this method only requires the collection of multi-view images to quickly produce 3D models and feature data. It can replace traditional manual surveying and mapping, shorten the survey cycle and improve data accuracy. INeRF-IPSA provides objective data support for protection level assessment and restoration priority sorting by quantitatively outputting indicators such as completeness of details and degree of weathering. In addition, this method can dynamically monitor the restoration effect and changes in the status of landscape elements, enabling timely adjustment of protection measures. The output 3D semantic model can be widely used across scenarios such as archiving, restoration, planning and display, without the need to repeatedly collect data, thereby reducing overall costs and improving work efficiency. Therefore, protection engineers can use this system to efficiently detect structural deterioration and quickly identify damaged parts and their severity through accurate 3D models and quantitative indicators, and then prioritize repair work and rationally allocate resources to ensure the stability and safety of key structures. At the same time, the system's real-time monitoring function enables engineers to continue tracking the restoration's progress, ensuring the quality and efficiency of the work and further improving the overall effectiveness of heritage protection. In summary, INeRF-IPSA uses quantitative data and visual models to help different users simplify decision-making, reduce operational costs, and truly realize the deep integration of technology and heritage management practices.

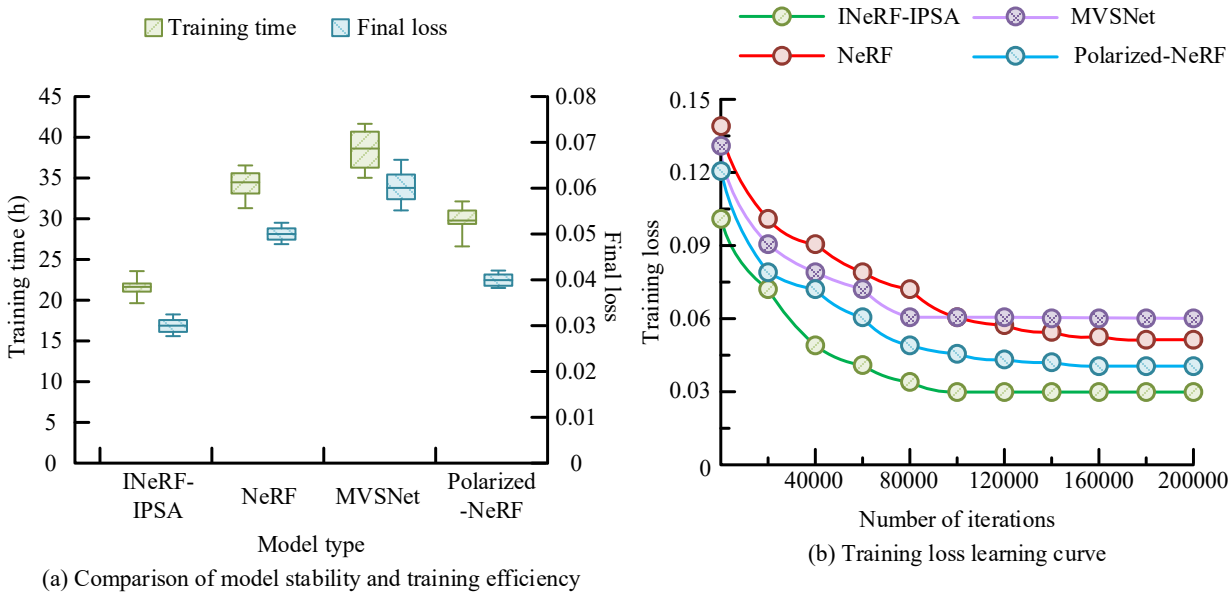
3. Results

3.1. Performance Validation and Comparative Analysis of the INeRF-IPSA Model

To validate the performance and superiority of the proposed INeRF-IPSA model, a comprehensive comparative analysis was conducted from multiple dimensions, including model convergence efficiency, element extraction accuracy, and 3D reconstruction quality. To ensure the scientific rigor and meticulousness of the experiments, three currently mainstream and representative models were selected as comparison benchmarks: the NeRF model, the Multi-view Stereo Network (MVSNet) model, and the Polarized Neural Radiance Fields (Polarized-NeRF) model. The ETH3D dataset was chosen for the experimental data, as it contained high-resolution architectural and outdoor scenes, making it suitable for fine-grained extraction and 3D reconstruction tasks of urban historical landscape elements. During data processing, the selected image sequences underwent uniform preprocessing, including image size normalization, illumination correction, and camera parameter calibration, to ensure consistent input data. All experiments were conducted on a server equipped with an NVIDIA RTX 3090 GPU, and the testing platform was built based on the PyTorch deep learning framework. The study initially focused on the training efficiency and stability of the models, verifying the robustness and efficiency of the INeRF-IPSA model by evaluating its performance across multiple independent training sessions. The specific test results are shown in Fig. 8.

As shown in Fig. 8(a), the INeRF-IPSA model demonstrated significant advantages in training efficiency and convergence stability. Across multiple experiments, its median training time was only 21.5 hours, substantially lower than that of Polarized-NeRF (29.8 hours), NeRF (34.2 hours), and MVSNet (38.5 hours). Meanwhile, the INeRF-IPSA model exhibited the smallest range for both the box and whiskers, with its median final loss value stabilizing at an extremely low level of 0.030. In contrast, the median final loss values for Polarized-NeRF, NeRF, and MVSNet were 0.040, 0.050, and 0.060, respectively, with more dispersed data distributions. This indicated that the INeRF-IPSA model not only converged

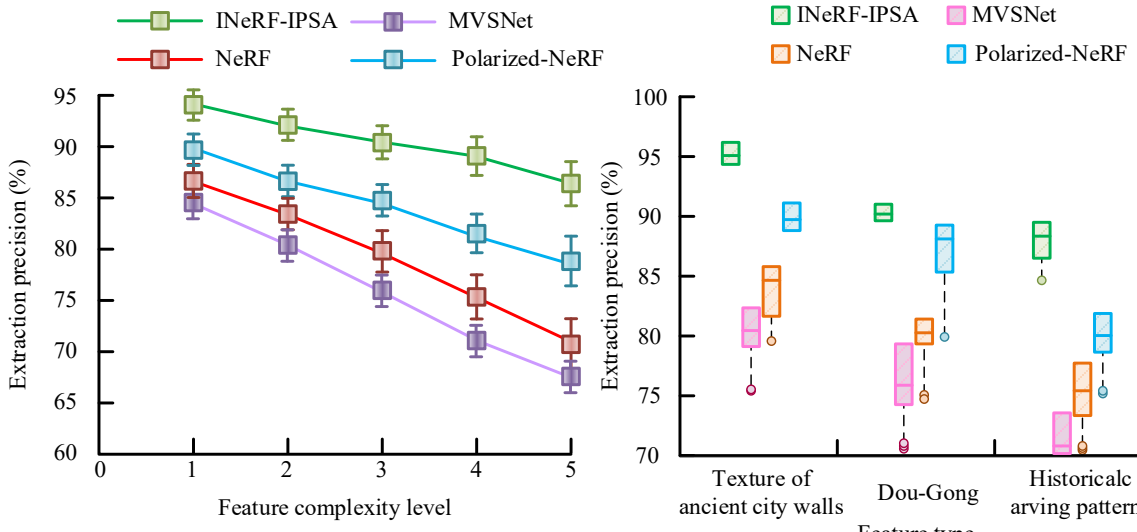
faster but also consistently achieved higher accuracy, demonstrating exceptional robustness. As illustrated in Fig. 8(b), during the initial training phase, the loss values of all models declined rapidly, but the curve for the INeRF-IPSA model dropped more steeply, indicating a faster convergence trend. Ultimately, the loss value of the INeRF-IPSA model stabilized at approximately 0.03, whereas the Polarized-NeRF, MVSNet, and NeRF models converged to approximately 0.04, 0.06, and 0.05, respectively. This demonstrated that the INeRF-IPSA model not only converged rapidly but also achieved lower loss values, highlighting its higher accuracy and stronger robustness in extracting elements from urban historical landscapes. The study further validated the extraction capabilities of each model for fine-grained elements, such as ancient architectural bracket sets, ancient city wall textures, and historical stele patterns, and quantified the impact of dynamically changing element complexity on each model's extraction accuracy. The specific test results are shown in Fig. 9.



(a) Comparison of model stability and training efficiency

(b) Training loss learning curve

Fig. 8. Comparison of training efficiency and stability across models



(a) Precision curves for fine-grained feature extraction of different models at various complexity levels

(b) Precision performance of models

Fig. 9. Comparison of fine-grained feature extraction performance among models

As shown in Fig. 9(a), the INeRF-IPSA model ranked first in pixel-level semantic matching accuracy across all element complexity levels from one to five. At level one, it achieved 94.2%, and at level five, it maintained 86.3%, representing only a 7.9% decline. In contrast, Polarized-NeRF, NeRF, and MVSNet experienced declines of 11.0%, 16.0%, and 16.8%, respectively, with lower accuracy at each complexity level compared to INeRF-IPSA. Additionally, the INeRF-IPSA model exhibited the shortest error bars, indicating optimal performance stability. As illustrated in Fig. 9(b), the INeRF-IPSA model achieved median accuracies of 90.1%, 95.0%, and 88.0% for extracting ancient architectural bracket sets, ancient city wall textures, and historical stele patterns, respectively, with the most concentrated data distribution. MVSNet demonstrated the lowest median accuracy across all elements and had the highest number of outliers. Polarized-NeRF and NeRF fell between

the two in terms of accuracy and data stability, fully validating the superiority of INeRF-IPSA in fine-grained element extraction. The study further conducted a quantitative evaluation of the reconstruction quality of each model in complex environments. This set of experiments aimed to assess the performance of each model under varying illumination, occlusion, and scene-complexity conditions. The specific test results are shown in Fig. 10.

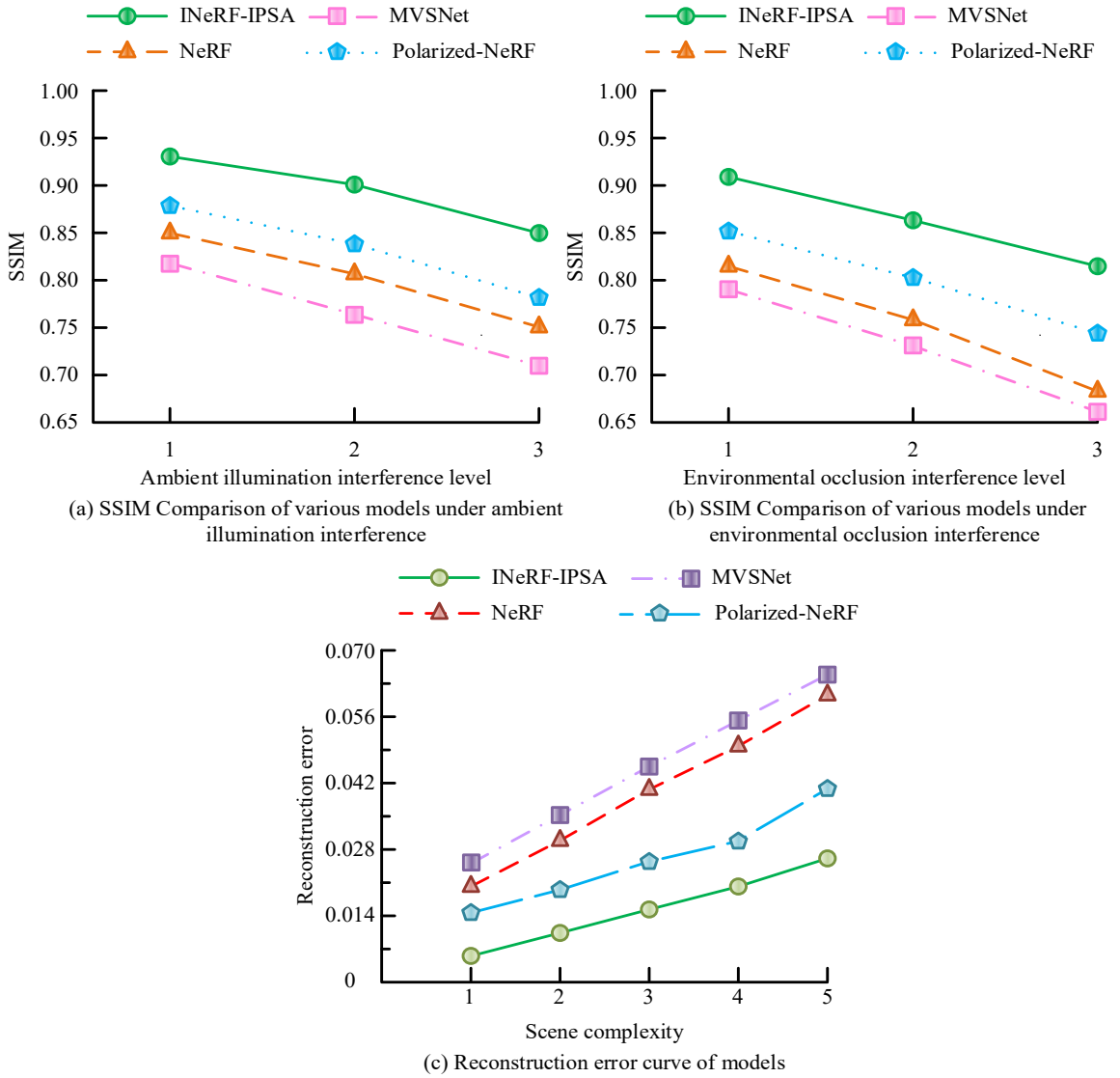


Fig. 10. Comparing 3D reconstruction of urban historical landscapes under environmental interference

As shown in Fig. 10(a), the INeRF-IPSA model consistently achieved the highest Structural Similarity Index Measure (SSIM) values. Under standard lighting conditions, it reached 0.93, which decreased to 0.90 in low light and 0.85 in strong backlight, representing a decline of only 8.6%. Polarized-NeRF followed, with an SSIM of 0.78 in a strong backlight. NeRF and MVSNet performed worse, with SSIM values of 0.75 and 0.71, respectively, in strong backlight, showing declines exceeding 13%. As illustrated in Fig. 10(b), when faced with environmental occlusion, the INeRF-IPSA model still outperformed the others, achieving SSIMs of 0.91 under mild occlusion and 0.82 under severe occlusion, representing a 9.9% decline. Polarized-NeRF reached 0.74 under severe occlusion, while NeRF and MVSNet achieved 0.68 and 0.66, respectively, with declines exceeding 15%. These results fully validated the outstanding resistance of the proposed model to interference from complex environments. As depicted in Fig. 10(c), as scene complexity increased, reconstruction errors rose for all models, but the INeRF-IPSA model exhibited the smallest increase in error, with the most gradual curve. Even in the most complex scenes, the reconstruction error of the INeRF-IPSA model was only 0.025, significantly lower than 0.04 for Polarized-NeRF and 0.06 and 0.065 for NeRF and MVSNet, respectively. This indicated that the INeRF-IPSA model, through its unique representation and feature enhancement techniques, could more effectively manage high-frequency details and complex geometric structures, ensuring high-quality 3D reconstruction in historical landscape scenes of varying complexity. To delve deeper into the performance improvements brought about by IPSA in the INeRF-IPSA model, the study focused on analyzing its ability to extract details of historical landscape elements by comparing it with the Polarized-NeRF model. The specific test results are shown in Fig. 11.

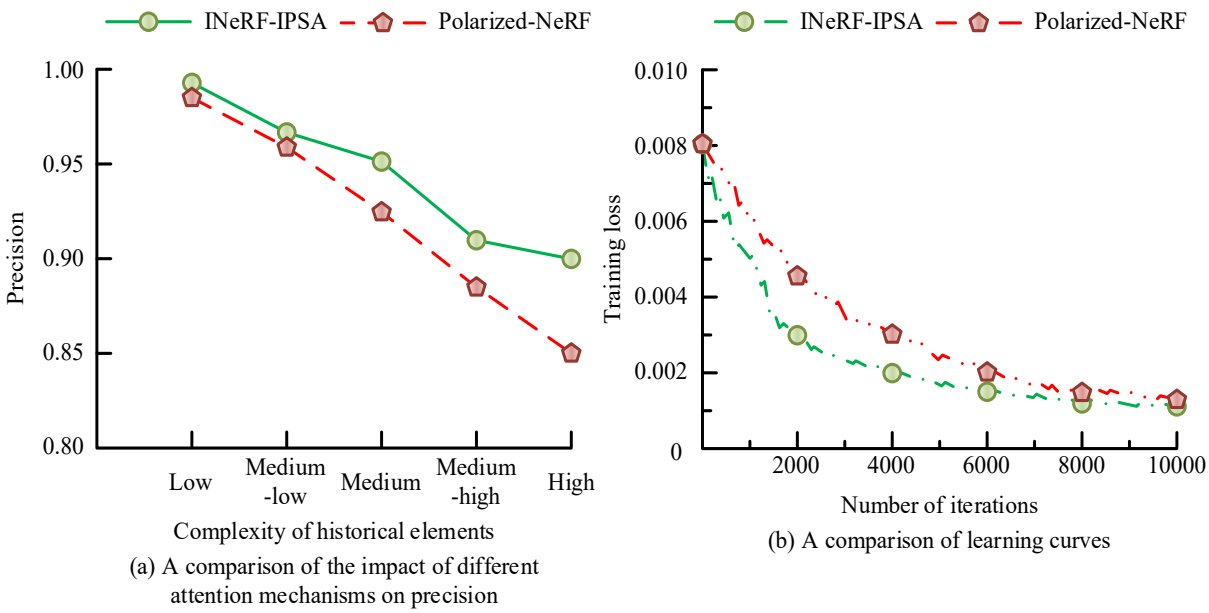


Fig. 11. Performance comparison of different attention mechanisms

As shown in Fig. 11(a), when managing historical elements of varying complexity, the INeRF-IPSA model consistently demonstrated superior accuracy compared to the Polarized-NeRF model. Particularly when dealing with highly complex and detailed hollow-carved decorations on ancient buildings, the INeRF-IPSA model maintained an accuracy of around 90%, whereas the Polarized-NeRF model's accuracy dropped to 85%. This indicated that the IPSA mechanism offered significant advantages in capturing and reconstructing complex geometric details. As illustrated in Fig. 11(b), the INeRF-IPSA model showed a faster decline in loss and ultimately converged to a lower loss level. By the 4000th iteration, the loss value of the INeRF-IPSA model had decreased to approximately 0.002, while that of the Polarized-NeRF model remained at 0.003. This demonstrated that the IPSA mechanism not only enhanced the model's ability to capture details but also accelerated convergence, thereby improving training efficiency and final reconstruction quality.

3.2. Analysis of the Application Effectiveness of the INeRF-IPSA Model in Real-World Scenarios

To verify the value and effectiveness of the INeRF-IPSA model in practical applications, a series of experiments was designed to document and restore 3D elements of the urban historical landscape. The study aimed to evaluate the model's performance in real, complex environments and the practicality of its outputs for cultural heritage preservation and urban planning management. Initially, ancient architectural bracket sets were selected as the primary target for validating the INeRF-IPSA model's performance in 3D digital documentation of complex components. As an iconic structural element in traditional Chinese architecture, the intricate and sophisticated design of bracket sets imposes stringent requirements on the accuracy and integrity of detail in 3D reconstruction. This experiment utilized the OpenHeritage 3D dataset, which contains a large volume of high-precision 3D point cloud and texture data of ancient buildings. During data processing, the collected bracket images underwent preprocessing, including illumination correction and camera parameter calibration, to ensure the quality of the input data. Subsequently, the INeRF-IPSA model was employed to reconstruct and extract features from these data, followed by a performance evaluation of the reconstructed 3D models. The final results are presented in Table 1.

As shown in Table 1, the INeRF-IPSA model excelled at 3D digital documentation of bracket sets in ancient architecture, achieving remarkable reconstruction accuracy and exceptional detail integrity. The model encompassed six typical installation scenarios for bracket sets, caisson ceilings, eaves areas, terraces, corners, corridor pillars, and lintels, utilizing a total of 34 datasets. The model's mean error remained consistently between 0.9 and 1.8mm. Even for corner bracket sets with intricate structures, the error remained exceedingly low, fully demonstrating its superior ability to restore the complex geometric configurations of bracket sets. Regarding detail representation, the model maintained a texture clarity rating of 4.2 to 5.0 and achieved a detail completeness rate ranging from 93.8% to 99.8%, enabling precise reproduction of the fine textures and minute components of the bracket sets. Although the reconstruction time (38 to 78 minutes) increased in tandem with the number of images (95 to 190), the quality of the 3D outputs consistently met the required standards, fully satisfying the demands of practical documentation. To further assess the robustness and detail-capturing ability of the INeRF-IPSA model in handling large-scale historical elements, an application experiment was conducted focusing on the texture analysis and damage detection of ancient city walls. The Cologne-Klasse-2015 dataset was chosen for this study, as it included complex stone textures and weathering patterns, making it suitable for simulating the real conditions of ancient city walls. The images underwent preprocessing and calibration to ensure data accuracy. Subsequently, 3D reconstruction and damage identification were performed on selected wall sections, and the final results are presented in Table 2.

Table 1. Application results of INeRF-IPSA model in 3D digital archiving of ancient building Dou-Gong

Dou-Gong ID	Archiving area	Image count	Model reconstruction time(min)	Average model error(mm)	Texture clarity (Level 1-5)	Detail integrity (%)	File size (GB)
A-01	Caisson ceiling	95	38	0.9	5.0	99.8	0.70
A-02		100	40	1.0	4.9	99.5	0.75
A-03		105	42	1.1	4.9	99.1	0.80
A-04		110	44	1.2	4.8	98.7	0.85
A-05		115	46	1.2	4.8	98.5	0.90
A-06		120	48	1.3	4.7	98.2	0.95
A-07		125	50	1.3	4.7	97.9	1.00
B-01	Eaves	130	52	1.4	4.6	97.6	1.05
B-02		135	55	1.4	4.6	97.4	1.10
B-03		140	57	1.5	4.5	97.1	1.15
B-04		145	59	1.5	4.5	96.8	1.20
B-05		150	61	1.6	4.4	96.5	1.25
B-06		155	63	1.6	4.4	96.2	1.30
B-07		160	65	1.7	4.3	95.9	1.35
C-01	Flat-seat	100	41	1.0	4.9	99.3	0.76
C-02		105	43	1.1	4.9	99.0	0.81
C-03		110	45	1.1	4.8	98.6	0.86
C-04		115	47	1.2	4.8	98.3	0.91
C-05		120	49	1.2	4.7	98.0	0.96
C-06		125	51	1.3	4.7	97.7	1.01
D-01	Corner	165	67	1.7	4.3	95.6	1.32
D-02		170	69	1.7	4.3	95.3	1.34
D-03		175	72	1.8	4.2	94.9	1.35
D-04		180	74	1.8	4.2	94.5	1.35
D-05		185	76	1.8	4.2	94.1	1.35
D-06		190	78	1.8	4.2	93.8	1.35
E-01	Corridor column	110	45	1.2	4.8	98.4	0.87
E-02		115	47	1.2	4.8	98.1	0.92
E-03		120	49	1.3	4.7	97.8	0.97
E-04		125	51	1.3	4.7	97.5	1.02
F-01	Lintel	135	56	1.4	4.6	97.0	1.12
F-02		140	58	1.5	4.5	96.7	1.17
F-03		145	60	1.5	4.5	96.4	1.22
F-04		150	62	1.6	4.4	96.1	1.27

In Table 2, the INeRF-IPSA model demonstrated exceptional performance and robustness in the tasks of ancient city wall texture analysis and damage detection. The model accurately identified complex wall textures with an impressive 98.5% precision and effectively distinguished several types of damage, including weathering, cracking, moss growth, and spalling. Moreover, the model maintained extremely low positioning errors for damage detection, ranging from 1.2–2.8cm. Additionally, the model could quantitatively assess the degree of weathering, providing precise data to support routine maintenance and scientific restoration of ancient city walls. Despite the relatively large data volume and reconstruction time, the model still achieved an elevated level of detail with a resolution of 0.4–0.6 millimeters per pixel, indicating that the research approach could effectively address the digitalization challenges of large-scale historical landscapes.

Subsequently, an application experiment was conducted to restore and manage 3D historical street layouts using the Stanford 3D Semantic Dataset. This dataset effectively simulated the complex environments of historical streets by including large-scale outdoor street scenes and architectural data, with specific application results presented in Table 3.

Table 2. Application results of INeRF-IPSA model in ancient city wall texture and damage detection

Wall section	Data points	Data size (GB)	Reconstruction time(h)	Texture recognition accuracy (%)	Damage type identified	Damage location error (cm)	Weathering quantified (%)	Model fineness(mm/pixel)
East-A	2500	5.2	6.5	98.2	Weathering, cracking	1.5	15.3	0.5
East-B	3000	6.5	8.0	97.5	Moss, spalling	2.1	8.9	0.6
South-A	2200	4.8	6.0	98.5	Weathering, spalling	1.2	12.1	0.4
South-B	2800	6.0	7.5	97.8	Cracking, moss	2.5	7.6	0.5
West-A	2000	4.5	5.5	98.0	Weathering, spalling	1.8	10.5	0.4
West-B	3200	7.0	8.5	96.9	Cracking, weathering	2.8	18.2	0.6
North-A	2600	5.8	7.2	97.6	Moss, weathering	2.0	9.7	0.5
North-B	2900	6.2	7.8	97.3	Cracking, spalling	2.3	11.4	0.5

In Table 3, the INeRF-IPSA model achieved high-precision reconstruction of large-scale street scenes extending up to 180 meters, with the accuracy of the 3D models consistently maintained at 2.1–2.8cm. Meanwhile, the model demonstrated strong semantic segmentation capabilities, achieving identification accuracy of over 94.7% for elements such as buildings, memorial arches, and stone-paved roads, providing precise geometric and semantic information for subsequent urban planning and landscape management. Although the reconstruction time increased with the number of images, the model's update efficiency remained above 91.5%, demonstrating its ability to provide rapid responses and dynamic updates. These results fully validated the robustness and efficiency of the INeRF-IPSA model in handling large-scale, complex historical scenes. To evaluate the model’s capability to restore high-precision, highly detailed elements, the study conducted application experiments in 3D pattern restoration and text recognition of historical stone inscriptions. The TUM-RGBD dataset was employed, which contained numerous small objects and textures, effectively simulating the fine textures and minute irregularities of stone inscriptions. The specific application results are presented in Table 4.

In Table 4, when the image resolution was set to 4K, the INeRF-IPSA model achieved reconstruction times of 32-40 minutes, with a model error of only 0.5-0.6mm. The texture restoration rate ranged from 97.5% to 98.5%, the text recognition accuracy ranged from 94.1% to 95.2%, the minimum recognizable font size was 1.2 to 1.8 mm, and the file size was 800 to 920 MB. After the resolution was upgraded to 8K, the model's reconstruction time extended to 55–65 minutes, and the file size increased to 1100–1200 MB. However, the model error decreased to 0.3–0.4mm, the texture restoration rate improved to 98.8%–99.1%, the text recognition accuracy reached 96.0%–96.5%, and the minimum recognizable font size was reduced to 0.9–1.0mm. The data indicated that the INeRF-IPSA model achieved significant improvements in precision with increased resolution. Even in 4K scenarios, it could meet the requirements for restoring fine features of stone inscriptions, fully validating its reliability in the digital documentation of historical stone tablets.

Table 3. Application results of INeRF-IPSA model in 3D reconstruction and management of historic streets

Street ID	Length (m)	Image sequence	Reconstruction time (h)	3D Model Accuracy(cm)	Semantic segmentation accuracy (%)	Recognizable elements	Model update efficiency (%)	Data storage (GB)
JL-01	150	500	8.5	2.5	95.8	Buildings, memorial archways, flagstones	92.1	12.5
JL-02	120	450	7.2	2.1	96.3	Buildings, ancient trees, manhole covers	94.5	10.8
JL-03	180	600	10.1	2.8	94.7	Buildings, gate towers, stone lanterns	91.5	15.3
JL-04	135	520	9.0	2.3	95.9	Buildings, memorial archways, streetlights	93.8	13.0
JL-05	160	550	9.5	2.6	95.4	Buildings, ancient trees, storefronts	92.9	14.1
JL-06	140	480	8.8	2.4	96.0	Buildings, streetlights, flagstones	93.5	11.8

4. Discussion

In the task of extracting urban historical landscape elements, the INeRF-IPSA model demonstrated significant performance advantages over comparative models. During tests of model convergence efficiency and 3D reconstruction accuracy, the INeRF-IPSA model consistently achieved the lowest training loss. This approach shared similarities with the YOLO-MS multimodal detection framework proposed by Xie et al. (2023), which enhanced detection accuracy through feature interaction and self-attention fusion. However, the INeRF-IPSA model demonstrated even greater strengths, achieving far superior convergence efficiency and model stability when managing multi-detailed, high-dimensional data characteristic of urban historical landscapes, compared to competing models. In comparative experiments, the INeRF-IPSA model followed a similar logic to the multi-view stereo vision framework proposed by Xu et al. (2024), yet outperformed it in fine-grained feature capture. While the multi-view stereo vision framework addressed generalization issues, it focused primarily on maintaining feature consistency across views without specific optimizations for fine-grained elements. In contrast, the INeRF-IPSA model enhanced multi-scale feature perception by introducing ASP to optimize PSA, enabling precise capture of subtle features such as ancient architectural bracket set textures and stone inscription patterns. This effectively addressed the limitations of traditional models in fine-grained element extraction. When applied to real-world urban historical landscape digitalization tasks, the INeRF-IPSA model further highlighted its performance advantages, aligning with the direction of expanding adaptability and practicality proposed by Lei et al (2024). through intelligent mesh generation technology. However, the INeRF-IPSA model surpassed it in scene adaptability and output precision. Although intelligent mesh generation technology expanded the scope of mesh generation through machine learning, it lacked optimizations for high-precision applications such as cultural heritage preservation. The INeRF-IPSA model, on the other hand, not only accommodated the processing requirements of diverse historical landscape elements, such as bracket sets, ancient city walls, streets, and stone inscriptions, but also produced high-precision 3D semantic models that fully met the stringent accuracy demands of urban historical landscape digital documentation. In summary, this model not only provided a superior solution for the extraction of urban historical landscape elements but also established a higher precision standard

for the application of computer vision technology in cultural heritage digitalization, with significant practical value. Future research could further optimize the model for extreme scenarios (e.g., severe weathering, dense occlusions) based on its existing strengths, thereby expanding its application advantages in cultural heritage preservation.

Table 4. INeRF-IPSA model results for 3D restoration and text recognition of historical inscriptions

Inscription ID	Image resolution	Recon. Time (min)	Model error (mm)	Texture rest. Degree (%)	Text recognition Accuracy (%)	Min. font size (mm)	File size (MB)
BK-01	4K	32	0.6	97.5	94.1	1.8	800
BK-02		33	0.6	97.6	94.3	1.7	820
BK-03		34	0.6	97.7	94.5	1.6	840
BK-04		35	0.5	97.9	94.8	1.5	850
BK-05		36	0.5	98.0	94.9	1.4	860
BK-06		37	0.5	98.1	95.0	1.3	880
BK-07		38	0.5	98.2	95.1	1.3	890
BK-08		39	0.5	98.4	95.2	1.2	910
BK-09		40	0.5	98.5	95.2	1.2	920
BK-10	8K	55	0.4	98.8	96.0	1.0	1100
BK-11		56	0.4	98.8	96.1	1.0	1110
BK-12		58	0.4	98.9	96.2	1.0	1130
BK-13		60	0.3	99.0	96.3	0.9	1150
BK-14		62	0.3	99.0	96.4	0.9	1180
BK-15		65	0.3	99.1	96.5	0.9	1200

5. Conclusion

To address the issues of inaccurate feature extraction and limited recognition accuracy exhibited by traditional image analysis methods and single deep learning models when processing high-resolution, multi-source heterogeneous urban historical landscape data, the study proposed an urban historical landscape element extraction method integrating INeRF and IPSA. In comparative experiments on urban historical landscape element extraction, the INeRF-IPSA algorithm achieved an extraction accuracy of 94.2% for elements at complexity level one, surpassing the Polarized-NeRF algorithm (89.7%) by 4.5 percentage points and outperforming the NeRF algorithm (87.2%) and MVSNet algorithm (84.1%) by 7.0 and 10.1 percentage points, respectively. Further analysis of fine-grained element extraction accuracy revealed that the algorithm achieved 95.0% precision in extracting ancient city wall textures, while maintaining extraction accuracies of 90.1% and 88.0% for intricately detailed ancient architectural bracket sets and historical stone inscription patterns, respectively. Notably, in the challenging task of extracting openwork carvings on complex ancient buildings, the algorithm demonstrated a precision of 90%, significantly higher than that of the Polarized-NeRF algorithm (85%). These results validated the reliability of the INeRF-IPSA algorithm for fine-grained feature extraction and complex element pattern classification, providing technical support for the precise mining of key elements in urban historical landscapes. The INeRF-IPSA model exhibited outstanding performance in terms of optimized efficiency, feature extraction accuracy, and practical application reliability, effectively overcoming the technical bottlenecks of traditional methods in extracting historical landscape elements under complex conditions. It offered a novel approach for in-depth analysis and precise modeling of urban historical landscape data. However, the research samples primarily relied on datasets such as ETH3D and OpenHeritage 3D, resulting in limited coverage of urban historical landscapes with distinct regional characteristics (e.g., streets in southern water towns and ancient buildings of ethnic minorities). Subsequent research could further expand the sample scope by incorporating more diverse historical landscape data to provide more comprehensive technical support for the preservation and transmission of urban historical contexts across a wider range of regions.

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Xiaofei Chen and Liqun Guo. The first draft of the manuscript was written by Xiaofei Chen. All authors commented on previous versions of the manuscript. All authors read and approved of the final manuscript.

Funding

The research is supported by Hubei Provincial Higher Education Teaching Research Project, Project Number: 2022313.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used ChatGPT for formatting and organizational assistance only. There were no AI tools used to generate scientific content, analysis, conclusions, or references. All content was reviewed and validated by the authors.

References

Chen, S., Xie, E., Ge, C., Chen, R., Liang, D., and Luo, P. (2023). Cyclemlp: A mlp-like architecture for dense visual predictions. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 45(12), 14284-14300. doi: 10.1109/TPAMI.2023.3303397.

Chen, Z., Wang, C., Guo, Y. C., and Zhang, S. H. (2023). Structnerf: Neural radiance fields for indoor scenes with structural hints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12), 15694-15705. doi: 10.1109/TPAMI.2023.3305295.

Das, P. K., Dash, A., and Meher, S. (2024). ACDSSNet: Atrous convolution-based deep semantic segmentation network for efficient detection of sickle cell anemia. *IEEE Journal of Biomedical and Health Informatics,* 28(10), 5676-5684. doi: 10.1109/JBHI.2024.3362843.

Feng, Y., Meng, X., Zhou, F., Lin, W., and Su, Z. (2023). Real-world non-homogeneous haze removal by sliding self-attention wavelet network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10), 5470-5485. doi: 10.1109/TCSVT.2023.3256414.

Guo, S., Wang, Q., Gao, Y., Xie, R., Li, L., Zhu, F., and Song, L. (2024). Depth-guided robust point cloud fusion NeRF for sparse input views. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9), 8093-8106. doi: 10.1109/TCSVT.2024.3385360.

Han, D., Ryu, J., Kim, S., Kim, S., Park, J., Yoo, and H. J. (2023). MetaVRain: A mobile neural 3-D rendering processor with bundle-frame-familiarity-based NeRF acceleration and hybrid DNN computing. *IEEE Journal of Solid-State Circuits*, 59(1), 65-78. doi: 10.1109/JSSC.2023.3291871.

Hasanvand, M., Nooshyar, M., Moharamkhani, E., and Selyari, A. (2023). Machine learning methodology for identifying vehicles using image processing. *Artificial Intelligence and Applications, 1*(3), 154-162. doi: 10.47852/bonviewAIA3202833.

Huang, X., Zhang, Q., Feng, Y., Li, H., and Wang, Q. (2024). Ltm-nerf: Embedding 3d local tone mapping in hdr neural radiance field. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 46(12), 10944-10959. doi: 10.1109/TPAMI.2024.3448620.

Jamshidi, B., Hakak, S., and Lu, R. (2023). A self-attention mechanism-based model for early detection of fake news. *IEEE Transactions on Computational Social Systems,* 11(4), 5241-5252. doi: 10.1109/TCSS.2023.3322160.

Jing, W., Wang, S., Zhang, W., and Li, C. (2023). Reconstruction of neural radiance fields with vivid scenes in the metaverse. *IEEE Transactions on Consumer Electronics*, 70(1), 3222-3231. doi: 10.1109/TCE.2023.3346870.

Lei, N., Li, Z., Xu, Z., Li, Y., and Gu, X. (2023). What's the situation with intelligent mesh generation: A survey and perspectives. *IEEE transactions on visualization and computer graphics*, 30(8), 4997-5017. doi: 10.1109/TVCG.2023.3281781.

Liu, S., He, N., Wang, C., Yu, H., and Han, W. (2023). Lightweight human pose estimation algorithm based on polarized self-attention. *Multimedia Systems,* 29(1), 197-210. doi: 10.1007/s00530-022-00981-z.

Matos, C. E. F., Junior, G. B., de Almeida, J. D. S., and de Paiva, A. C. (2024). Cpp-unet: combined pyramid pooling modules in the u-net network for kidney, tumor and cyst segmentation. *IEEE Latin America Transactions*, 22(8), 642-650. doi: 10.1109/TLA.2024.10620387.

Naumann, J., Xu, B., Leutenegger, S., and Zuo, X. (2024). NeRF-VO: Real-time sparse visual odometry with neural radiance fields. *IEEE Robotics and Automation Letters,* 9(8), 7278-7285. doi: 10.1109/LRA.2024.3421192.

Qi, Z., Kongfa, H., Tianshu, W., and Tao, Y. (2024). Lightweight and polarized self-attention mechanism for abnormal morphology classification algorithm during traditional Chinese medicine inspection. Digital Chinese Medicine, 7(3), 256-263. doi: 10.1016/j.dcmed.2024.12.005.

Qu, Q., Liang, H., Chen, X., Chung, Y. Y., and Shen, Y. (2024). Nerf-nqa: No-reference quality assessment for scenes generated by nerf and neural view synthesis methods. *IEEE Transactions on Visualization and Computer Graphics*, 30(5), 2129-2139. doi: 10.1109/TVCG.2024.3372037.

Sun, J., Xu, Y., Ding, M., Yi, H., Wang, C., Wang, J., Zhang, L., and Schwager, M. (2023). NeRF-Loc: Transformer-based object localization within neural radiance fields. *IEEE Robotics and Automation Letters*, 8(8), 5244-5250. doi: 10.1109/LRA.2023.3293308.

Wang, C., Jiang, R., Chai, M., He, M., Chen, D., and Liao, J. (2023). Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics,* 30(8), 4983-4996. doi: 10.1109/TVCG.2023.3283400.

Wang, N., Wang, Y., Feng, Y., and Wei, Y. (2024). MDD-ShipNet: Math-data integrated defogging for fog-occlusion ship detection. *IEEE Transactions on Intelligent Transportation Systems*, 25(10), 15040-15052. doi: 10.1109/TITS.2024.3394573.

Wu, Y., Yuan, H., Liu, Z., Wei, X., Dai, H. (2023). Multitasks Joint Prediction of Fuel Cells Based on Self-Attention Residual Network. *IEEE Transactions on Transportation Electrification*, 10(3), 6867-6879. doi: 10.1109/TTE.2023.3343519.

Xie, J., Shi, Y., Ni, D., Milling, M., Liu, S., Zhang, J., Qian, K., and Schuller, B. W. (2024). Automatic bird sound source separation based on passive acoustic devices in wild environment. *IEEE Internet of Things Journal*, 11(9), 16604-16617. doi: 10.1109/JIOT.2024.3354036.

Xie, Y., Zhang, L., Yu, X., and Xie, W. (2023). YOLO-MS: Multispectral object detection via feature interaction and self-attention guided fusion. *IEEE Transactions on Cognitive and Developmental Systems*, 15(4), 2132-2143. doi: 10.1109/TCDS.2023.3238181.

Xu, H., Chen, W., Sun, B., Xie, X., and Kang, W. (2024). Robustmvs: Single domain generalized deep multi-view stereo. *IEEE Transactions on Circuits and Systems for Video Technology,* 34(10), 9181-9194. doi: 10.1109/TCSVT.2024.3399458.

Yao, X., Wang, Y., Wu, Y., He, G., and Luo, S. (2023). MLP-based efficient convolutional neural network for lane detection. *IEEE Transactions on Vehicular Technology,* 72(10), 12602-12614. doi: 10.1109/TVT.2023.3275571.

Yarza Pérez, A. J. and Verbakel, E. (2025). The role of adaptive reuse in historic urban landscapes towards cities of inclusion. The case of acre. *Journal of cultural heritage management and sustainable development,* 15(2), 306-339. doi: 10.1108/JCHMSD-05-2022-0074.

Zhang, J., Li, X., Wan, Z., Wang, C., and Liao, J. (2024). Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(12), 7749-7762. doi: 10.1109/TVCG.2024.3361502.

Zhou, P., Xie, L., Ni, B., and Tian, Q. (2023). Cips-3d++: End-to-end real-time high-resolution 3d-aware gans for gan inversion and stylization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10), 11502-11520. doi: 10.1109/TPAMI.2023.3285648.

Zhou, Z., Islam, M. T., and Xing, L. (2023). Multibranch CNN with MLP-mixer-based feature exploration for high-performance disease diagnosis. IEEE Transactions on Neural Networks and Learning Systems, 35(6), 7351-7362. doi: 10.1109/TNNLS.2023.3250490.

Zhou, Z., Zhong, T., Liu, M., and Ye, Y. (2023). Evaluating building color harmoniousness in a historic district intelligently: An algorithm-driven approach using street-view images. *Environment and Planning B: Urban Analytics and City Science*, 50(7), 1838-1857. doi: 10.1177/23998083221146539.

Zhu, J., Wang, H., Hogg, D., and Kelly, T. (2025). Learning to sculpt neural cityscapes. *The Visual Computer,* 41(4), 2233-2249. doi: 10.1007/s00371-024-03528-7.

Xiaofei Chen received her Master's degree in Landscape Architecture from the School of Urban Planning and Architecture at Huazhong University of Science and Technology in 2008. She is currently a faculty member at the School of Art and Design, Wuhan Institute of Technology. She conducted visiting research at the School of Architecture, Southeast University in Nanjing as a domestic Young Backbone Teacher, focusing on landscape architecture. In 2022, she was funded by the Special Art Talent Training Program of the China Scholarship Council and served as a research fellow in the Interior Architecture program at Tokyo Zokei University, Japan, where she conducted research on co-working space design. In terms of research, she has led one provincial-level teaching research project in Hubei, one youth project funded by the Hubei Provincial Department of Humanities and Social Sciences, one key project of the Humanities and Social Sciences Research Base of Higher Education Institutions in Hubei, and one university-level humanities and social sciences project. She has also developed a university-level first-class course.

Liqun Guo holds a Ph.D. from Wuhan University of Technology and is currently a professor at Wuhan Institute of Technology. Her primary research interests focus on architectural environmental design, both interior and exterior. She has long been engaged in related research and teaching reform, having led multiple projects, including key topics in Hubei Provincial Education Science Planning, provincial-level teaching research projects, research projects funded by the Hubei Provincial Department of Education, and projects supported by provincial key research bases in humanities and social sciences. She has also undertaken the development of a provincial high-quality resource-sharing course and several postgraduate teaching reform projects. In addition, she has participated in the National Social Science Foundation (Art Studies) project "Research on Cultural and Creative Product Development Based on Museum Grading Evaluation Standards," as well as several major provincial research projects, including "Landscape Spatial Restoration of Traditional Villages in Southeastern Hubei" and "Research on the Spatial Form Pattern of Wuhan City and Strategies for Regional Cultural Development."