Annals of Emerging Technologies in Computing (AETiC)
Vol. 10, No. 3, 2026

Research Article

Spatiotemporal Transformer-Based Video Generation Method for Multi-Camera Smart City Scenarios

Bingchao Sun

Zhejiang College of Construction, Hangzhou, China
190133@zjjs.edu.cn

Received: 13 January 2026; Accepted: 12 May 2026; Published: 01 July 2026

Abstract: Traditional video acquisition and generation methods are limited by sensor layout, data transmission costs, and modeling capabilities, rendering it challenging to simultaneously capture local structure and global temporal dependencies in complex, multi-scale, and irregular urban scenes, resulting in inconsistent video quality and stability. To address the problem of urban video frame reconstruction and generation under complex monitoring conditions, this study introduces a smart city video generation method that integrates spatiotemporal transformer networks and graph convolution networks. First, a spatiotemporal Transformer is used to capture spatiotemporal semantics. Then, graph convolution is introduced to model structural relationships among spatial patches within video frames, enabling effective spatial information propagation. Furthermore, gated recurrent units and multi-scale attention mechanisms are combined to enhance dynamic modeling and feature fusion capabilities. Experiments are conducted on urban monitoring video data containing three representative scenarios, including urban road traffic monitoring, public square surveillance, and environmental monitoring scenes. The performance of the proposed method is compared with convolutional neural network based video reconstruction methods and spatiotemporal transformer based models under the same experimental settings. In these scenarios, the introduced approach achieves a peak signal-to-noise ratio improvement of 2.14 dB, 3.01 dB, and 3.40 dB compared to the convolutional neural network based reconstruction method, respectively, while the structural similarity index increases by approximately 1.52% to 2.97% across different scenes. Moreover, compared to existing video modeling methods reported in the literature, the proposed method exhibits the least performance degradation under various types of noise and occlusion disturbances. The findings demonstrate that the introduced approach can achieve higher quality and more robust video generation in complex dynamic scenes, particularly in urban monitoring environments with noise interference and partial occlusion, providing a new technical path and implementation means for visual perception in smart cities.

Keywords: ***Graph Neural Network; Multi-Camera Video Generation; Multi-Scale Feature Fusion; Smart City; Spatiotemporal Transformer***

1. Introduction

With the continuous advancement of smart city construction, urban infrastructure, public transportation, security monitoring and environmental governance are rapidly becoming digitalized and intelligent. As an important component of multi-source sensing data in smart cities, video carries the spatiotemporal information expression of complex dynamic scenes and is of great significance for urban event perception, environmental monitoring, traffic scheduling and emergency response [1]. With the development of internet of things technology, urban sensing systems have gradually formed a sensing network composed of multi-source sensors and intelligent analysis algorithms. The integration of machine learning and IoT technology has become an important research direction for the development of smart cities [2]. However, traditional video acquisition methods are limited by sensor layout, data transmission and storage costs, making it challenging to fulfil the video demands for high spatiotemporal resolution and strong robustness. Therefore, deep

Bingchao Sun, "Spatiotemporal Transformer-Based Video Generation Method for Multi-Camera Smart City Scenarios", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, 1 July 2026, Vol. 10, No. 3, pp. 1-15, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2026.03.001, Available: http://aetic.theiaer.org/archive/v10/v10n3/p1.html.

learning-based video generation and reconstruction methods have become one of the key technologies in smart city sensing systems [3]. However, existing video generation methods either focus on pixel reconstruction and are difficult to model global spatiotemporal dependences [4]. In addition, existing methods focus on temporal modeling and lack characterization of local spatial topology, rendering it challenging to adapt to multi-scale, irregular and complex urban scenes [5]. Therefore, achieving efficient integration of spatial structure and temporal dependencies has become a key technical bottleneck.

In multi-camera urban video sensing networks, there are often obvious spatial connections and structural dependencies between different monitoring nodes. However, existing video generation methods ignore this cross-node topological information, making it difficult for models to simultaneously characterize the urban spatial structure and the long-term evolutionary features of video sequences. Furthermore, single temporal modeling methods are prone to semantic inconsistencies or structural distortions in complex dynamic scenes. Therefore, it is necessary to construct a unified video generation framework capable of simultaneously modeling spatial topological relationships and long-term temporal dependencies.

To address the aforementioned issues, this research puts forward a spatiotemporal Transformer (STT) video generation method for smart cities, aiming to fully extract and efficiently model the spatiotemporal features of smart city videos. This study utilizes STT to capture spatiotemporal semantic information from videos and employs graph convolutional network (GCN) to perform topological modeling of non-Euclidean spatial structures. The primary innovations and notable contributions of this research are outlined below: (1) A video generation framework GCN-STT, which integrates GCN and STT, is proposed, effectively combining local topological structure and long temporal information modeling; (2) A multi-scale feature fusion mechanism is designed for complex dynamic scenarios in smart cities, improving the adaptability to multi-target and multi-scale motion; (3) In the temporal modeling stage, a gated recurrent unit (GRU) is introduced to dynamically model node features, and a multi-scale attention mechanism is used to achieve weighted fusion of features at different time scales and structural levels, thereby enhancing the temporal stability and structural consistency in the video generation process; (4) The effectiveness of the method is verified through multi-dimensional indicators of pixel, structure, perception, and temporal consistency, providing a new technical path for high-quality video generation in smart city visual perception systems.

2. Related Works

Against the backdrop of the rapid development of smart city construction, high-quality video generation technology has become an important foundation for supporting intelligent perception and decision-making in various domains such as urban traffic monitoring, public safety, and environmental governance. Existing research has explored this topic extensively from the perspective of urban visual perception and video surveillance systems. For example, to solve the urgent needs of citizens for safety and real-time anomaly monitoring in the context of accelerated global urbanization, Sharma *et al*. proposed a standardized smart city monitoring framework that integrates cutting-edge artificial intelligence detection and massive video data. By constructing a unified video perception model, accurate perception of key events such as abnormal behavior, fire, and tampering was achieved [6]. Zhou *et al*. introduced a unified fusion framework to address the problem of the separation of low-level and high-level visual perception in existing traffic monitoring systems. By introducing advanced perception enhancement, efficient learning, knowledge-driven understanding, collaborative perception and lightweight computing, the accuracy of scene understanding and practical adaptability were comprehensively improved [7]. In addition, Zhang *et al*. proposed a defogging-reconstruction integrated solution based on deep learning to tackle the challenges of color distortion and low contrast caused by outdoor haze images. By estimating the transmission map through a fine transmission depth convolutional regression network, the accuracy and efficiency of 3D reconstruction were improved [8]. Hu *et al*. introduced a new multi-view stereo method grounded in deep learning to address the problem of high computational complexity and memory usage of monocular cameras. By efficiently estimating depth pixel by pixel through a lightweight network, real-time 3D reconstruction of large scenes was achieved [9]. These studies provide an important foundation for visual perception and complex scene understanding in smart cities.

With the development of deep learning, video generation and prediction technologies have gradually evolved from traditional convolutional structures to deep learning frameworks that integrate spatial and temporal features. For example, to solve the problem of limited reconstruction quality of high-frame sequences from single-frame measurements in video snapshot compression imaging, Wang *et al*. proposed Spatial-Temporal Transformer (STFormer). By using token generation, spatiotemporal dual-branch self-attention, and fusion network to jointly mine spatial details and temporal correlations, efficient real data reconstruction and simulation are achieved [10]. To handle the issue of difficult detection of anomalies in the remote sensing perspective of UAVs, Tran *et al*. proposed an unsupervised Transformer future frame prediction network. By jointly modeling the spatiotemporal sequence through encoder-decoder and locating anomalies with reconstruction errors, they achieved accurate early warning of events such as traffic accidents and traffic jams [11]. Kaddar *et al*. introduced a hybrid framework of convolutional neural network (CNN) + vision Transformer (ViT) to address the problem that CNN Deepfake detectors only focus on local artifacts and have weak generalization across generation methods. By extracting local features through CNN and feeding them into ViT for self-attention to learn global dependencies, the detection accuracy and generalization ability of unknown forgery algorithms were improved [12]. Tang *et al*. designed a Human-centric Spatiotemporal Video Grounding (HC-STVG) method to address the problem that it is difficult to accurately locate the spatiotemporal trajectory of a specific person in long-term monitoring text descriptions. By extracting cross-modal representations through ViT, they jointly located the spatial tube and time period, thereby achieving high-precision positioning of the spatiotemporal tube of the target person [13].

In summary, while current research on smart city video generation has made significant progress in areas such as dehazing reconstruction, surveillance fusion, 3D scene reconstruction, and anomaly detection, existing methods still suffer from limitations, including insufficient spatial topology characterization, limited global spatiotemporal dependency modeling capabilities, and poor adaptability to complex dynamic scenes. Particularly in multi-objective, multi-scale, and irregular urban scenarios, single convolutional or temporal modeling methods struggle to simultaneously capture local details and global correlations. To address this gap, this research proposes a smart city video generation method based on the fusion of GCN and STT. By introducing topology-aware spatial structure modeling and a multi-scale temporal feature fusion mechanism, it aims to improve the semantic consistency, detail reproduction, and temporal stability of the generated video. The innovative aspect of this research lies in estabishng a spatial and temporal collaborative modeling mechanism and a comprehensive verification system based on multi-dimensional indicators.

3. Research Methodology

The research focuses on video generation and reconstruction tasks in multi-camera urban surveillance scenarios, namely, generating target video frames with spatiotemporal consistency given historical video sequences and multi-node spatial structure information. First proposes a video feature modeling and encoding method based on STT. Second, it constructs a video generation framework integrating GCN and STT to enhance spatial topology modeling and global temporal dependency capture capabilities. Finally, through multi-scale feature fusion and temporal dynamic modeling, it achieves high-quality video generation and reconstruction of complex smart city scenarios.

3.1. Video Feature Modeling and Coding Method Based on STT

With the continuous improvement of smart city infrastructure, the deployment of nodes such as road monitoring, pedestrian density analysis, and environmental monitoring has formed a large-scale video stream with high spatiotemporal coupling. However, traditional CNNs are limited by local receptive fields and have difficulty capturing global dependencies [14]. Recurrent Neural Networks (RNNs) are prone to gradient vanishing and information decay problems in long-term sequence modeling, making it difficult to fully characterize complex dynamic processes [15]. Therefore, this study proposes an STT feature modeling method to capture long-range dependencies in the time dimension and preserve the topological structure of the urban scene in the spatial dimension. The specific details are shown in Figure 1.

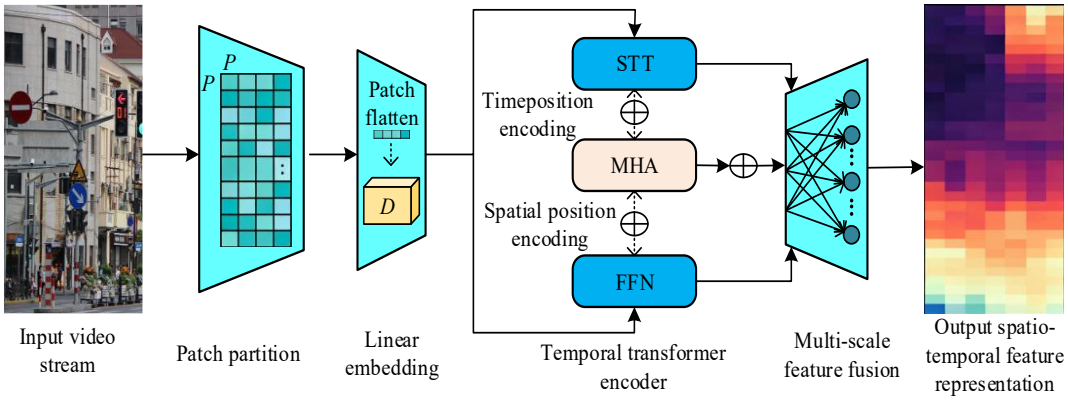


Figure 1. Schematic diagram of the overall framework for video feature modeling based on STT[1]

Figure 1 illustrates the video feature modeling framework based on STT. First, the input video stream is divided into several $P\times P$ sub-blocks and projected into a D-dimensional feature space through linear embedding, achieving a preliminary transformation from image to feature. Second, the features enter the temporal Transformer encoder, which combines temporal and spatial location encoding, and uses a multi-head attention (MHA) mechanism and a feedforward neural network (FNN) to model global temporal dependencies and enhance feature representation [16-17]. Based on this, features at different scales are fused through multi-scale feature fusion, taking into account both local details and overall structure. Assume that each frame of the video sequence is divided into multiple non-overlapping sub-blocks and embedded into a D-dimensional feature space through linear mapping, as shown in equation (1).

$$\begin{cases} X=\{x_1,x_2,...,x_T\} \\ z_{t,i}=W_p\cdot \text{Flatten}(p_{t,i})+b_q \end{cases} \tag{1}$$

In equation (1), X represents the video sequence; T represents the number of frames in the sequence. $x_t\in\mathbb{R}^{H\times W\times C}$ represents the image of the t th frame, where H, W, and C are the image height, width, and number of channels. $z_{t,i}\in\mathbb{R}^D$ represents the feature representation of the i th sub-block in the t th frame. $W_p\in\mathbb{R}^{D\times(P^2\cdot C)}$ represents the linear mapping matrix, P is the size of the sub-block; $p_{t,i}$ is the non-overlapping sub-block; $b_q\in\mathbb{R}^D$ is the bias term; and $\text{Flatten}(\cdot)$ is the flattening operation. After obtaining the patch embedding, to capture the long-term dependencies between frames, the self-attention mechanism of Transformer is introduced. First, the query (Q), key (K), and value (V) are obtained through linear transformation, as shown in equation (2).

$$\begin{cases} Q=ZW_Q \\ K=ZW_K \\ V=ZW_V \end{cases} \tag{2}$$

In equation (2), Z represents the feature representation vector; W_Q, W_K, and W_V are the learnable parameter matrices, respectively. The self-attention calculation formula is shown in equation (3).

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}})V \tag{3}$$

In equation (3), d_k represents the dimension of the key vector. Based on this, the study further enhances the feature representation capability through residual connections and FNN, as presented in equation (4).

$$\begin{cases} H^{(t)}=\text{MHA}(Z)+Z \\ H_{\text{out}}^{(t)}=\text{FFN}(\text{LayerNorm}(H^{(t)}))+H^{(t)} \end{cases} \tag{4}$$

In equation (4), $H^{(t)}$ is the feature representation after MHA. $H_{\text{out}}^{(t)}$ represents the output feature after further processing by LayerNorm and FNN. LayerNorm represents the layer normalization operation.

[1] https://pixabay.com/zh/photos/street-shanghai-city-china-road-6569760

However, using only the temporal dimension information cannot fully express the structural features in urban scenes. Therefore, the study adds spatial and temporal location encoding to the patch features to ensure that the features retain semantic location information during the temporal modeling process [18]. For the i th sub-block of the t th frame, the update formula is presented in equation (5).

$$\hat{z}_{t,i} = z_{t,i} + P_s(i) + P_t(t) \tag{5}$$

In equation (5), $P_s(i) \in \mathbb{R}^D$ represents spatial location coding information, reflecting the fixed position of the corresponding patch in the image. $P_t(t) \in \mathbb{R}^D$ represents temporal location coding, which mainly represents the video frame sequence information. The structure of spatiotemporal location coding and multi-scale feature fusion mechanism is shown in Figure 2.

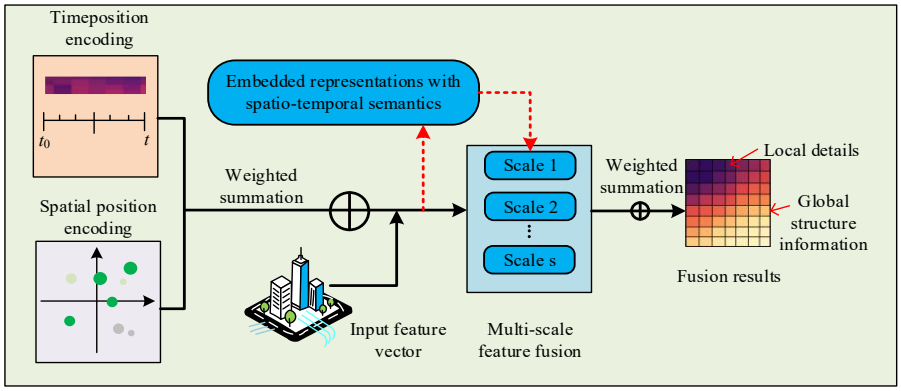


Figure 2. Schematic diagram of spatial and temporal location coding and multi-scale feature fusion mechanism

In Figure 2, temporal and spatial location encodings supplement the feature vectors with temporal and spatial semantics, respectively, ensuring that frame order and geometric location information are preserved in subsequent modeling. These two encodings are then fused with the input feature vector through a weighted sum to obtain an embedded representation with spatiotemporal semantics. Next, this representation is input to a multi-scale feature fusion module, which extracts local and global feature information from different scale branches. The features at each scale are weighted and summed to form the final fusion result. Assuming F_s represents the feature representation at the s th scale, and there are S scales in total, the mathematical expression for the final fused feature is shown in equation (6).

$$F_{\text{multi}} = \sum_{s=1}^{S} \alpha_s F_s \tag{6}$$

In equation (6), F_{multi} represents the final fused features; α_s represents a learnable weight parameter used to control the contribution ratio of features at each scale. Through this fusion mechanism, the model can not only identify small-scale pedestrian targets, but also simultaneously perceive large-scale structural information such as traffic flow and roads.

3.2. Smart City Video Generation Framework Based on GCN-STT

While STT-based video feature modeling and coding methods can identify and perceive urban features at different scales, they struggle to comprehensively represent urban regional interactions and structural dependencies. Therefore, this study further introduces a fusion mechanism of GCN and temporal modeling to construct a video generation framework, GCN-STT, with global topology awareness and temporal dynamic modeling capabilities. The details are shown in Figure 3.

In Figure 3, video stream data is first collected from distributed urban cameras and then processed by the STT module of the spatiotemporal feature modeling layer. Global temporal dependencies and local spatial structure features are extracted using MHA and time-space location coding. At the same time, the GCN graph convolution module uses urban video nodes as vertices and implements feature propagation between nodes based on the adjacency matrix to characterize the urban spatial topology. Secondly, the outputs of STT and GCN are fused and entered into the temporal modeling and video reconstruction layer. The Gated Recurrent Unit (GRU) module further models the temporal dynamics, and the decoder restores the target video frame

through deconvolution and upsampling. In the proposed framework, spatial patches extracted from video frames are treated as nodes in a graph $G=(V,E,A)$ [19]. Among them, the node set V represents N spatial patches within video frames; the edge set E represents the physical or functional association between nodes, and the adjacency matrix $A \in \mathbb{R}^{N\times N}$ is used to characterize the association. Considering the strong structural correlation between adjacent spatial regions, this study constructs an adjacency matrix based on spatial neighborhood relationships.

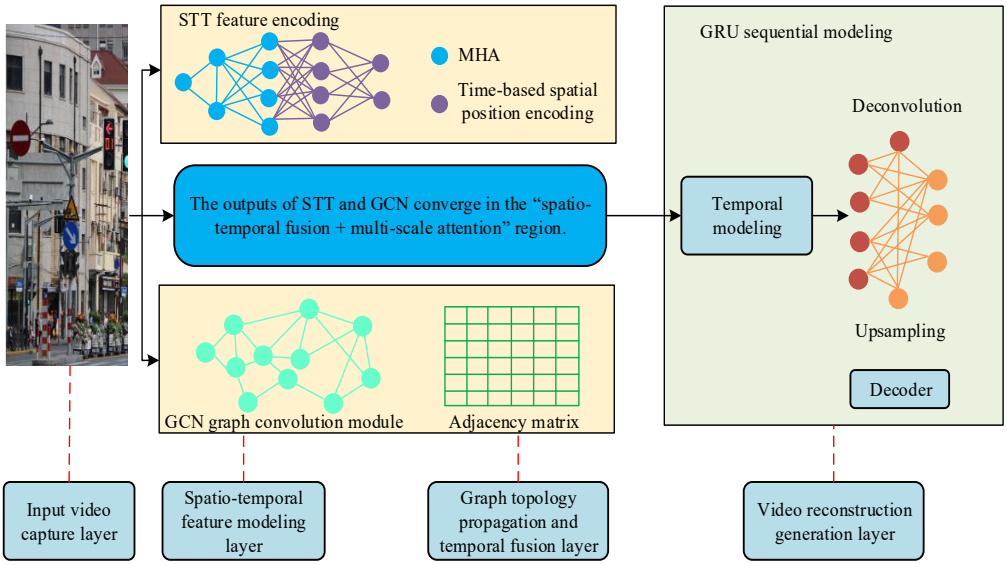


Figure 3. Smart city video generation framework based on GCN-STT

When two patches are spatially adjacent, their corresponding nodes are connected; otherwise, no connection is established. Therefore, the adjacency matrix is constructed using binary adjacency, where A_{ij}=1 when two nodes are adjacent, and A_{ij}=0 otherwise. Furthermore, the graph structure maintains a static topology during model training. The self-connection term I_N is added to the adjacency matrix to get equation (7).

$$\begin{cases} \tilde{A} = A + I_N \\ \tilde{D}_{ii} = \sum_j \tilde{A}_{ij} \end{cases} \tag{7}$$

In equation (7), $\tilde{D}_{ii}$ represents the degree matrix corresponding to the adjacency matrix, and the calculation process of performing a standard graph convolution layer is shown in equation (8).

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{8}$$

In equation (8), $H^{(l)}$ represents the node feature matrix of the l th layer, and the initial input $H^{(0)}$ comes from the encoding result of STT. $W^{(l)}$ represents the learnable weight matrix; σ represents the nonlinear activation function. By normalizing the adjacency matrix, graph convolution can smoothly aggregate the neighborhood features of each node, enabling the model to explicitly perceive the urban topology and its impact. To further enhance the neighborhood perception capability of the model, a multi-order propagation mechanism is introduced to include K -order neighbors in the convolution range, and the specific mathematical expression is shown in equation (9).

$$H^{(l+1)} = \sigma(\sum_{k=0}^{K}\theta_k T_k(\hat{L})H^{(l)}) \tag{9}$$

In equation (9), $\hat{L}$ represents the normalized Laplace matrix; θ_k represents the learnable Chebyshev coefficients; and $T_k(\cdot)$ represents the Chebyshev polynomial. The model can capture both local and long-range structural relationships through multi-order propagation. Considering that node features still have obvious temporal evolution patterns during the propagation of topological information, in order to efficiently

model the dynamic changes in the time dimension, the study introduces GRU to model the evolution trajectory of node features over time. For node feature X_t, its update process is shown in equation (10).

$$\begin{cases} z_t = \sigma(W_z X_t + U_z H_{t-1} + b_z) \\ r_t = \sigma(W_r X_t + U_r H_{t-1} + b_r) \\ \hat{H}_t = \tanh(W_h X_t + U_h (r_t \odot H_{t-1}) + b_h) \\ H_t = (1 - z_t) \odot H_{t-1} + z_t \odot \hat{H}_t \end{cases} \tag{10}$$

In equation (10), z_t and r_t are the update gate and reset gate; W_z, U_z, W_r, U_r, W_h and U_h all represent learnable parameters; b_z, b_r and b_h represent bias terms. $\odot$ represents the Hadamard product. GRU can flexibly control the retention and forgetting of information, and is superior to traditional RNN and Long Short-Term Memory Network (LSTM) in modeling the changes in node state over time. It has high computational efficiency and is suitable for the temporal modeling of large-scale urban sensing networks [20]. Based on this, in order to achieve deep coupling between spatial topological information and temporal dynamics, a multi-scale attention mechanism is introduced on the basis of graph convolution + GRU to fuse the features of different time steps and different graph convolutional layers. Its core idea is to assign weights α_s on the feature representations F_s at different scales, as shown in equation (11).

$$\begin{cases} \alpha_s = \dfrac{\exp(\mathbf{q}^{\mathrm{T}} \mathbf{k}_s)}{\sum_{j=1}^{S} \exp(\mathbf{q}^{\mathrm{T}} \mathbf{k}_j)} \\ F_{out} = \sum_{s=1}^{S} \alpha_s F_s \end{cases} \tag{11}$$

In equation (11), $\mathbf{q}$ represents the global query vector; $\mathbf{k}_s$ represents the key vector of each scale feature. Through the attention weight allocation mechanism, the model can adaptively select the scale information that contributes best to video generation, thereby achieving dynamic fusion and enhancement of features. In addition, to avoid the gradient decay problem when modeling long sequences, the study uses residual connections and layer normalization operations to optimize the output feature representation, as shown in equation (12).

$$\tilde{F}_{out} = \mathrm{LayerNorm}(F_{out} + H_t) \tag{12}$$

After completing topology propagation, temporal modeling, and feature fusion, the final video generation is achieved through the decoder module. The decoder adopts a structure combining deconvolution and upsampling to convert the spatiotemporal feature tensor $\tilde{F}_{out}$ into the target video frame $\hat{Y}_t$. The reconstruction process can be expressed as equation (13).

$$\hat{Y}_t = f_{\text{decoder}}(\tilde{F}_{out}) \tag{13}$$

In equation (13), f_{decoder} represents the decoder mapping function. To constrain the quality and stability of the generated video, a joint objective function of reconstruction loss and perceptual loss is introduced, as shown in equation (14).

$$Loss = \lambda_1 L_{\text{MSE}}(Y_t, \hat{Y}_t) + \lambda_2 L_{\text{Perceptual}}(Y_t, \hat{Y}_t) \tag{14}$$

In equation (14), L_{MSE} represents the pixel-level mean square error (MSE), used to ensure video clarity. $L_{\text{Perceptual}}$ represents the perceptual loss, used to measure semantic consistency. λ_1 and λ_2 represent weighting coefficients. Combining the above, the overall video generation process based on GCN-STT is shown in Figure 4.

In Figure 4, first, multi-source videos of the city are acquired. After preprocessing such as shading and brightness normalization, the video frames are segmented and linearly embedded, and spatiotemporal location encoding is injected. Second, the STT module models the video sequence through a self-attention mechanism, mainly for extracting long-distance temporal dependencies and global spatial semantic information. The GCN module uses an adjacency matrix to model the structural relationships between spatial nodes, realizing topological feature propagation, enabling the model to characterize the structural associations between different spatial regions. The outputs of both are integrated in a spatiotemporal fusion layer and weighted fusion is performed through a multi-scale attention mechanism. The multi-scale features come from

the feature representations output by different graph convolutional layers, corresponding to coarse-grained, medium-grained, and fine-grained structural information. The fused features are input in to the GRU for dynamic temporal modeling to capture the state evolution of video features in the temporal dimension. Finally, the decoder generates video frames through upsampling and deconvolution, and outputs the final video result by combining loss constraints and post-processing.

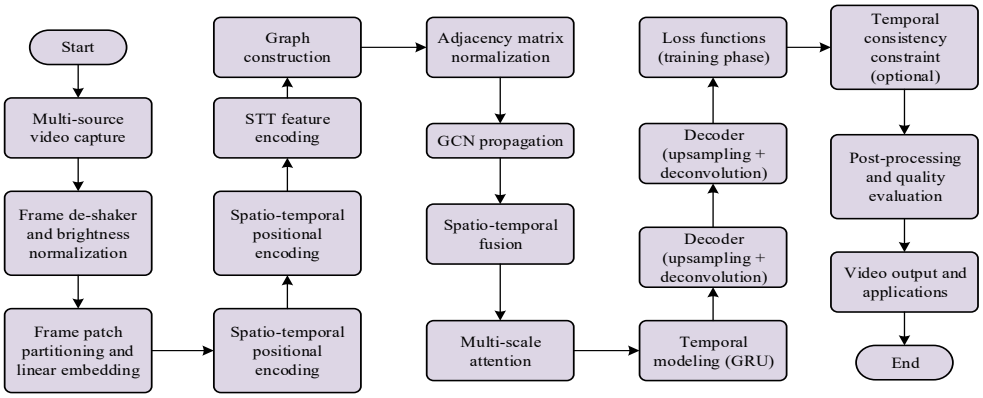


Figure 4. Video generation process based on GCN-STT

4. Results and Discussion

First, the experimental setup and relevant parameters for model training were defined. Second, the performance of the GCN-STT model was validated based on CNN and STT. Finally, the performance of the proposed method was compared with existing video generation methods.

4.1. Experimental Setup

To confirm the validity of the introduced approach, this study used multi-source video sensing scenarios of smart city traffic and public space monitoring as validation data. The data comes from the city's public monitoring system and open video resource platform, and has been uniformly organized and anonymized. This dataset covered three typical scenarios: road traffic monitoring, public square areas, and environmental monitoring nodes. Road traffic monitoring videos were collected from intersections of main urban roads, including dynamic scenes involving multiple targets such as vehicle flow, pedestrian flow, and traffic lights. Public square area videos were collected from city center squares and densely populated pedestrian areas, mainly reflecting crowd movement and environmental changes. Environmental monitoring node videos were collected from areas surrounding public facilities, including typical dynamic features such as weather changes and day-night light differences. The dataset contains 180 video sequences, with a total duration of approximately 3.6 hours. Each video is approximately 30-90 seconds long and covers 12 different monitoring areas. All videos were anonymized and desensitized, with a uniform sampling frequency of 25 fps and a resolution of 1280×720, and were divided into a 70% training set, 15% validation set, and 15% test set. Before formal training, the video frames underwent preprocessing steps including frame segmentation and linear embedding (each frame was divided into 16×16 patches and mapped to a high-dimensional feature space), brightness and contrast normalization, spatiotemporal location coding injection, and adjacency matrix construction.

The experimental hardware and software environment and the details of the hyperparameter settings for the GCN-STT model are shown in Table 1:

Table 1. Experimental Hardware and Software Environment and Model Hyperparameter Settings

Category	Item	Configuration/Parameter
Hardware	CPU	Intel Core i7-13700K 3.4 GHz
	GPU	NVIDIA RTX 4090 (24 GB VRAM)
	Memory	64 GB DDR5
	Storage	2 TB NVMe SSD
Software	Operating system	Ubuntu 22.04 LTS
	Deep learning framework	PyTorch 2.1.0
	Libraries	CUDA 12.2, cuDNN 8.9, OpenCV 4.8

	Programming language	Python 3.10
Training settings	Batch size	8
	Learning rate	1×10-4 (Cosine annealing)
	Optimizer	Adam (β1=0.9, β2=0.999)
	Epochs	200 (with early stopping)
Model hyperparameters	Positional encoding dimension	768
	Graph convolution order K	3
	Number of attention heads	8
	Feature dimension per head	96
	Transformer encoder layers	6
	GCN layers	3
	GRU hidden dimension	512
Loss function dettings	MSE weight λ_1	1.0
	Perceptual loss weight λ_2	0.2

4.2. Performance Verification of STTVideo Generation Method

To verify the performance of the GCN-STT smart city video generation framework in different application scenarios, the study compared the performance of a traditional video reconstruction model based on CNN and STT without graph convolution and multi-scale fusion mechanisms. First, the MSE, Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) of the three methods were analyzed in traffic road scenarios, public square scenarios, and environmental monitoring scenarios, as shown in Figure 5.

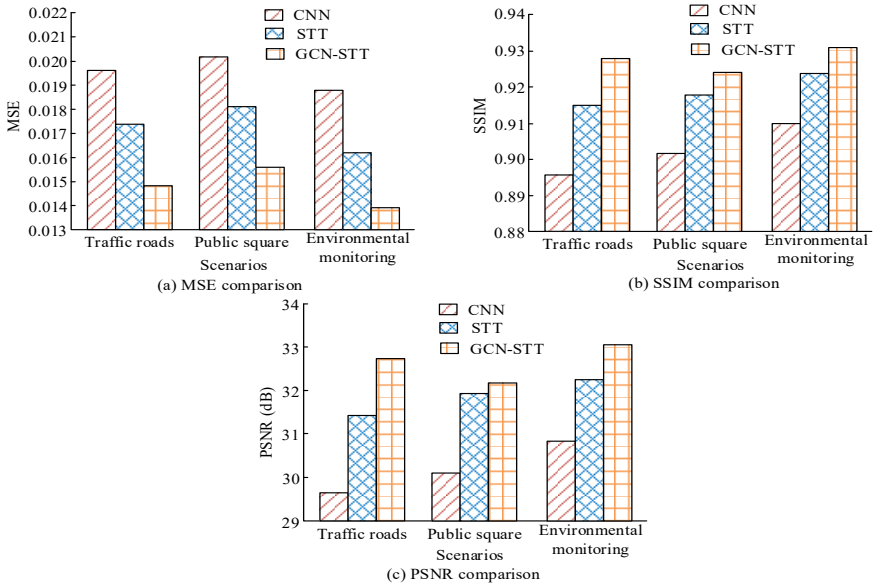
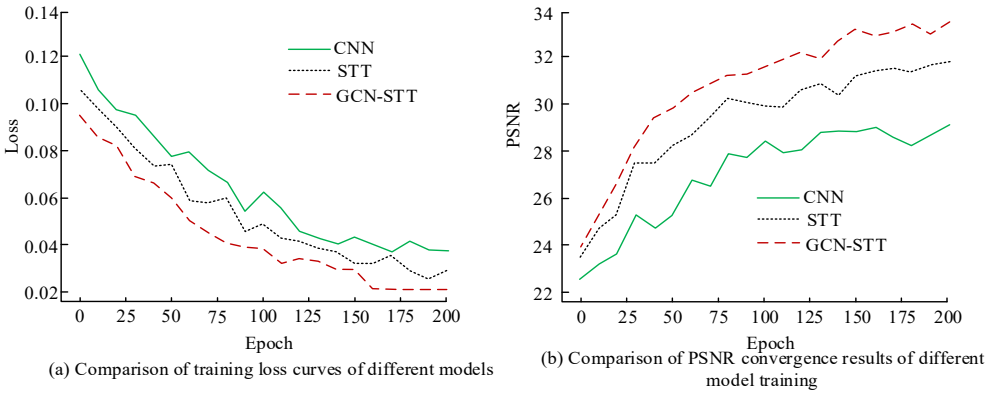


Figure 5. Model performance verification under different smart city scenarios

As shown in Figure 5(a), the MSE values of GCN-STT in the three scenarios were 0.0148, 0.0156, and 0.0139, respectively, which were 24.49%, 22.77%, and 26.06% lower than CNN, and 14.94%, 13.81%, and 14.20% lower than STT. This indicated that graph convolutional topology propagation effectively enhanced the spatial structure modeling capability and reduced pixel-level reconstruction errors. In Figure 5(b), the SSIM index of GCN-STT in the three smart city scenarios was 2.77% and 0.94% lower on average than the other two models, respectively. This may be because SSIM focuses more on the overall structural similarity of the image, while GCN-STT, in its modeling process, propagates spatial topological information through graph convolution and combines it with STT to extract global temporal features, thus prioritizing the reduction of pixel-level reconstruction errors and noise interference. In complex urban scenarios, this feature optimization method may produce slight changes in local structural consistency, leading to a slight decrease in the SSIM metric. As shown in Figure 5(c), GCN-STT achieves PSNRs of 32.74 dB, 32.18 dB, and 33.05 dB in the three scenarios, significantly outperforming CNN and STT, indicating that this method performs better in terms of pixel restoration accuracy and noise suppression. Based on multiple evaluation metrics such as MSE, SSIM, and PSNR, it can be seen that GCN-STT can effectively improve video reconstruction quality while maintaining

structural consistency, and demonstrates relatively stable video generation performance in the verified urban surveillance scenarios. The training effects of different models are shown in Figure 6.

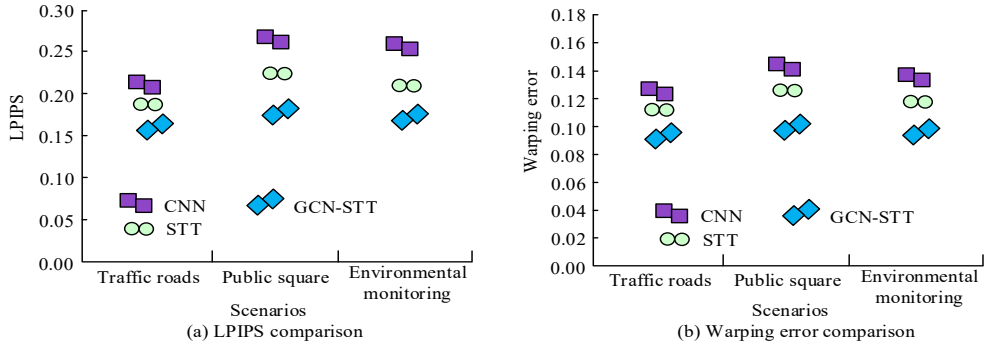


(a) Comparison of training loss curves of different models

(b) Comparison of PSNR convergence results of different model training

Figure 6. Comparison of training results of different models

As shown in Figures 6(a) and (b), the CNN curve exhibited the most significant fluctuations, with a slow decrease in training loss before stabilizing at a relatively high level, resulting in the lowest PSNR and relatively poor convergence quality. STT showed improved convergence speed in the early stages, but still experienced fluctuations and delays in the mid-to-late stages, ultimately achieving a PSNR of 31.92 dB. In contrast, the GCN-STT curve showed the fastest overall decrease and the smallest fluctuations, reaching a stable state after 150 rounds, ultimately achieving the lowest loss and the highest PSNR. Overall, GCN-STT not only improved the final performance but also enhanced the convergence efficiency and stability of the training process, which was of great significance for the deployment of large-scale smart city video generation tasks. Finally, the study further analyzed the Learned Perceptual Image Patch Similarity (LPIPS) and Warping Error of the three models, as shown in Figure 7.



(a) LPIPS comparison

(b) Warping error comparison

Figure 7. Verification of perceived quality and temporal consistency

As shown in Figures 7(a) and (b), compared with CNN, GCN-STT reduced LPIPS by 23.46%, 46.41%, and 48.28% in the three scenarios, respectively. Regarding warping error, GCN-STT reduced it by 35.79%, 44.12%, and 40.82% compared to CNN, and by 18.95%, 23.53%, and 20.41% compared to STT, respectively. This indicated that GCN-STT exhibited significant advantages in both perceptual quality and temporal consistency. Based on this, the study performed model ablation validation using Transformer as the baseline. The multi-scale attention mechanism is mainly used for feature fusion, and its effect is reflected in the final performance improvement of the complete model; therefore, it was not performed as a separate ablation module. Specific results are shown in Table 2.

Table 2. GCN-STT AblationValidation

STT	GCN	GRU	MSE	SSIM	PSNR (dB)
×	×	×	0.019	0.912	31.2
√	×	×	0.0181	0.918	31.92
×	√	×	0.0178	0.916	31.68
×	×	√	0.0176	0.919	31.77
√	√	×	0.0164	0.921	32.14
√	×	√	0.0168	0.923	32.06
×	√	√	0.0167	0.92	32.02
√	√	√	0.0148	0.931	32.74

As shown in Table 2, after introducing the spatiotemporal fusion mechanism to improve the Transformer, the MSE of the model was significantly reduced, and the PSNR was improved by 0.72 dB. This indicated that STT effectively enhanced the global modeling capability of video features by introducing spatial location information and spatiotemporal self-attention mechanism. After introducing GCN alone, the model further utilized the adjacency matrix to realize topological information propagation, reducing the MSE to 0.0178 and improving the PSNR by 0.48 dB. This showed that structural information had a significant gain for local region feature extraction. As a temporal state modeling module, GRU could capture inter-frame dynamic changes and further optimize feature representation, reducing the MSE to 0.0176. The performance was further improved when each module was combined in pairs, with STT+GCN showing the most significant effect, indicating a good complementary relationship between spatiotemporal modeling and topological propagation. Overall, STT mainly improved global perception and semantic consistency, GCN strengthened spatial structure and region association, and GRU stabilized temporal modeling and dynamic changes. The three worked together to achieve the performance gain of the model. Based on this, the study further compared the model complexity and inference time of GCN-STT with CNN and STT, as shown in Table 3.

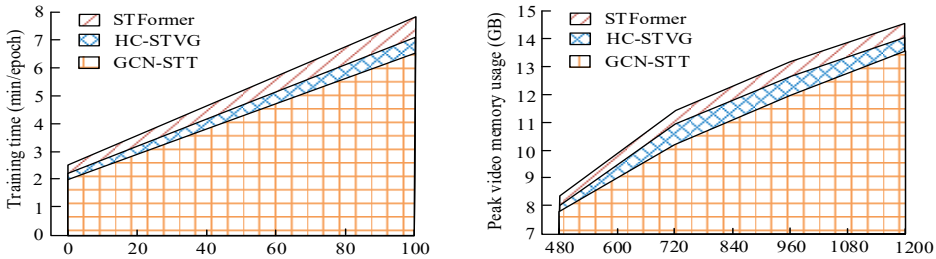
Table 3. Comparison of Model Complexity and Inference Time

Model	Parameters (M)	FLOPs (G)	Inference Time (s/frame)
CNN	18.32	32.71	0.029
STT	21.47	37.26	0.031
GCN-STT	23.64	41.83	0.034

As shown in Table 3, the GCN-STT model has a slightly increased parameter size and computational complexity compared to CNN and STT. Specifically, GCN-STT has 23.64 M parameters and a computational complexity of 41.83 G Floating Point Operations Per Second (FLOPs). This is mainly due to the introduction of graph convolution and multi-scale feature fusion mechanisms on top of STT, which enhances feature representation capabilities. Despite the increased complexity, the model's average inference time is only 0.034 s/frame, an increase of only about 0.003 s compared to STT, maintaining high computational efficiency overall and demonstrating good real-time processing potential under the current experimental setup.

4.3. Performance Comparison of Different Video Generation Methods

To further verify the computational efficiency and performance of the proposed GCN-STT framework in large-scale urban scenarios, STFormer proposed in reference [10] and HC-STVG proposed in reference [12] were selected as comparison methods to compare and analyze the performance of different models. The study first gradually increased the size of the training set and improved the input resolution, recording the training time per epoch and peak memory usage of the model under each configuration. The specific verification results are shown in Figure 8.



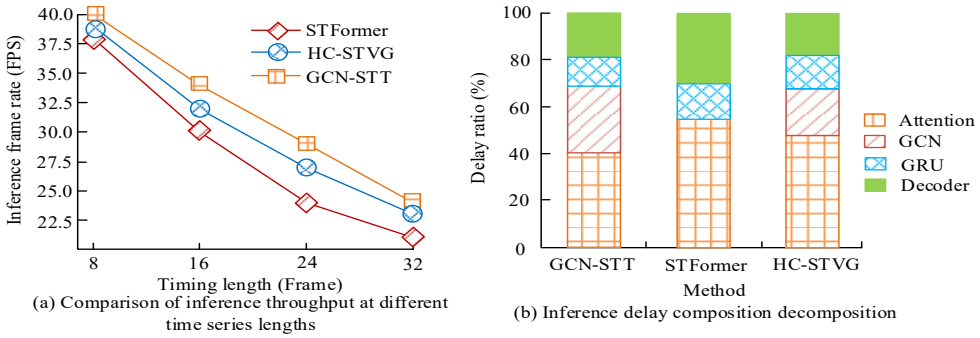
(a) Comparison of average training time per epoch for different methods

(b) Comparison of peak memory usage by different methods

Figure 8. Comparison of training efficiency and GPU memory usage

As shown in Figure 8(a), the training time per epoch for all three methods increased linearly with the data size increasing from 0% to 100%, but GCN-STT showed the slowest increase, significantly lower than STFormer and HC-STVG. Figure 8(b) shows that the memory usage of all three methods increased with the increase in input resolution, but GCN-STT showed the smallest increase, STFormer showed the fastest increase, and HC-STVG was in the middle. This indicated that GCN-STT effectively reduced the computational and memory overhead of global attention through sparse graph topology propagation, exhibiting better scalability and resource efficiency in large-scale and high-resolution scenarios. The study further compared the inference

throughput and latency composition decomposition of the three methods under different sequence lengths, as shown in Figure 9.

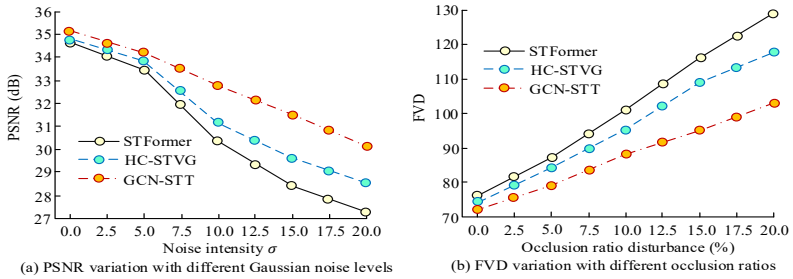


(a) Comparison of inference throughput at different time series lengths

(b) Inference delay composition decomposition

Figure 9. Comparison of inference delay components

As shown in Figure 9(a), the inference frame rate of all three methods decreased with increasing temporal length, but GCN-STT showed the smallest decrease, maintaining a consistently high throughput. This indicated that GCN-STT had better scalability under long sequence inputs. Figure 9(b) shows the latency composition decomposition results of the three methods. STFormer's latency was mainly concentrated in attention computation, while HC-STVG and GCN-STT had more balanced latency distributions. Among them, GCN-STT had the most reasonable proportion of each module and the lowest overall latency. In summary, GCN-STT combined high throughput and low latency in long temporal scenarios, resulting in the best inference efficiency. Finally, the study further analyzed the performance degradation of the three methods under conditions of increasing noise disturbance and occlusion ratio. In the experimental setup, environmental noise interference was simulated by adding Gaussian noise (noise intensity range of 0-20) with gradually increasing intensity to the test video frames, and occlusion scenes were constructed by superimposing occlusion areas at random locations, with the occlusion area ratio gradually increasing from 0 to 20%. The specific experimental results are shown in Figure 10.



(a) PSNR variation with different Gaussian noise levels

(b) FVD variation with different occlusion ratios

Figure 10. Comparison of robustness of different methods

As shown in Figure 10(a), the PSNR of all three methods decreased as the noise intensity σ increased from 0 to 20. GCN-STT exhibited the smallest degradation amplitude and the flattest curve, while STFormer decreased the fastest, with HC-STVG falling in the middle, indicating that GCN-STT was more stable against pixel-level noise. Combined with Figure 10(b), GCN-STT showed the smallest increase in Fréchet Video Distance (FVD), while STFormer increased the fastest. FVD is an indicator that measures the difference in distribution between the generated video and the real video; the lower the value, the closer the generated video is to the real video, and the higher the quality. This showed that GCN-STT could better maintain temporal consistency and structural stability under occlusion and noise interference, and its overall robustness was superior to the other two methods.

4.4. Statistical Reliability Analysis

To further verify the stability of the experimental results of different methods under random occlusion perturbation conditions where the occlusion area accounts for approximately 20% of the image region, the study conducted multiple repeated experiments on each model and statistically analyzed the mean and standard deviation of the generated video frames in terms of PSNR, SSIM, and LPIPS. Each experiment was repeated three times under the same environment, and the statistical results are shown in Table 4.

Table 4. Comparison of Video Generation Performance of Different Methods Under Noise and Occlusion Conditions

Method	PSNR (dB)	SSIM	LPIPS
CNN	29.84 ± 0.31	0.874 ± 0.012	0.183 ± 0.009
STT	31.05 ± 0.27	0.886 ± 0.010	0.162 ± 0.007
STFormer	31.42 ± 0.24	0.891 ± 0.009	0.158 ± 0.006
HC-STVG	31.36 ± 0.26	0.889 ± 0.010	0.160 ± 0.007
GCN-STT	32.18 ± 0.22	0.881 ± 0.009	0.148 ± 0.006

As shown in Table 4, GCN-STT achieves the highest PSNR value, indicating that the proposed method performs better in pixel reconstruction accuracy. It also achieves the lowest LPIPS value, suggesting that the generated video frames outperform the comparison methods in terms of perceptual quality. Although GCN-STT is slightly lower than some comparison models in SSIM, the overall difference is small, indicating that the proposed method focuses on reducing pixel-level errors and improving visual perceptual quality while maintaining structural similarity. Furthermore, the standard deviation shows that GCN-STT exhibits small fluctuations in its metrics across multiple experiments, indicating that the method has good stability and robustness in complex urban scenes and under conditions of noise and occlusion disturbances. The study selected surveillance video frames with local occlusion characteristics for qualitative comparative analysis, as shown in Figure 11.

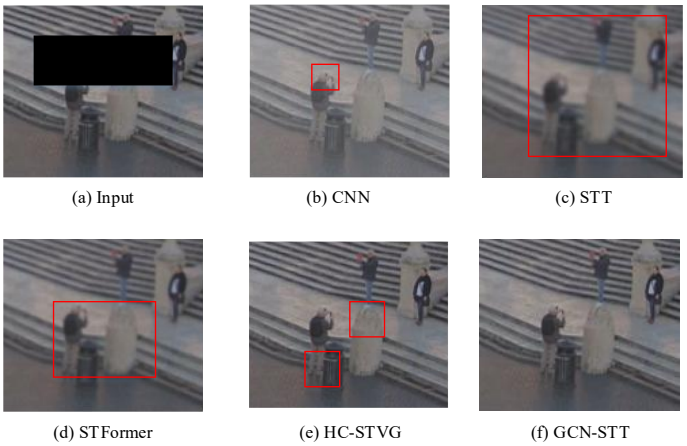


Figure 11. Qualitative comparison of generated video frames under occlusion conditions[1]

As shown in Figures 11(a)-(f), the CNN method exhibits significant blurring in local areas, making it difficult to recover the structural information of occluded targets. While the STT method improves the overall contour, it still suffers from noticeable loss of detail. STFormer and HC-STVG enhance the representation of local structures to some extent, but they still suffer from local distortion or blurring of details in complex scenes. In contrast, the GCN-STT method can better recover the target structurenear the occluded area and maintain spatial consistency between the background and the target.

5. Conclusion

To address the shortcomings of traditional video generation methods in complex smart city scenarios, such as insufficient spatial structure modeling, limited global temporal capture capabilities, and weak anti-interference performance, a smart city video generation framework integrating GCN and STT was proposed. Spatial topology propagation was achieved through graph convolution, spatiotemporal dependencies were captured using STT, and a multi-scale attention mechanism is combined with GRU for dynamic feature fusion. Experiments showed that in road, square, and environmental monitoring scenarios, GCN-STT reduced MSE by 24.49%, 22.77%, and 26.06% compared to CNN, and by 14.94%, 13.81%, and 14.20% compared to STT, respectively. In the LPIPS perception index, GCN-STT achieved the highest reduction of 48.28%. Compared to STFormer and HC-STVG, GCN-STT exhibited the best robustness to noise and occlusion perturbations. The results demonstrated that GCN-STT significantly enhanced the generation capability in complex dynamic scenarios through deep fusion of spatial topology and spatiotemporal features. The

[1] https://www.skylinewebcams.com/zh/webcam/italia/lazio/roma/piazza-di-spagna.htm

proposed video generation method achieved both high throughput and low latency in long-term temporal scenarios, exhibiting optimal inference efficiency and performance.

However, the research experiments were primarily validated using limited urban surveillance scenario data. When extreme occlusion, complex weather changes, or a significant increase in the number of camera nodes occur, the model's generated results may still exhibit local detail blurring or incomplete structural reconstruction. Furthermore, the proposed method relies to some extent on the graph structure construction strategy, and the definition of adjacency relationships may affect model performance. Simultaneously, as the number of camera nodes increases, the model's computational complexity may further increase. Therefore, future research will further explore more adaptive graph structure construction methods and conduct validation on larger-scale, multi-scenario urban video data to improve the model's scalability and generalization ability.

CRediT Author Contribution Statement

Bingchao Sun: Conceptualization, Methodology, writing – original draft, writing – review & editing, Supervision, Investigation.

Acknowledgement

This work was supported by the Teaching Reform Projects of Higher Vocational Education in Zhejiang Province (Project No. JG20240189) and the Zhejiang College of Construction Urban Lifeline Special Project (Project No. Z202503).

References

[1] Bing Liu, Zixuan Liu and Libo Fang, "Innovative Approaches to Assessing Urban Space Quality: A Multi-Source Big Data Perspective on Knowledge Dynamics", *Journal of the Knowledge Economy*, Print ISSN: 1868-7865, Online ISSN: 1868-7873, 28 March 2024, Vol. 16, No. 2, pp. 6803-6841, Published by Springer, DOI: 10.1007/s13132-024-01803-5, Available: https://link.springer.com/article/10.1007/s13132-024-01803-5.

[2] Ayushi Chahal, Santosh Reddy Addula, Anurag Jain, Preeti Gulia, Nasib Singh Gill *et al.*, "Systematic Analysis based on Conflux of Machine Learning and Internet of Things using Bibliometric analysis", *Journal of Intelligent Systems and Internet of Things*, Print ISSN: 2769-786X, Online ISSN: 2690-6791, June 2024, Vol. 13, No. 1, pp. 196-224, Published by American Scientific Publishing Group (ASPG), DOI: 10.54216/JISIoT.130115, Available: https://www.americaspg.com/articleinfo/18/show/2890.

[3] Xiana Chen, Junxian Yu, Yingying Zhu, Ruonan Wu and Wei Tu, "Short video-driven deep perception for city imagery", *Environment and Planning B: Urban Analytics and City Science*, Print ISSN: 2399-8083, Online ISSN: 2399-8091, March 2024, Vol. 51, No. 3, pp. 689-704, Published by SAGE Publications, DOI: 10.1177/23998083231193236, Available: https://journals.sagepub.com/doi/10.1177/23998083231193236.

[4] Ambreen Sabha and Arvind Selwal, "Towards Machine Vision-Based Video Analysis in Smart Cities: A Survey, Framework, Applications and Open Issues", *Multimedia Tools and Applications*, Print ISSN: 1380-7501, Online ISSN: 1573-7721, 9 August 2023, Vol. 83, No. 22, pp. 62107-62158, Published by Springer, DOI: 10.1007/s11042-023-16434-2, Available: https://link.springer.com/article/10.1007/s11042-023-16434-2.

[5] Asma Zahra, Mubeen Ghafoor, Kamran Munir, Ata Ullah and Zain Ul Abideen, "Application of Region-Based Video Surveillance in Smart Cities Using Deep Learning", *Multimedia Tools and Applications*, Print ISSN: 1380-7501, Online ISSN: 1573-7721, 27 December 2021, Vol. 83, No. 5, pp. 15313-15338, Published by Springer, DOI: 10.1007/s11042-021-11468-w, Available: https://link.springer.com/article/10.1007/s11042-021-11468-w.

[6] Himani Sharma and Navdeep Kanwal, "Video Surveillance in Smart Cities: Current Status, Challenges & Future Directions", *Multimedia Tools and Applications*, Print ISSN: 1380-7501, Online ISSN: 1573-7721, 24 June 2024, Vol. 84, No. 16, pp. 15787-15832, Published by Springer, DOI: 10.1007/s11042-024-19696-6, Available: https://link.springer.com/article/10.1007/s11042-024-19696-6.

[7] Wei Zhou, Li Yang, Lei Zhao, Runyu Zhang, Yifan Cui *et al.*, "Vision Technologies with Applications in Traffic Surveillance Systems: A Holistic Survey", *ACM Computing Surveys*, Print ISSN: 0360-0300, Online ISSN: 1557-7341, August 2025, Vol. 58, No. 3, pp. 1-47, Published by Association for Computing Machinery (ACM), DOI: 10.1145/3760525, Available: https://dl.acm.org/doi/10.1145/3760525.

[8] Jing Zhang, Xin Qi and Zheng Wen, "Deep-Learning-Empowered 3D Reconstruction for Dehazed Images in IoT-Enhanced Smart Cities", *Computers, Materials & Continua*, Print ISSN: 1546-2218, Online ISSN: 1546-2226, 13 April

2021, Vol. 68, No. 2, pp. 2807-2824, Published by Tech Science Press, DOI: 10.32604/cmc.2021.017410, Available: https://www.techscience.com/cmc/v68n2/42214/html.

[9] Yuxi Hu, Taimeng Fu, Guanchong Niu, Zixiao Liu and Man-On Pun, "3D Map Reconstruction Using a Monocular Camera for Smart Cities", *The Journal of Supercomputing*, Print ISSN: 0920-8542, Online ISSN: 1573-0484, 7 May 2022, Vol. 78, No. 14, pp. 16512-16528, Published by Springer, DOI: 10.1007/s11227-022-04512-5, Available: https://link.springer.com/article/10.1007/s11227-022-04512-5.

[10] Lishun Wang, Miao Cao, Yong Zhong and Xin Yuan, "Spatial-Temporal Transformer for Video Snapshot Compressive Imaging", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Print ISSN: 0162-8828, Online ISSN: 1939-3539, 29 November 2022, Vol. 45, No. 7, pp. 9072-9089, Published by IEEE Computer Society, DOI: 10.1109/TPAMI.2022.3225382, Available: https://ieeexplore.ieee.org/document/9965744.

[11] Tung Minh Tran, Doanh Cao Bui, Tam V. Nguyen and Khang Nguyen, "Transformer-Based Spatio-Temporal Unsupervised Traffic Anomaly Detection in Aerial Videos", *IEEE Transactions on Circuits and Systems for Video Technology*, Print ISSN: 1051-8215, Online ISSN: 1558-2205, September 2024, Vol. 34, No. 9, pp. 8292-8309, Published by IEEE, DOI: 10.1109/TCSVT.2024.3376399, Available: https://ieeexplore.ieee.org/document/10466765.

[12] Bachir Kaddar, Sid Ahmed Fezza, Zahid Akhtar, Wassim Hamidouche, Abdenour Hadid *et al.*, "Deepfake Detection Using Spatiotemporal Transformer", *ACM Transactions on Multimedia Computing, Communications, and Applications*, Print ISSN: 1551-6857, Online ISSN: 1551-6865, January 2024, Vol. 20, No. 11, pp. 1-21, Published by Association for Computing Machinery (ACM), DOI: 10.1145/3643030, Available: https://dl.acm.org/doi/10.1145/3643030.

[13] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin *et al.*, "Human-Centric Spatio-Temporal Video Grounding With Visual Transformers", *IEEE Transactions on Circuits and Systems for Video Technology*, Print ISSN: 1051-8215, Online ISSN: 1558-2205, 1 December 2021, Vol. 32, No. 12, pp. 8238-8249, Published by IEEE, DOI: 10.1109/TCSVT.2021.3085907, Available: https://ieeexplore.ieee.org/document/9446308.

[14] Jian Hua Qiao, Yu Ting Guan and Jin Ning Zhi, "Transformer-CNN Hybrid Network for Underwater Image Enhancement", *The Journal of Supercomputing*, Print ISSN: 0920-8542, Online ISSN: 1573-0484, 11 July 2025, Vol. 81, No. 10, pp. 1-17, Published by Springer, DOI: 10.1007/s11227-025-07612-0, Available: https://link.springer.com/article/10.1007/s11227-025-07612-0.

[15] Jongseon Kim, Hyungjoon Kim, HyunGi Kim, Dongjun Lee and Sungroh Yoon, "A Comprehensive Survey of Deep Learning for Time Series Forecasting: Architectural Diversity and Open Challenges", *Artificial Intelligence Review*, Print ISSN: 0269-2821, Online ISSN: 1573-7462, April 2025, Vol. 58, No. 7, pp. 1-95, Published by Springer, DOI: 10.1007/s10462-025-11223-9, Available: https://link.springer.com/article/10.1007/s10462-025-11223-9.

[16] Taiqu Lai, Lianglun Cheng, Depei Wang, Haiming Ye and Weiwen Zhang, "RMAN: Relational Multi-Head Attention Neural Network for Joint Extraction of Entities and Relations", *Applied Intelligence*, Print ISSN: 0924-669X, Online ISSN: 1573-7497, 1 July 2021, Vol. 52, No. 3, pp. 3132-3142, Published by Springer, DOI: 10.1007/s10489-021-02600-2, Available: https://link.springer.com/article/10.1007/s10489-021-02600-2.

[17] Diyuan Li, Zida Liu, Peng Xiao, Jian Zhou and Danial Jahed Armaghani, "Intelligent Rockburst Prediction Model with Sample Category Balance Using Feedforward Neural Network and Bayesian Optimization", *Underground Space*, Print ISSN: 2096-2754, Online ISSN: 2467-9674, October 2022, Vol. 7, No. 5, pp. 833-846, Published by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. and Tongji University, DOI: 10.1016/j.undsp.2021.12.009, Available: https://www.sciencedirect.com/science/article/pii/S2467967422000137.

[18] Muhammad Salman Latif, Rafaqat Kazmi, Nadia Khan, Rizwan Majeed and Sunnia Ikram, "Pest Prediction in Rice Using IoT and Feed Forward Neural Network", *KSII Transactions on Internet and Information Systems*, Print ISSN: not clearly stated on current journal site, Online ISSN: 1976-7277, January 2022, Vol. 16, No. 1, pp. 133-152, Published by Korean Society for Internet Information (KSII), DOI: 10.3837/tiis.2022.01.008, Available: https://itiis.org/digital-library/25250.

[19] Anam Fatima, Yi Yu, Janak Kapuriya, Julien Lalanne and Jainendra Shukla, "Semantic Frame Aggregation-Based Transformer for Live Video Comment Generation", *IEEE Transactions on Multimedia*, Print ISSN: 1520-9210, Online ISSN: 1941-0077, January 2025, Vol. 27, pp. 7821-7833, Published by IEEE, DOI: 10.1109/TMM.2025.3604921, Available: https://ieeexplore.ieee.org/document/11146668.

[20] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei *et al.*, "End-to-End Video Scene Graph Generation With Temporal Propagation Transformer", *IEEE Transactions on Multimedia*, Print ISSN: 1520-9210, Online ISSN: 1941-0077, 2024, Vol. 26, pp. 1613-1625, Published by IEEE, DOI: 10.1109/TMM.2023.3283879, Available: https://ieeexplore.ieee.org/document/10145598.