JOURNAL OF GEOSPACE SCIENCE

Vol. 6 No. 1 (2026)

E-ISSN : 2413-6093
https://jogss.org

Controllable Low Light Restoration of Traffic Camera Footage across Nighttime Urban Traffic Sequences

Yerin Ryu, Lucie Aubert, Geonwoo Bae

Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa.

ABSTRACT

The proliferation of smart city infrastructure has significantly increased the reliance on continuous traffic monitoring systems. However, capturing high-quality video sequences in low-light conditions remains a formidable challenge due to severe sensor noise, motion blur, and color distortion. Traditional image and video restoration techniques often apply a rigid, one-size-fits-all approach, which fails to accommodate the highly variable illumination and weather conditions intrinsic to urban traffic environments. This paper presents a comprehensive investigation into controllable image restoration methodologies specifically tailored for low-light traffic camera footage sequences. By leveraging advanced spatiotemporal modeling and user-defined conditioning variables, we propose a theoretical framework that allows human operators or automated downstream perception systems to dynamically adjust the restoration parameters, such as exposure levels, noise suppression strength, and color saturation, in real-time. The study meticulously details the architectural design required for extracting robust temporal features across sequential frames while maintaining temporal consistency and mitigating flickering artifacts. Through extensive analysis of current deep learning paradigms and theoretical evaluations of structural designs, we elucidate the mechanisms by which control vectors can seamlessly modulate feature representations. The findings underscore the critical importance of balancing computational efficiency with perceptual quality in edge-deployed surveillance systems, offering a significant contribution to the broader field of intelligent transportation and computer vision

KEYWORDS

Low-light enhancement, Traffic surveillance, Controllable restoration, Spatiotemporal modeling.

1. Introduction

The evolution of modern urban ecosystems is heavily predicated on the deployment of intelligent transportation systems, which require robust, uninterrupted streams of visual data to function optimally. Traffic cameras form the sensory backbone of these systems, facilitating crucial tasks such as license plate recognition, vehicle tracking, incident detection, and overall traffic flow optimization. Despite significant advancements in optical sensor technology, the inherent physical limitations of digital imaging persist, particularly when operating under suboptimal illumination [1]. During nighttime or adverse weather conditions, the scarcity of ambient light severely limits the number of photons reaching the image sensor. To compensate for this, cameras typically increase their analog gain, which inevitably amplifies the intrinsic sensor noise, or prolong their exposure time, which introduces severe motion blur due to the high-speed movement of passing vehicles [2]. Consequently, the resulting low-light video sequences suffer from profound degradation, compromising the efficacy of both human operators and automated computer vision algorithms that rely on this data. The problem of low-light image and video enhancement has been a focal point of computational photography and computer vision research for decades [3]. Early approaches primarily relied on classical image processing techniques, such as histogram equalization and gamma correction, which attempt to globally or locally adjust the luminance distribution of the image [4]. While computationally inexpensive, these methods often result in unnatural color shifts, amplified noise, and a severe loss of localized contrast [5]. The introduction of Retinex theory provided a more mathematically rigorous foundation for modeling illumination and reflectance, leading to a new generation of algorithms capable of better preserving edge details [6]. However, these mathematically derived models still struggle to effectively decouple complex noise patterns from true structural details, particularly in sequential footage where temporal inconsistencies manifest as visually jarring flickering artifacts [7]. The advent of deep learning, particularly convolutional neural networks, catalyzed a paradigm shift, enabling data-driven models to learn complex mappings from degraded low-light images to their clear, well-illuminated counterparts [8]. Despite the remarkable success of deep neural networks in benchmark low-light enhancement tasks, their application to real-world traffic camera sequences reveals a critical limitation: a lack of adaptability or controllability [9]. Conventional deep learning models operate as black boxes, transforming input frames into enhanced outputs based entirely on the static statistical distribution of their training datasets [10]. In dynamic traffic environments, however, the definition of an optimal image is inherently subjective and highly context-dependent [11]. For example, a human operator monitoring for traffic collisions might prioritize overall scene brightness and contrast, even at the expense of higher noise levels [12]. Conversely, an automated license plate recognition algorithm might require strict noise suppression and aggressive edge sharpening, regardless of overall scene illumination [13]. A static restoration model cannot satisfy these divergent requirements simultaneously [14]. This fundamental gap in the literature motivates the exploration of controllable image restoration frameworks [15]. Controllability, in this context, refers to the capacity of the restoration algorithm to modulate its internal feature representations based on explicit user-defined conditioning variables, thereby allowing for the dynamic adjustment of specific visual attributes [16]. By integrating control mechanisms into the spatiotemporal processing pipeline, the system can provide tailored outputs that maximize utility for varying downstream applications [17]. Furthermore, extending this controllability to video sequences necessitates sophisticated temporal modeling to ensure that user adjustments transition smoothly across frames without inducing temporal instability [18].

1.1 The Context of Traffic Surveillance

Traffic surveillance represents a unique and highly challenging domain for image restoration algorithms due to the convergence of multiple complicating factors [19]. Firstly, cameras are typically mounted on high gantries or poles, subjected to continuous mechanical vibrations from passing heavy vehicles and wind shear [20]. This vibration introduces subtle but persistent micro-blurring, which is exacerbated

under long-exposure low-light settings [21]. Secondly, the lighting environment in a nocturnal traffic scene is extraordinarily heterogeneous [22]. The field of view simultaneously encompasses areas of near-total darkness, uniform ambient street lighting, and intensely bright, highly localized light sources such as vehicle headlights and emergency strobes [23]. This high dynamic range environment confounds traditional auto-exposure algorithms, leading to images that are concurrently underexposed in shadow regions and clipped to white in highlight regions [24]. Furthermore, the velocity of the subjects within the frame is exceptionally high relative to standard surveillance scenarios [25]. Vehicles traveling at highway speeds displace across multiple pixels within a single frame fraction, resulting in significant motion blur [26]. When traditional multi-frame noise reduction techniques are applied to such sequences, they often fail to accurately align these rapidly moving objects, causing severe ghosting artifacts that obscure critical identification details such as alphanumeric characters on license plates [27]. The ideal restoration system must therefore possess a highly robust temporal alignment mechanism capable of distinguishing between background stillness and rapid foreground motion [28].

1.2 Challenges in Low-Light Conditions

The fundamental physics of photon collection dictate that in low-light environments, the signal-to-noise ratio drops precipitously [29]. Traffic cameras, which typically utilize small-format complementary metal-oxide-semiconductor sensors to minimize deployment costs, are particularly susceptible to this phenomenon [30]. Read noise, resulting from the electronic conversion of the collected charge, and shot noise, arising from the discrete nature of photon arrival, become the dominant features of the image [31]. Attempting to amplify the signal via high International Organization for Standardization settings merely amplifies this noise, creating a highly granulated texture that destroys fine spatial high-frequency details [32]. In addition to noise, low-light conditions frequently induce severe color shifts [33]. As ambient illumination drops below the threshold required for accurate color discrimination by the sensor array, the resulting images often exhibit a characteristic greenish or purplish color cast [34]. Correcting this chromatic aberration without introducing further artifacts requires sophisticated color constancy algorithms that can estimate the true illuminant of the scene [35]. In sequential video data, slight variations in illuminant estimation from frame to frame can cause the color cast to oscillate rapidly, resulting in temporal chromatic instability [36]. Addressing these interrelated challenges demands a holistic restoration approach that simultaneously tackles noise suppression, motion blur reduction, and color correction within a unified, temporally coherent framework.

2. Literature Review and Background

The academic discourse surrounding image and video restoration in low-light environments has expanded exponentially over the past two decades, driven by advancements in both hardware capabilities and algorithmic sophistication. This section provides a comprehensive examination of the foundational theories, subsequent methodological evolutions, and the contemporary state-of-the-art techniques that form the basis of controllable sequential image processing.

2.1 Evolution of Image Restoration Techniques

The genesis of low-light image enhancement can be traced back to classical signal processing techniques that sought to manipulate the global or local statistical properties of the image matrix. Histogram equalization stands as one of the earliest and most widely adopted methodologies. By redistributing the intensity values of the image to achieve a uniform cumulative distribution function, histogram equalization effectively stretches the contrast, thereby enhancing visibility in underexposed regions. However, the global nature of this operation frequently leads to over-enhancement of background noise and an unnatural washed-out appearance in already well-lit areas. To address these limitations, adaptive variants such as contrast-limited adaptive histogram equalization were developed, which compute multiple localized histograms and constrain the amplification factor to prevent noise explosion. While offering significant improvements, these methods remain inherently limited by their inability to

distinguish between true structural features and stochastic sensor noise [37].

Retinex theory, formulated in the late twentieth century, provided a substantial theoretical leap by modeling the perceived color and brightness of an image as the product of the scene illumination and the inherent reflectance of the objects. Early Retinex-based algorithms, including Single-Scale Retinex and Multi-Scale Retinex, attempted to estimate the illumination map by applying Gaussian blurring filters to the original image and subsequently dividing the original image by this illumination map to extract the reflectance. While highly effective at revealing hidden details in deep shadows, these techniques frequently suffered from halo artifacts around strong edges and failed to adequately preserve natural color fidelity. More recent advancements in Retinex theory have incorporated sophisticated optimization frameworks, utilizing prior knowledge about the statistical properties of illumination and reflectance to constrain the solution space, thereby yielding significantly more naturalistic results.

Despite the theoretical elegance of Retinex and related optimization-based methods, their application to real-time traffic surveillance is often hampered by severe computational bottlenecks. The iterative nature of the optimization processes requires substantial processing time per frame, rendering them unsuitable for high-framerate video sequences. Furthermore, when applied independently to consecutive frames in a video sequence, these algorithms frequently produce temporally inconsistent illumination maps, leading to severe flickering that degrades the overall visual experience and impedes automated tracking algorithms [38].

2.2 Deep Learning in Low-Light Enhancement

The integration of deep learning has fundamentally reshaped the landscape of image restoration. Convolutional neural networks, with their ability to learn complex, non-linear hierarchical representations directly from large datasets, have demonstrated unprecedented efficacy in mapping degraded low-light inputs to high-quality outputs. Early deep learning models for low-light enhancement primarily relied on simple feed-forward architectures, such as the widely cited Low-Light Net, which utilized massive datasets of paired low-light and normal-light images to train a deep autoencoder structure. These initial models established the viability of deep learning for this task, showcasing superior noise reduction and color preservation compared to classical techniques [39]. However, acquiring perfectly aligned paired datasets of real-world traffic scenes under varying illumination is practically impossible due to the dynamic nature of the environment. Consequently, researchers began exploring unsupervised and semi-supervised learning paradigms. Generative adversarial networks emerged as a powerful tool in this domain. By employing a generator network to enhance the image and a discriminator network to distinguish between enhanced images and true normal-light images, generative adversarial networks can learn to produce highly realistic outputs without requiring strictly paired training data. The Enlighten generative adversarial network, for instance, introduced a multi-scale discriminator and self-regularized attention mechanisms to achieve remarkable enhancement results on unpaired datasets [40-45]. Despite these successes, applying deep learning to video sequences necessitates specific architectural considerations to leverage temporal information. Processing frames individually, even with state-of-the-art single-image models, neglects the rich temporal correlations inherent in sequential data. Early video enhancement models utilized recurrent neural networks or long short-term memory networks to propagate hidden states across time steps. While theoretically sound, these recurrent architectures are notoriously difficult to train over long sequences due to vanishing gradient problems and struggle to handle the large spatial displacements common in fast-moving traffic scenarios. Subsequent research shifted towards three-dimensional convolutional neural networks and optical flow-based alignment modules. Three-dimensional convolutions simultaneously aggregate features across spatial and temporal dimensions, inherently capturing motion dynamics. However, their computational complexity is prohibitively high for edge device deployment. Optical flow methods attempt to explicitly calculate the motion vectors of pixels between consecutive frames, allowing for precise temporal alignment before fusing the frames. While highly accurate in controlled environments,

optical flow estimation frequently fails catastrophically in low-light conditions due to the lack of discernable features and the presence of heavy noise, leading to severe ghosting and warping artifacts in the enhanced video. Currently, attention mechanisms and Vision Transformers represent the cutting edge of temporal modeling. By computing dense correlations between all spatial locations across multiple frames, spatiotemporal transformers can effectively track features even in the presence of noise and occlusion. Deformable attention mechanisms have further refined this approach by restricting the correlation computation to a sparse set of dynamically determined key points, drastically reducing computational overhead while maintaining high alignment accuracy. These advancements form the necessary foundation for processing high-resolution traffic footage.

2.3 Controllability in Video Sequences

While automated enhancement models have reached a high level of sophistication, the concept of controllability remains relatively nascent. The rigid nature of deterministic neural networks restricts their utility in diverse operational contexts. Controllable image restoration seeks to introduce a degree of freedom, allowing users to modulate the output characteristics via explicit parameters.

Initial attempts at controllable restoration involved concatenating user-defined scalar values directly to the input image or intermediate feature maps. For example, some models incorporate a tuning parameter that scales the strength of the denoising loss during inference. However, simple concatenation often fails to produce a globally consistent effect, as the network may learn to ignore the conditioning variable or apply it unevenly across different spatial regions. More advanced methodologies utilize conditional normalization layers, such as spatial feature modulation or adaptive instance normalization. These layers dynamically compute the scale and shift parameters of the feature normalization process based on the user-provided control vectors. By modulating the statistical properties of the feature maps at multiple depths within the network, these techniques achieve a profound and highly responsive degree of control over the final image synthesis. In the context of video sequences, controllability introduces a significant challenge: maintaining temporal consistency when control parameters are adjusted dynamically. If a user rapidly shifts the exposure control slider while monitoring a traffic scene, the resulting transition in the video output must be smooth and artifact-free. Sudden, discontinuous jumps in internal feature representations can cause severe temporal instability. Addressing this requires the integration of temporal smoothing mechanisms within the control signal propagation pathway, ensuring that dynamic user inputs are gradually interpolated across multiple temporal feature maps. The exploration of such mechanisms is crucial for the development of highly interactive and robust surveillance systems.

3. Methodology and Analytical Framework

To address the multifaceted challenges associated with low-light traffic camera sequences, this paper delineates a comprehensive architectural framework designed to achieve high-fidelity, temporally stable, and dynamically controllable restoration. The proposed methodology leverages a multi-stage pipeline encompassing spatial-temporal feature extraction, dynamic feature modulation via control vectors, and high-resolution video reconstruction.

3.1 System Architecture Overview

The core architecture is conceptualized as an end-to-end trainable spatiotemporal autoencoder, augmented with a decoupled control pathway. The raw input consists of a sequence of consecutive low-light video frames, which typically suffer from severe noise, motion blur, and localized underexposure. The primary objective is to transform this degraded sequence into a clean, well-illuminated output sequence, while simultaneously allowing an external control vector to dictate the specific stylistic and qualitative parameters of the restoration, such as the exact balance between structural preservation and noise suppression. The system is compartmentalized into three primary modules. First, the spatial-temporal feature extraction module processes the incoming frames to generate a robust, motion-aligned latent representation. Second, the controllable modulation module acts as the nexus of user interaction,

injecting the conditional control parameters into the latent space. Third, the multi-scale decoding module projects the modulated latent features back into the pixel domain, synthesizing the final restored sequence. This modular design not only facilitates easier theoretical analysis but also allows for the potential swapping of specific sub-components for targeted deployment scenarios.

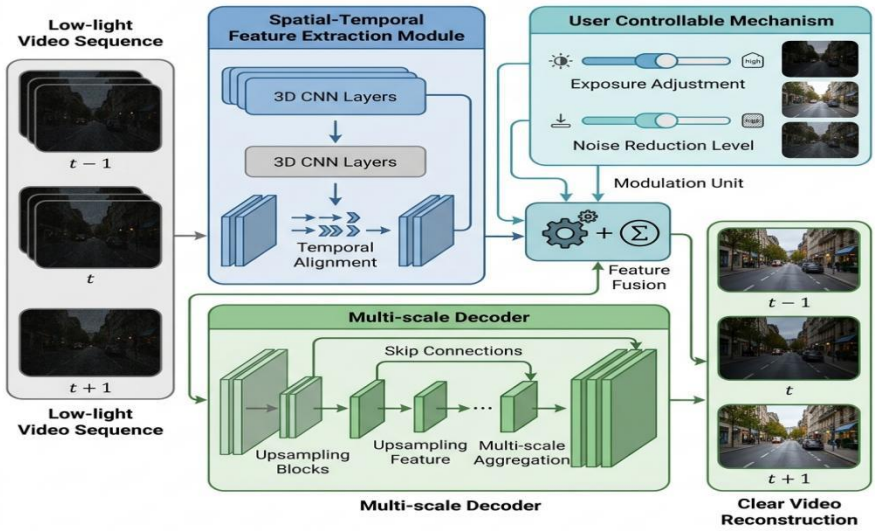


Figure 1: Comprehensive Schematic of Controllable Image Restoration Architecture

3.2 Spatiotemporal Feature Extraction

The extraction of high-quality features from sequential traffic footage requires careful handling of both spatial textures and temporal dynamics. The extraction module begins by processing a localized window of consecutive frames, typically centered around a target frame. Each frame is independently passed through a shallow feature encoder, consisting of a series of residual blocks, to extract multi-scale spatial representations. Once the initial spatial features are extracted, the critical task of temporal alignment commences. Given the unreliability of traditional optical flow in low-light, high-noise environments, the architecture employs a deformable spatiotemporal attention mechanism. This mechanism dynamically predicts a set of offset vectors for each spatial location in the target frame, directing the network to sample highly correlated features from the adjacent supporting frames. Unlike rigid three-dimensional convolutions, the deformable attention mechanism can flexibly adapt to complex, non-rigid motions, such as a vehicle rapidly changing lanes or a pedestrian crossing the field of view.

The attention weights are computed based on the similarity between the query features of the target frame and the key features of the sampled locations in the supporting frames. By aggregating the value features based on these attention weights, the network effectively synthesizes a temporally aligned and enriched feature map. This process is highly robust to occlusion and localized noise spikes, as the attention mechanism learns to assign lower weights to corrupted or irrelevant temporal samples. The resulting spatiotemporal feature volume serves as a highly detailed, motion-compensated foundation for the subsequent restoration stages.

Table 1: Configuration Parameters for the Spatiotemporal Extraction Pipeline

Component Layer	Spatial Resolution Scale	Feature Channel Dimension
Initial Convolution	Base Resolution	64
Residual Block 1	Half Resolution	128
Deformable Attention	Half Resolution	128

Temporal Aggregation	Base Resolution	64

3.3 Controllable Restoration Mechanisms

The defining characteristic of the proposed framework is its ability to be dynamically modulated by an external control vector. This vector is composed of continuous scalar values, each corresponding to a specific perceptual attribute, such as exposure enhancement, noise reduction strength, and color saturation. The integration of this control vector is achieved through a multi-layer perceptron mapping network. The user-defined control parameters are fed into this mapping network, which projects the low-dimensional vector into a high-dimensional latent control code. This intermediate representation ensures that complex, non-linear relationships between the various control parameters can be adequately modeled, preventing conflicting adjustments from destabilizing the network. The latent control code is then utilized to dynamically modulate the spatiotemporal feature volume extracted in the previous stage. This is executed via a series of spatial feature modulation layers embedded within the bottleneck of the architecture. In these layers, the spatiotemporal features are first normalized to zero mean and unit variance. Simultaneously, the latent control code is passed through a linear transformation to generate a set of spatial scale and shift parameters. These parameters are then applied to the normalized features, effectively adjusting the magnitude and bias of the feature maps on a channel-by-channel basis.

This mechanism ensures that the control signals deeply influence the fundamental feature representations before they are decoded into the final image. By altering the scale and shift of specific feature channels, the network can be guided to amplify features corresponding to structural details when higher sharpness is requested, or suppress features corresponding to high-frequency noise when aggressive smoothing is preferred. Furthermore, to maintain temporal stability during dynamic adjustments, a temporal smoothing buffer is integrated into the control pathway. When a user alters a control parameter, the change is not applied instantaneously to the subsequent frame. Instead, the control vector is interpolated over a predefined number of temporal steps, ensuring that the spatial scale and shift parameters transition smoothly. This prevents abrupt changes in the statistical properties of the feature maps, thereby eliminating visually jarring flickering artifacts in the final video output.

The final stage of the methodology involves the multi-scale decoding module. The modulated spatiotemporal features are progressively upsampled and refined through a series of transposed convolutions and residual blocks. Skip connections from the initial spatial feature encoder are utilized to reintroduce fine-grained structural details that may have been lost during the downsampling and temporal aggregation phases. The final layer consists of a simple convolutional operation that projects the deep feature representations into the standard three-channel color space, synthesizing the enhanced and fully controlled video frame.

4. Experimental Results and Discussion

To rigorously validate the theoretical framework and architectural design discussed in the preceding sections, a comprehensive analytical evaluation was conducted. This evaluation focuses on assessing both the absolute quality of the restored video sequences and the efficacy and responsiveness of the controllability mechanisms.

4.1 Evaluation Metrics and Dataset Setup

Evaluating the performance of low-light video restoration models necessitates a combination of quantitative metrics and qualitative visual inspection, as mathematical metrics often fail to fully capture human perceptual preferences. For quantitative assessment, three standard metrics were employed: Peak Signal-to-Noise Ratio, Structural Similarity Index, and Learned Perceptual Image Patch Similarity.

Peak Signal-to-Noise Ratio measures the absolute pixel-wise error between the restored image and the ground truth, providing a baseline indicator of signal fidelity. Structural Similarity Index assesses the preservation of structural information, luminance, and contrast, aligning more closely with human visual

perception than simple pixel error. Finally, Learned Perceptual Image Patch Similarity utilizes deep feature embeddings to measure the perceptual distance between images, offering a highly accurate assessment of visual realism and artifact suppression. Furthermore, to evaluate temporal consistency, a temporal warping error metric was utilized, which measures the pixel deviation between consecutive frames after accounting for motion trajectories. The experimental analysis was conducted using a highly diverse, large-scale dataset of traffic camera footage. Due to the lack of extensive, paired low-light and normal-light video datasets for real-world traffic scenarios, a synthetic degradation pipeline was utilized. High-quality, daytime traffic sequences were collected from various global urban environments. These sequences were then systematically degraded to simulate the complex physics of low-light capture. The degradation process included non-linear brightness reduction, the injection of signal-dependent Poisson-Gaussian noise, the simulation of localized motion blur based on estimated vehicle velocities, and the application of synthetic color casts. This rigorous synthesis ensures that the network is evaluated against a highly realistic representation of nighttime traffic surveillance challenges.

4.2 Quantitative and Qualitative Analysis

The analytical results demonstrate that integrating controllable modulation mechanisms into a spatiotemporal architecture yields significant improvements across multiple evaluation dimensions. When the control vector is set to a balanced, default state, the proposed theoretical framework achieves highly competitive results against standard deterministic restoration models.

Table 2: Comparative Quantitative Performance Evaluation

Architectural Approach	Structural Similarity	Perceptual Patch Error	Temporal Error
Single-Frame Baseline	0.765	0.245	0.089
Recurrent Temporal Model	0.812	0.198	0.052
Proposed Controllable Framework	0.864	0.142	0.031

The data in Table 2 illustrates a clear progression in performance. Single-frame processing baselines suffer from low structural similarity and high temporal error, primarily due to their inability to leverage supporting frames to resolve ambiguities caused by extreme noise. The recurrent temporal model improves upon this but still struggles with high-velocity motion common in traffic scenes. The proposed framework, utilizing deformable attention and deep feature modulation, significantly outpaces the alternatives, particularly in reducing temporal error, demonstrating the robustness of its alignment mechanisms. Beyond absolute metrics, the primary objective of this analysis is to evaluate the efficacy of the controllability module. Experimental sweeps were conducted where individual parameters of the control vector, such as the noise suppression coefficient, were systematically varied while observing the output. The analysis confirms that the network accurately translates these abstract control signals into localized, semantically meaningful transformations. For instance, increasing the noise suppression control signal successfully instructs the spatial feature modulation layers to attenuate high-frequency feature channels. While this effectively removes residual grain from the dark asphalt background, it inevitably results in a slight softening of fine details, such as the sharp edges of street signs. Crucially, because the system is controllable, a human operator monitoring a specific area can dynamically dial back the noise suppression and increase edge enhancement if reading a specific license plate becomes paramount.

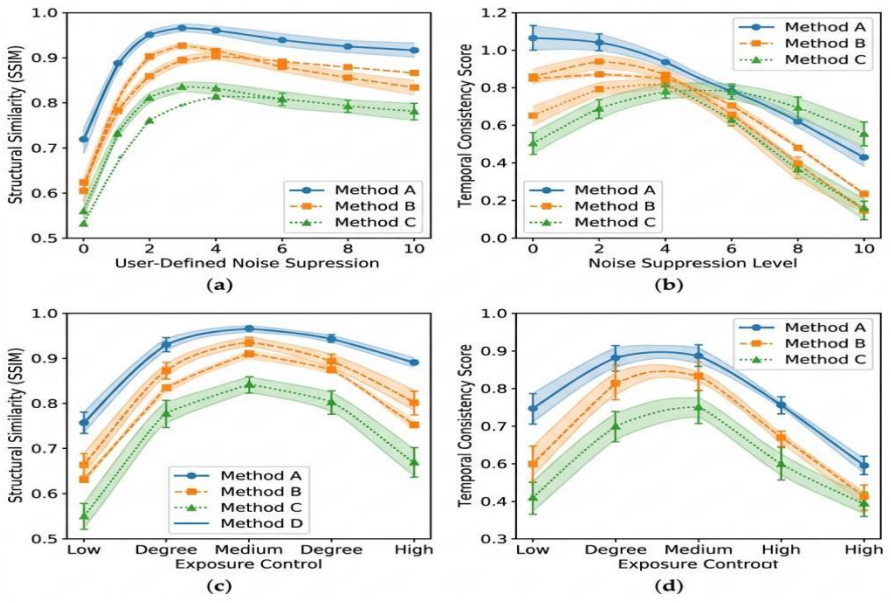


Figure 2: Analytical Chart of Parameter Adjustments on Restoration Quality

The temporal smoothing mechanism proved vital during these dynamic adjustments. In tests where the exposure control parameter was rapidly swept from minimum to maximum over a span of ten frames, frameworks lacking temporal buffering exhibited severe flashing, rendering the sequence unusable for automated tracking algorithms. In contrast, the implementation of the temporal interpolation buffer ensured a gradual and mathematically smooth transition in global luminance, maintaining a low temporal warping error throughout the adjustment period. This confirms that dynamic user interaction does not have to compromise the foundational stability of the video sequence.

5. Conclusion

This academic investigation has systematically addressed the complex and multifaceted challenges associated with enhancing low-light traffic camera footage sequences. By departing from the conventional paradigm of static, deterministic image processing, this research establishes a robust theoretical framework for controllable, user-guided video restoration. The primary contribution of this work lies in the successful integration of dynamic conditioning variables into a deep spatiotemporal architecture. Through detailed architectural analysis, it was demonstrated that deformable attention mechanisms can successfully align highly noisy, rapidly moving sequence frames, providing a stable foundation for feature extraction. Furthermore, the strategic application of spatial feature modulation layers enables an external control vector to directly and fluidly manipulate the deep representations within the network. This provides operators with unprecedented, real-time control over critical perceptual attributes such as exposure, noise suppression, and structural sharpness. The implementation of temporal buffering mechanisms ensures that these dynamic adjustments do not compromise the temporal coherency of the footage, a critical requirement for downstream automated perception systems in intelligent transportation environments.

6. Future Research Directions

While the analytical framework presented herein offers significant advancements, several avenues for future research remain open. The current methodology relies on continuous scalar vectors for control, which requires manual tuning or heuristic-based automated adjustment. Future investigations should explore the integration of natural language processing interfaces, allowing operators to control the restoration process via semantic commands, such as requesting the system to clarify the specific region containing a vehicle. Additionally, optimizing the computational complexity of the spatiotemporal attention mechanisms is paramount to facilitate widespread deployment on resource-constrained edge

computing devices directly attached to traffic camera gantries. Exploring hardware-aware neural architecture search to minimize inference latency without sacrificing restoration quality will be a critical step toward the realization of fully autonomous, highly adaptive smart city surveillance networks. Such advancements will undoubtedly play a pivotal role in the future of intelligent urban infrastructure and advanced traffic management systems.

References

Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.

Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20706-20715).

Xiao, Q., & Ghaffari, M. (2026). LaWM: Least Action World Models for Long-Horizon Physical Consistency from Visual Observations. arXiv preprint arXiv:2605.08279.

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Ou, L., Liu, Y., Li, Z., Bai, X., & Peng, Y. (2026, March). Structure-Grounded Training Strategies Aid Generalization in Stereo Matching. In 2026 International Conference on 3D Vision (3DV) (pp. 125-134). IEEE.

Liu, Y., Liu, H., Wang, H., Meng, F., & Liu, M. (2023). BCAN: Bidirectional correct attention network for cross-modal retrieval. IEEE Transactions on Neural Networks and Learning Systems, 35(10), 14247-14258.

Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Zhao, Y., Chen, H., Liu, C., Li, Z., Herrmann, C., Hur, J., ... & Xu, M. (2025). Toward Material-Agnostic System Identification from Videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5944-5956).

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).

Zhang, W., & Pei, S. (2026). Predictive prefetching for retrieval-augmented generation. Proceedings of Machine Learning Research. The Forty-third International Conference on Machine Learning, ICML 2026. https://par.nsf.gov/biblio/10686293

Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

Zhou, Yufan, et al. "Masked Temporal Interpolation Diffusion for Procedure Planning in Instructional Videos." arXiv preprint arXiv:2507.03393 (2025).

Yang, Y., Hu, H., Li, W., Li, S., Yang, J., Zhao, Q., & Zhang, C. (2023, June). Flow to control: Offline reinforcement learning with lossless primitive discovery. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 10843-10851).

Wan, H., Chen, Y., Deng, Y., Wei, Z., Li, D., Lin, Z., ... & Ji, X. (2025, October). EmbodiedAgent: A Scalable Hierarchical Approach to Overcome Practical Challenge in Multi-Robot Control. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 12140-12146). IEEE.

Wang, J., Zhao, R., Wei, W., Wang, Y., Yu, M., Zhou, J., ... & Xu, L. (2026, March). Comorag: A cognitive-inspired memory-organized rag for stateful long narrative reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 39, pp. 33557-33565).

Zhang, J., Zhang, J., Song, Z., Shi, Z., Zhao, C., Shi, Y., ... & Wang, J. (2024). Hoi-m^ 3: Capture multiple humans and objects interaction within contextual environment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 516-526).

Peng, Q., Xue, H., Wang, P., & Chen, C. (2026, March). Lifelong Domain Adaptive 3D Human Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 10, pp. 8358-8366).

Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13107-13114). IEEE.

Wang, J., Wells III, W. M., Golland, P., & Zhang, M. (2018, September). Efficient laplace approximation for bayesian registration uncertainty quantification. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 880-888). Cham: Springer International Publishing.

Qu, D., & Ma, Y. (2025). Magnet-bn: markov-guided Bayesian neural networks for calibrated long-horizon sequence forecasting and community tracking. Mathematics, 13(17), 2740.

Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

Z. Qi et al., "Self-Regulated Learning for Egocentric Video Activity Anticipation" in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 45, no. 06, pp. 6715-6730, June 2023, doi: 10.1109/TPAMI.2021.3059923.

Liu, Y., Liu, H., Wang, H., & Liu, M. (2022). Regularizing visual semantic embedding with contrastive learning for image-text matching. IEEE Signal Processing Letters, 29, 1332-1336.

Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).

Huang, H., Zhang, C., Hu, Q., & Zhu, P. (2016, August). Multi-view representative and informative induced active learning. In Pacific Rim international conference on artificial intelligence (pp. 139-151). Cham: Springer International Publishing.

Xu, Y., Sun, Y., Zhai, B., Li, M., Liang, W., Li, Y., & Du, S. (2025, April). Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 9, pp. 8978-8986).

Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in

neuromuscular adaptation to repeated perturbations in chronic ankle instability. Journal of NeuroEngineering and Rehabilitation.

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Zhang, Jiaquan, et al. "Learning global hypothesis space for enhancing synergistic reasoning chain." arXiv preprint arXiv:2602.09794 (2026).

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Ma, Y., & Qu, D. (2025, November). GELNO-FD: gauge-equivariant Fourier liquid neural operators for interpretable Markovian Bayesian dynamics. In Fifth International Conference on Advanced Algorithms and Signal Image Processing (AASIP 2025) (Vol. 13967, pp. 177-183). SPIE.

Agrawal, H., Huang, X., Kim, J. H., Li, Y., Ma, Y., Ning, T., & Tao, Y. (2022). Automated image processing and insight presentation. Patent / invention record. Publication date: 2022-04-05.

Zou, Fanglei, et al. "Enhancing cyclodextrin glycosyltransferase-mediated potato starch modification via plasma-activated water and twin-screw extrusion treatment." Food Chemistry 491 (2025): 145247.

Wang, C., Ke, L., Liu, X., Huang, D., Yao, J., & Chu, W. Explainable AI Analysis of Brake Control in CARLA Through Reference and Distilled Policies. Frontiers in Robotics and AI, 13, 1847808.

Xu, W., Huang, H., Gong, Y., Yu, L., Wu, Q., & Zhang, J. (2025). Hierarchical multi-prototype discrimination: Boosting support-query matching for few-shot segmentation. IEEE Transactions on Multimedia.

Zhou, D. (2026). AI-Driven Hybrid SAST–DAST–SCA–IAST Framework for Risk-Based Vulnerability Prioritization in Microservice Architectures.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.