ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

HS-VidNet: An Efficient Video Denoising Network for Autonomous Driving Based on Frequency-Spatial Reconstruction Mechanism

Pan Wang[a,*], Lei Ding[b]

School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, Shaanxi, China
[a]lding@sust.edu.cn, [b]1538916116@qq.com

Abstract: *Visual perception is critical for Autonomous Driving Systems (ADS) under extreme weather conditions such as rain, fog, and low illumination. This paper proposes HS-VidNet, an efficient and lightweight video denoising network. The method integrates the Spatial and Channel Reconstruction Convolution (SCConv) module within a U-Net architecture for feature reconstruction. This module utilizes Spatial Reconstruction Units (SRU) and Channel Reconstruction Units (CRU) to reshape feature flows. It suppresses non-discriminative redundancy in regions like the sky and road surface while concentrating limited computational resources on critical semantic topologies, such as road edges and lane lines. This significantly reduces computational overhead. Furthermore, the HiLo attention mechanism is introduced to compensate for the loss of high-frequency details during denoising. The high-frequency branch extracts fine geometric textures within local windows. Concurrently, the low-frequency branch models global long-range dependencies through a down-sampling strategy. This enhances the preservation of critical structural information and maintains feature consistency. Experiments were conducted on the CARLA-AWC dataset using the CARLA simulator. Results demonstrate that HS-VidNet achieves a stable inference speed of 72 FPS with a computational cost of only 87.2 GFLOPs. Its efficiency outperforms existing SCUNet and SwinIR-Light algorithms. In terms of accuracy, the model achieves an SSIM of 0.912, effectively balancing environmental noise removal with the preservation of critical structures.*

Keywords: *Video Denoising; Autonomous Driving; Lightweight Network; Frequency Decoupling; SCConv; HiLo Attention*

1. Introduction

Visual perception is the cornerstone of environmental modeling for Autonomous Driving Systems (ADS). Its reliability is directly linked to the safety of decision-making systems. In real-world road scenarios, adverse weather such as night-time low illumination, heavy rain, or dense fog often couples with inherent sensor thermal noise. This leads to significant degradation of the video stream. Such visual degradation severely undermines the feature representation of critical targets, including lane lines, traffic signs, and pedestrians. Consequently, it weakens the robustness of downstream perception tasks [1]. Therefore, designing a video denoising algorithm that balances real-time requirements for edge devices with the ability to restore high-frequency textures has become a core demand in computer vision and intelligent transportation.

Existing video denoising methods are mainly categorized into traditional filtering algorithms and deep learning-based algorithms [2]. Traditional methods, represented by Block-matching and 3D filtering (BM3D), possess good mathematical interpretability [1-3]. However, they tend to over-smooth images and lose critical textures when processing non-stationary noise. In recent years, Convolutional Neural Networks (CNN) have achieved significant progress in denoising performance due to their powerful non-linear feature fitting capabilities [4]. For instance, DnCNN introduces residual learning to simplify the fitting of noise mappings. RIDNet utilizes a residual-in-residual structure and channel attention mechanisms to improve the extraction accuracy of salient features. Furthermore, symmetric encoder-decoder architectures represented by U-Net have become mainstream backbones for video reconstruction tasks through skip connections [5]. Nevertheless, the deployment of CNNs on vehicles is still limited by inherent operator constraints. The weight-sharing sliding window mechanism fails to distinguish content importance. This results in computational power being wasted on non-discriminative redundant regions

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

such as the sky and road surfaces. Additionally, the inductive bias of local receptive fields limits the network's global context modeling. Consequently, it is difficult for the model to maintain the geometric consistency of road topology while suppressing large-scale noise [6].

To overcome the limitations of local modeling, Vision Transformer (ViT) and its variants have been introduced to low-level vision tasks. Typical methods like SwinIR employ a shifted window mechanism to achieve feature interaction between non-overlapping regions [7]. Restormer optimizes inference efficiency in high-resolution scenarios by calculating transposed attention in the channel dimension. Although self-attention mechanisms can effectively capture long-range dependencies, their computational complexity increases quadratically with image resolution. This leads to a substantial increase in memory usage and inference latency. Moreover, the inherent low-pass filtering bias of this mechanism often induces edge blurring. This damages critical high-frequency details required for driving decisions. Therefore, a core challenge in current video enhancement tasks is how to collaboratively optimize inference efficiency, global consistency, and structural fidelity under limited computational power [8].

To address these constraints and challenges, this paper proposes HS-VidNet, an efficient video denoising network based on a frequency-spatial reconstruction mechanism. The architecture focuses on improving feature utilization through modular reconstruction. The Spatial and Channel Reconstruction Convolution (SCConv) is integrated symmetrically throughout the path to form the network backbone [9]. By explicitly reconstructing spatial and channel distributions in the feature flow, it suppresses non-discriminative redundancy while significantly reducing Floating Point Operations (FLOPs). A HiLo attention mechanism based on orthogonal frequency decoupling is introduced at the bottleneck layer [10]. This mechanism utilizes a dual-branch architecture to extract fine geometric textures and macro topological context in parallel, ensuring the fidelity of critical information such as lane edges. Experiments based on the CARLA-AWC adverse weather simulation dataset demonstrate that this method achieves an ideal balance between objective metrics (PSNR/SSIM) and inference speed, showing strong potential for practical application.

2. The HS-VidNet Network Model

This chapter first introduces the overall architecture based on U-Net and discusses the design considerations regarding real-time performance and fidelity for autonomous driving [11]. Subsequently, it provides a detailed analysis of how the SCConv module effectively suppresses and utilizes feature redundancy through spatial and channel reconstruction. The chapter also discusses the implementation logic of the HiLo attention mechanism, which balances global perception and local details using a frequency decoupling strategy. Finally, a multi-constraint hybrid loss function is defined to ensure the stability and convergence of model training.

2.1 Overall Architecture

As a video denoising method designed for autonomous driving, HS-VidNet adopts a lightweight symmetric U-shaped encoder-decoder architecture, as shown in Figure 1. This architecture balances real-time processing speed with feature restoration quality. Given a single-frame noisy input I_{in} (with dimensions $H\times W\times 3$), the network predicts a clean estimate I_{out} by fitting a non-linear mapping function F. The system architecture follows these functional divisions:

(1) Encoder Path: The encoder extracts deep semantic features through a cascade of four downsampling levels. The spatial resolution decreases progressively (from $H_1\times W_1$ to $H_4\times W_4$). Meanwhile, the channel dimension expands layer by layer (from $C_1$to C_4). To suppress computational redundancy throughout the path, SCConv modules are integrated at each level. An explicit reconstruction mechanism filters non-discriminative redundant features. This maintains critical geometric topology while compressing computational overhead.

(2) Bottleneck Layer Representation (Bottleneck): A HiLo attention mechanism with frequency decoupling is embedded in the low-level representation space ($H_5\times W_5\times C_5$). Based on an orthogonal frequency decoupling strategy, this module captures global context and fine local textures in parallel. This process compensates for the loss of high-frequency details caused by resolution reduction in deep mappings.

(3) Decoder Path: The structure is perfectly symmetric to the encoder. SCConv modules are integrated at each level to ensure continuous feature reconstruction efficiency. This path restores the

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

original resolution through step-by-step upsampling. The channel dimension is symmetrically reduced from C_4 to C_1. Multi-scale features from the encoding path are fused via skip connections. This guides the faithful restoration of critical edges, such as road signs and vehicle boundaries, ultimately generating a high-quality denoised image I_{out}.

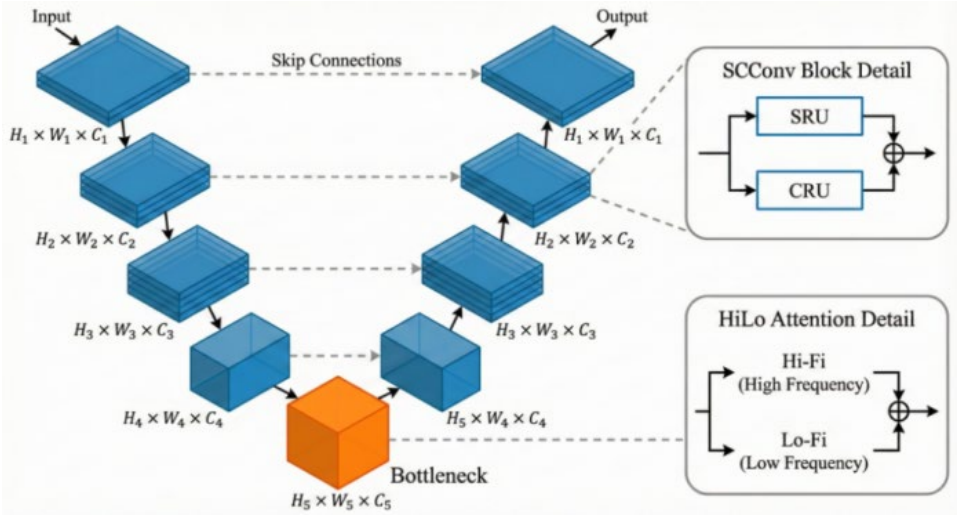


Figure 1: Overall architecture of HS-VidNet.

2.2 Spatial and Channel Reconstruction Module

To address on-board computational constraints and efficiency requirements, this paper introduces the Spatial and Channel Reconstruction Convolution (SCConv) module to replace standard convolutions throughout the encoder-decoder path. Standard convolutions apply a uniform sliding pattern even in low-information-entropy regions like the road and sky. This often results in significant spatial and channel redundancy, leading to a waste of computational resources. As shown in Figure 2, SCConv explicitly suppresses redundant information and reconstructs feature flows through parallel Spatial Reconstruction Units (SRU) and Channel Reconstruction Units (CRU). Its mathematical expression is:

$$F_{out}=SRU(F_{in}) \oplus CRU(F_{in}) \tag{1}$$

Where F_{in} and F_{out} represent the input and output feature maps, respectively, and " $\oplus$ " denotes element-wise addition.

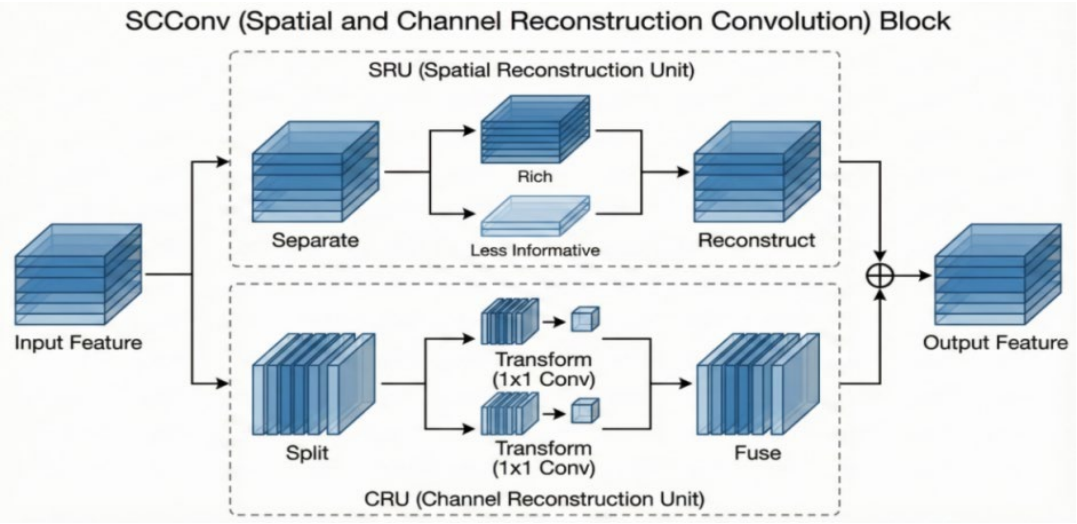


Figure 2: Internal structure of the SCConv module.

(1) Spatial Reconstruction Unit (SRU): The SRU utilizes a separation-reconstruction mechanism to suppress spatial redundancy. This unit quantifies feature weights using Group Normalization scaling factors. It explicitly separates low-information backgrounds from discriminative foregrounds. In

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

conjunction with weighted mapping, it guides the network to focus on high-frequency texture details in Regions of Interest (ROI). This maintains the topological integrity of the features.

(2) Channel Reconstruction Unit (CRU): The CRU follows the principles of splitting, transforming, and fusing to compress channel computational overhead. It dynamically splits channels and applies lightweight 1×1 convolutions to extract core representations. Combined with an adaptive fusion mechanism, it reconstructs cross-channel interactions. This design significantly reduces Floating Point Operations (FLOPs) while maintaining feature discriminative power.

(3) SCConv collaborates with these dual units to refine feature flows, balancing inference efficiency with representation robustness. Its inherent distillation logic filters non-discriminative interference. This guides the network to achieve faithful reconstruction of critical geometric details under complex conditions. This design principle ensures consistent computational gains across the entire path. It also demonstrates the scientific validity of maintaining information purity through representation reconstruction. Ultimately, this provides theoretical support for the robust application of HS-VidNet in autonomous driving scenarios.

2.3 HiLo Attention Mechanism

While SCConv reconstructs features, the HiLo attention mechanism is introduced to overcome the limitations of traditional convolutions' local receptive fields in modeling long-range topology. Conventional global attention can capture global context. However, its quadratic computational load impairs inference efficiency and often leads to over-smoothing of high-frequency edges, such as lane lines. Inspired by the human visual system, the HiLo mechanism employs orthogonal frequency decoupling. It splits multi-head attention into high-frequency (Hi-Fi) and low-frequency (Lo-Fi) branches for parallel processing (see Figure 3). This achieves collaborative encoding of macro-topology and micro-details with minimal overhead.

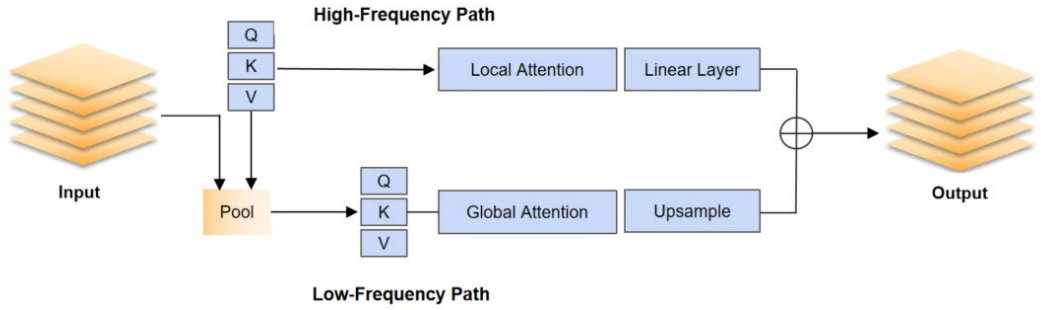


Figure 3: Schematic structure of the Frequency-Decoupled Attention (HiLo Attention) mechanism.

(1) High-Frequency Branch: To avoid feature smoothing risks induced by global aggregation, this branch adopts a local attention mechanism based on non-overlapping windows (as shown in Figure 3). The input feature X is spatially partitioned into multiple local grid windows of size s×s. Self-attention operations are performed independently within each window. This design forces the model to focus on local pixel correlations. While precisely restoring high-frequency topological structures, it optimizes computational complexity from $O((HW)^2)$ to $O(HW \times s^2)$. This ensures that the computational load grows linearly with resolution.

(2) Low-Frequency Branch: This branch utilizes a downsampling-global processing-upsampling bottleneck structure to capture global context dependencies, as illustrated in Figure 3. High-frequency interference is suppressed via the low-pass filtering property of average pooling to extract global contours. Subsequently, Query (Q), Key (K), and Value (V) vectors are generated in the compressed spatial dimension to perform global self-attention calculations. This effectively reduces the total operations. The processed global semantic features are then restored to the original resolution via bilinear interpolation to achieve spatial alignment with the high-frequency features.

(3) Feature Fusion: The outputs of the Hi-Fi and Lo-Fi branches are fused through element-wise addition. This strategy effectively integrates the complementary advantages of local textures and global semantics. It enables HS-VidNet to collaboratively encode macro road topology and micro edge details with low computational loss. This achieves a balance between denoising accuracy and inference speed.

(4) The HiLo attention mechanism adopts an orthogonal frequency decoupling strategy. It utilizes a

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

bottleneck structure to capture long-range dependencies from low-frequency redundancy, bypassing the computational bottleneck of global self-attention. This design achieves collaborative representation of macro-topology and micro-textures. While improving denoising quality, it maintains high inference speed, providing core algorithmic support for the engineering deployment of HS-VidNet in autonomous driving scenarios.

2.4 Multi-Constraint Hybrid Loss Function

To fully leverage the performance of HS-VidNet in frequency-decoupled feature extraction, this paper constructs a multi-dimensional perceptual hybrid optimization objective. This objective collaboratively achieves structural fidelity and detail restoration across three dimensions: the pixel, structural, and gradient domains. The total objective function, L_{total}, is formed by the weighted coupling of the pixel reconstruction loss L_{rec}, the structural similarity loss L_{ssim}, and the edge-aware loss L_{edge} :

$$L_{total}=\lambda_1 L_{rec}+\lambda_2 L_{ssim}+\lambda_3 L_{edge} \tag{2}$$

Here,$\lambda_1,\lambda_2,\lambda_3$are hyperparameters used to balance the weights of each term. By adjusting these weight coefficients, the contribution of each constraint dimension to model convergence can be coordinated.

1) Robust Pixel Reconstruction Loss: To improve the network's robustness against sensor thermal noise and outliers, this paper adopts the Charbonnier loss instead of the standard L2 loss (Mean Squared Error, MSE). Compared to MSE, which tends to produce smooth but blurry predictions, the Charbonnier loss combines the sparsity advantages of the L1 loss with numerical stability. It optimizes training by introducing a small constant ϵ (set to 10^{-3}):

$$L_{rec}=\sqrt{\|I_{out}-I_{gt}\|^2+\epsilon^2} \tag{3}$$

Where I_{out} represents the enhanced video frame and I_{gt} denotes the corresponding clean Ground Truth. This loss term provides basic convergence constraints for model training while ensuring color fidelity.

2) Structural Consistency Loss: To coordinate the modeling of the global topological structure by the Lo-Fi branch, a Structural Similarity Index Measure (SSIM) loss is introduced. This loss explicitly constrains the perceptual quality of the image in terms of luminance, contrast, and structure. Given the stringent requirements for road geometric continuity in autonomous driving scenarios, L_{ssim} effectively suppresses potential structural artifacts during denoising. This ensures that the spatial logic of lane lines and road surfaces remains undistorted:

$$L_{ssim}=1-SSIM(I_{out},I_{gt}) \tag{4}$$

3) Edge-Aware Auxiliary Loss: For the high-frequency details captured by the Hi-Fi branch, an edge loss based on the Laplacian operator is introduced. By calculating the difference between the predicted and ground truth images in the gradient domain, this loss forces the network to focus on the spatial alignment of high-frequency gradients:

$$L_{edge}=\sqrt{\|\Delta(I_{out})-\Delta(I_{gt})\|^2+\epsilon^2} \tag{5}$$

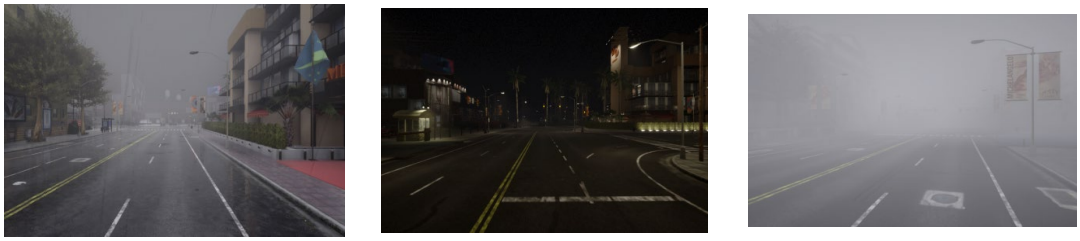
Where $\Delta(\cdot)$represents the Laplacian edge extraction operation. This loss term acts as an attention-guiding mechanism. By optimizing the spatial distribution of gradient weights, it explicitly strengthens the feature response in high-frequency texture regions. Consequently, it significantly enhances the sharpness and discriminative power of critical edges for downstream lane detection tasks.

3. Experimental results and analysis

3.1 Dataset and Implementation Details

This paper constructs a comprehensive dataset named CARLA-AWC (All-Weather Conditions) based on the high-fidelity open-source autonomous driving simulator CARLA (0.9.13). To cover typical environments such as urban roads, highways, and intersections, the data collection scope encompasses the Town01 to Town05 maps in CARLA[12]. Representative visual samples illustrating these diverse and complex environmental conditions are depicted in Figure 4.

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304



(a) (b) (c)

Figure 4: Sample images from the CARLA-AWC dataset. The dataset covers a variety of complex environments, including (a) rainy, (b) night, and (c) foggy scenarios

By configuring dynamically coupled environmental variables, this paper simulates the signal degradation process of sensors under various extreme working conditions. Specifically, weather conditions include clear scenarios as a baseline, along with rainfall of varying intensities (0% to 100%) and fog of varying densities (0% to 100%). This covers a diverse range from clear visibility to complete occlusion. Lighting conditions encompass high-intensity daylight, transitional twilight, and low-light night scenarios with solar altitude angles between -30° and 10°. The final dataset contains a total of 5,489 image pairs with a resolution of 640 x 480. Following a common split ratio, the dataset was randomly divided into a training set (4,391 images), a validation set (549 images), and a test set (549 images) at an 8:1:1 ratio.

All experiments in this study were conducted on a high-performance computing platform. The specific hardware environment includes an Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz, 128GB of RAM, and a single NVIDIA GeForce RTX 3090 GPU (24GB VRAM). The software environment is based on the Ubuntu 20.04 LTS operating system. We utilized the PyTorch 1.10 deep learning framework, accelerated by CUDA 11.3 and CuDNN 8.2. During the training phase, input images were randomly cropped into 128 x 128 patches, and horizontal flipping was applied to enhance data diversity. Network parameters were initialized using the He initialization method. We employed the AdamW optimizer with momentum parameters set to $\beta_1=0.9$, $\beta_2=0.999$, and a weight decay of 1×10^{-4}. The initial learning rate was set to $2\times10^{-4}$and gradually decayed to 1×10^{-6} over 150 epochs using a cosine annealing strategy. The batch size during training was 16.

3.2 Evaluation Metrics

To quantitatively evaluate the image restoration quality of HS-VidNet in complex autonomous driving scenarios, this paper adopts Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the core accuracy metrics. Furthermore, given the high sensitivity of autonomous driving tasks to real-time performance, this study further introduces Frames Per Second (FPS), Number of Parameters (Params), and Floating Point Operations (GFLOPs). These metrics are used to comprehensively evaluate the model's computational efficiency and deployment potential[13].

1) Peak Signal-to-Noise Ratio (PSNR): PSNR measures the pixel error between the denoised image and the original clean image in decibels (dB). A higher PSNR indicates lower signal distortion and superior denoising quality. Its calculation first requires the definition of Mean Squared Error (MSE), as shown in the following formulas:

$$MSE=\frac{1}{HW}\sum_{i,j}(I_{gt}(i,j)-I_{out}(i,j))^2 \tag{6}$$

$$PSNR=10\cdot\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{7}$$

Where I_{gt} denotes the clean ground-truth image and I_{out} represents the denoised image generated by the network. H and $W$are the height and width of the image, respectively. MAX_I is the maximum possible pixel value (typically 255), and MSE is the mean squared error.

2) Structural Similarity (SSIM):SSIM measures the perceptual quality of an image across three dimensions: luminance, contrast, and structure. The value of this metric ranges from [0, 1]. A value closer to 1 indicates higher structural fidelity. Its evaluation logic maintains strong consistency with the subjective perception of the Human Visual System (HVS). The calculation formula is as follows:

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)} \tag{8}$$

In this formula, μ_x and μ_y represent the mean values of images x and y, respectively. σ_x^2 and σ_y^2 denote the variances, while σ_{xy} represents the covariance . c_1 and c_2 are constants used to maintain divisional stability.

3.3 Comparative experiment

To comprehensively evaluate the restoration performance of HS-VidNet, this paper performs a quantitative comparison with four typical denoising algorithms on the CARLA-AWC test set. The selected methods cover various technical approaches. BM3D represents traditional transform-domain filtering algorithms based on non-local self-similarity. DnCNN is a pioneering work in deep learning-based denoising that employs a residual learning mechanism. SwinIR-Light is a lightweight model based on the Swin Transformer architecture, which excels at capturing global features. SCUNet is a current state-of-the-art (SOTA) model that combines the dual advantages of Convolutional Neural Networks and Transformers[14].

The results are presented in Table 1. The traditional BM3D method performs poorly when handling complex dynamic rain and fog noise. Although the deep learning method DnCNN has a small number of parameters (0.56M), its fully convolutional structure leads to high computational cost (185.7 GFLOPs) and an inference speed of only 45 FPS. The Transformer-based SwinIR-Light shows some competitiveness in SSIM (0.894). However, it is constrained by a high computational overhead of 278.3 GFLOPs, resulting in an inference speed of only 22 FPS. This fails to meet the real-time requirements of edge devices. In contrast, HS-VidNet (Ours) demonstrates highly competitive comprehensive performance. Thanks to the redundancy suppression of SCConv, our model achieves a high inference speed of 72 FPS with a computational load of only 87.2 GFLOPs. This allows it to be easily deployed on existing automotive hardware. Although the PSNR (31.45 dB) is slightly lower than that of the SOTA model SCUNet (31.52 dB), which has more parameters, our model achieves the best SSIM result (0.912). These results indicate that HS-VidNet achieves high-precision denoising while significantly improving inference speed. It reaches an ideal balance between restoration quality and computational efficiency. Additionally, the HiLo attention mechanism effectively preserves critical high-frequency information, such as lane lines and road geometric topology.

To evaluate the performance of the proposed algorithm, comparative experiments were conducted against several mainstream object detection algorithms, including YOLOv5s, YOLOv7-tiny, YOLOv8n, RT-detr, and NanoDet, as well as classic methods such as Faster R-CNN and SSD. For all algorithms used in the experiments, the input images were standardized to a resolution of 640 × 640 pixels. Each model was initialized with pretrained weights on public datasets and trained for 150 epochs. The evaluation metrics included the number of parameters, computational complexity (FLOPs), precision, recall, and mean average precision at IoU threshold 0.5 (mAP@0.5) to comprehensively assess model performance.

Table 1: Quantitative comparison of different methods on the CARLA-AWC dataset (Input size: 640 x 480)

Method	Params(M)	GFLOPs(G)	FPS	PSNR(dB)	SSIM
BM3D	-	-	-	25.64	0.706
DnCNN	0.56	185.7	45	28.18	0.819
SwinIR-Light	0.98	278.3	22	30.72	0.894
SCUNet	3.25	331.5	36	31.52	0.908
HS-VidNet	1.68	87.2	72	31.45	0.912

3.4 Ablation Study

To verify the effectiveness of the core components in HS-VidNet, this paper uses a lightweight U-Net as the baseline model. By progressively integrating the SCConv module and the HiLo attention mechanism, we quantitatively analyze the independent contribution of each component to the model's

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

performance.

The experimental results are presented in Table 2. While maintaining a relatively fast inference speed, the baseline model based on the lightweight U-Net shows limited performance in PSNR and SSIM when handling noise coupled with complex weather conditions. This is due to its restricted feature extraction mechanism. After introducing the Spatial and Channel Reconstruction Convolution (SCConv) module (Model A), the inference speed increased from 55 FPS to 82 FPS, and GFLOPs decreased significantly. This confirms that SCConv effectively breaks through the computational bottleneck by suppressing feature redundancy. Subsequently, with the introduction of the frequency-decoupled attention mechanism (Model B), the SSIM metric increased significantly from 0.852 to 0.898. This validates the effectiveness of the HiLo mechanism in capturing long-range dependencies and preserving high-frequency texture details, despite the drop in inference speed caused by additional computational overhead. Finally, the complete HS-VidNet combines both components. The channel compression characteristics of SCConv effectively alleviate the computational pressure of HiLo. The model achieves a reconstruction accuracy of 31.45 dB while maintaining a real-time inference performance of 72 FPS. This confirms the synergy and complementarity between the modules.

Table 2: Ablation study of core components (Input size: 640 x 480)

Model	SCConv	HiLoAttention	Params(M)	GFLOPs(G)	FPS	PSNR(dB)
Baseline	×	×	3.50	151.6	55	29.34
ModelA	√	×	1.55	69.9	82	29.18
ModelB	×	√	3.63	168.9	48	30.84
HS-VidNet	√	√	1.68	87.2	72	31.45
Model	SCConv	HiLoAttention	Params(M)	GFLOPs(G)	FPS	PSNR(dB)
Baseline	×	×	3.50	151.6	55	29.34
ModelA	√	×	1.55	69.9	82	29.18
ModelB	×	√	3.63	168.9	48	30.84
HS-VidNet	√	√	1.68	87.2	72	31.45
Model	SCConv	HiLoAttention	Params(M)	GFLOPs(G)	FPS	PSNR(dB)

3.5 Visual Analysis

Figure 5 illustrates a visual comparison of various algorithms on typical frames from the CARLA-AWC test set, which include rainy, foggy, and low-light scenarios. Figures 5(a)-(d) correspond to the ground truth, noisy input, SCUNet results, and the results from the proposed HS-VidNet, respectively. The Regions of Interest (ROI) indicated by red boxes are extracted using a pre-trained YOLOv8 model. By performing targeted sampling of core semantic targets such as traffic signs and lane lines, this strategy establishes a logical mapping between the restoration task and downstream perception.

(a) *(b)* *(c)* *(d)*

Figure 5: Visual comparison of various denoising algorithms on the CARLA-AWC test set: (a) ground truth; (b) noisy input; (c) SCUNet; (d) HS-VidNet (ours).

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

Experimental observations show that the noisy input in Fig. 5(b) suffers from severe visual occlusion and texture loss. The comparison method SCUNet [Fig. 5(c)] improves the overall Signal-to-Noise Ratio (SNR). However, it faces a significant trade-off bottleneck in preserving details. Observations in the red-boxed regions reveal that SCUNet over-suppresses some high-frequency signals while filtering strong noise. This leads to slight blurring of sign contours and building boundaries. These results reflect that large-parameter regression models struggle to balance noise suppression and edge sharpness without frequency feature guidance.

The proposed HS-VidNet achieves a balance between noise suppression and detail preservation. This success is attributed to the HiLo attention mechanism's ability to capture high-frequency components effectively. The model filters background noise while maintaining the clarity of critical road topologies. This includes features such as lane line continuity and sign contours. These qualitative results align with the quantitative metrics in Table 1. They demonstrate the model's potential for engineering applications in autonomous driving scenarios.

4. Conclusion

To meet the perception requirements of autonomous driving under extreme weather conditions, this paper proposes HS-VidNet, a lightweight video denoising network. By integrating Spatial and Channel Reconstruction Convolution (SCConv) and the HiLo attention mechanism, this architecture enhances critical topological fidelity during feature reconstruction while improving inference efficiency. Experiments on the CARLA-AWC dataset demonstrate that HS-VidNet achieves a real-time inference speed of 72 FPS on an RTX 3090 GPU, with a computational cost of only 87.2 GFLOPs. Its operational efficiency outperforms mainstream algorithms such as SCUNet and SwinIR-Light. Simultaneously, the model achieves a superior SSIM of 0.912. This confirms its fidelity in restoring critical topological information, such as lane lines and traffic signs. In summary, the proposed method achieves an effective balance between denoising performance and computational efficiency. Future research will focus on model deployment on embedded edge devices and the end-to-end joint optimization of the denoising module with downstream object detection tasks[15].

References

[1] Xu, C., & Sankar, R. (2024). A Comprehensive Review of Autonomous Driving Algorithms: Tackling Adverse Weather Conditions, Unpredictable Traffic Violations, Blind Spot Monitoring, and Emergency Maneuvers. Algorithms, 17(11), 526.

[2] Sheth, D. Y., Mohan, S., Vincent, J. L., et al. (2021). Unsupervised deep video denoising. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1759-1768.

[3] Tassano, M., Delon, J., & Veit, T. (2020). FastDVDnet: Towards Real-Time Deep Video Denoising via Recurrence and Multi-Depth-Separable Convolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1765-1774.

[4] Chen, L., Chu, X., Zhang, X., & Sun, J. (2022). Simple Baselines for Image Restoration. Proceedings of the European Conference on Computer Vision (ECCV), 17–33.

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-241.

[6] Zhao, H., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging, 3(1), 47–57.

[7] Liang, J., Cao, J., Sun, G., et al. (2021). SwinIR: Image restoration using swin transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 1833-1844.

[8] Han, K., Wang, Y., Xu, Q., et al. (2020). GhostNet: More features from cheap operations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[9] Li, J., Wen, Y., He, L., et al. (2023). SCConv: Spatial and channel reconstruction convolution for feature redundancy. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6153-6162.

[10] Pan, Z., Zhuang, B., Liu, J., et al. (2022). Fast vision transformers with HiLo attention. Advances in Neural Information Processing Systems (NeurIPS), 35, 14541-14554.

[11] Chan, K. C., Wang, X., Yu, K., & Loy, C. C. (2022). BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5934-5943.

ISSN 2616-5775 Vol. 9, Issue 3: 30-39, DOI: 10.25236/AJCIS.2026.090304

[12] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An Open Urban Driving Simulator. Proceedings of the 1st Annual Conference on Robot Learning (CoRL), 1–16.
[13] Bianco, S., Cadene, R., Celona, L., & Napoletano, P. (2018). Benchmark Analysis of Representative Deep Neural Network Architectures. IEEE Access, 6, 64270-64277. doi: 10.1109/ACCESS.2018.2877890.
[14] Zhang, K., Li, Y., Liang, J., et al. (2023). Practical Blind Image Denoising via Swin-Conv-UNet and Data Synthesis. Machine Intelligence Research, 20(6), 822–836.
[15] Yue, Z., Wang, J., & Loy, C. C. (2024). Efficient diffusion model for image restoration by residual shifting. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).