Low-Light Video Enhancement via Fast–Slow Dual Branches and Flow-Guided Attention

Tianzhi Jia , Shikui Wei * , Yao Zhao

Posted Date: 17 April 2026

doi: 10.20944/preprints202604.1276.v1

Keywords: low-light video enhancement; optical flow; Transformer; multi-rate architecture; temporal consistency

Article

Low-Light Video Enhancement via Fast–Slow Dual Branches and Flow-Guided Attention

Tianzhi Jia [1,2], Shikui Wei [1,2,*] and Yao Zhao [1,2]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2] Visual Intelligence + X International Joint Laboratory of the Ministry of Education, Beijing 100044, China
* Correspondence: shkwei@bjtu.edu.cn

Abstract

Low-light video enhancement aims to restore clear, color-faithful, and temporally consistent visual content from video sequences captured under extremely low signal-to-noise ratios and high dynamic range constraints. Existing multi-frame enhancement methods typically adopt uniform spatio-temporal sampling and feature extraction strategies for all frames, making it challenging to simultaneously achieve long-range temporal denoising and accurate fast-motion modeling. To address this trade-off, we propose a low-light video enhancement framework based on a Fast–Slow dual-branch architecture. The video signal is decomposed into two complementary feature streams: a Slow branch with sparse temporal sampling and high spatial resolution, built on a Vision Transformer backbone, which focuses on long-range temporal denoising and high-frequency texture restoration for static and slow-moving regions; and a Fast branch with dense temporal sampling and low spatial resolution, built on a ViT-Tiny backbone, which efficiently captures large-scale motion and rapid illumination changes. To mitigate the discrepancy in sampling rates and spatial resolutions between the two branches, we further introduce a flow branch based on a pre-trained StreamFlow model and design a Flow-Guided Cross-Attention (FGCA) module. FGCA first uses optical flow to geometrically modulate and progressively align Fast-branch features, and then injects the flow-enhanced Fast features into the Slow branch at each space-time location via lightweight pixel-wise cross-attention. This mechanism achieves a cascade of coarse geometric alignment and fine semantic fusion. Experiments on two real-world low-light video datasets, SDSD-indoor and SDSD-outdoor, demonstrate that our method consistently outperforms several representative approaches in terms of PSNR, SSIM, AB(Var), and MABD, while effectively suppressing motion blur and ghosting artifacts in dynamic night scenes, yielding temporally stable and perceptually pleasing results.

Keywords: low-light video enhancement; optical flow; Transformer; multi-rate architecture; temporal consistency

1. Introduction

With the rapid development of imaging hardware and mobile devices, video capture under low-light conditions has become a core requirement in many applications, such as smartphone photography, nighttime autonomous driving, and all-day surveillance. However, due to the physical limitations of photoelectric conversion, low-light videos often suffer from multiple degradations. On the one hand, to obtain sufficient exposure, the sensor gain (ISO) is usually significantly increased, leading to strong random noise dominated by a mixed Poisson–Gaussian distribution, with noise intensity fluctuating dramatically over time. On the other hand, the scarcity of photons causes severe color distortion and dynamic range compression. In addition, longer exposure times can increase signal strength but inevitably introduce motion blur in dynamic scenes [1]. These factors combined make low-light video enhancement (LLVE) a challenging yet practically important problem.

Compared with single-image low-light enhancement (LLIE), LLVE must improve image quality while maintaining temporal consistency. Directly applying single-frame enhancement methods (e.g.,

 doi:10.20944/preprints202604.1276.v1

RetinexFormer [2]) frame-by-frame typically ignores inter-frame dependencies, resulting in flickering brightness and temporally unstable noise patterns. Therefore, effectively exploiting multi-frame temporal information for joint denoising and structure reconstruction is the main direction of current research.

Many existing multi-frame LLVE methods (e.g., BasicVSR++ [3], SDSDNet [4]) adopt a *uniform* spatio-temporal modeling strategy: regardless of whether the content is a static background or a fast-moving object, the network uses the same sampling density and computational depth for every frame. This design has two fundamental limitations in complex dynamic scenes.

First, from the perspective of statistical signal processing, for additive noise, the quality of signal recovery is positively correlated with the number of observations. An ideal denoising network would use densely sampled, long temporal windows to fully exploit multi-frame averaging for noise suppression. However, in the presence of substantial motion, long temporal windows imply large physical displacements between the first and last frames. Under extremely low SNR, optical flow estimation or deformable convolution-based alignment becomes unreliable for such large disparities, leading to ghosting artifacts and motion blur in the fused results [5,6].

Second, video content is highly heterogeneous in the spatial and temporal frequency domains. Static backgrounds and slowly moving regions contain rich high-frequency textures and structural details, which require deep networks and sufficient temporal context for accurate restoration [7]. In contrast, fast-moving objects in low-light conditions are often already heavily motion-blurred and behave more like low-frequency structures with displacements, making them more suitable for modeling by shallow networks with larger temporal strides [8]. Existing methods treat all regions uniformly, using the same sampling and modeling complexity for the entire sequence. This leads to suboptimal allocation of computation: they fail to fully exploit multi-frame redundancy for denoising while also struggling to robustly capture large dynamics.

To address these issues, we propose a low-light video enhancement framework based on a *Fast–Slow* dual-branch architecture, inspired by multi-rate signal processing and geometric prior integration. The core idea is to decouple spatio-temporal features into complementary streams: a Slow branch with sparse temporal sampling and high spatial resolution, built on a ViT-Base backbone, which focuses on high-fidelity denoising and texture reconstruction for static and slow-motion regions; and a Fast branch with dense temporal sampling and low spatial resolution, built on a lightweight ViT-Tiny backbone, which focuses on modeling large motions and rapid illumination changes.

Since the Fast and Slow branches differ in both temporal sampling rate and spatial resolution, high-quality feature fusion under controllable alignment errors becomes critical. To this end, we further introduce a flow branch based on a pre-trained StreamFlow model [9] and design a Flow-Guided Cross-Attention (FGCA) module. FGCA leverages optical flow to provide explicit geometric priors for the Fast branch, progressively injecting encoded flow features into the Fast features at multiple stages. It then aligns Fast+Flow features to the Slow feature plane via temporal downsampling and spatial upsampling, finally fusing them with Slow features at each space-time location via lightweight pixel-wise cross-attention [10]. This achieves a cascade of coarse geometric alignment and fine semantic fusion with controlled computational complexity.

The main contributions of this work are summarized as follows:

- We propose a Fast–Slow dual-branch architecture for low-light video enhancement. By introducing asymmetric multi-rate sampling in time and space and employing ViT-Base and ViT-Tiny as the Slow and Fast backbones, respectively, we effectively decouple long-range high-fidelity denoising from efficient large-motion modeling, enabling targeted allocation of computational resources.
- We design a Flow-Guided Cross-Attention (FGCA) module that leverages pre-estimated optical flow as an explicit geometric prior to modulate Fast-branch features and performs pixel-wise cross-attention with the Slow branch after multi-stage alignment. FGCA effectively alleviates

 doi:10.20944/preprints202604.1276.v1

feature misalignment problems caused by multi-rate sampling and heterogeneous branches, significantly reducing motion artifacts.

- We conduct extensive experiments on two real-world low-light video datasets: SDSD-indoor and SDSD-outdoor. The results show that our method consistently outperforms several representative baselines in terms of PSNR, SSIM, AB(Var), and MABD, providing visually pleasing and temporally consistent results in challenging low-light dynamic scenes.

2. Related Work

2.1. Deep Learning-Based Low-Light Image Enhancement

Low-light image enhancement (LLIE) has evolved from early models based on Retinex theory to modern end-to-end deep learning approaches. RetinexNet [1] combines the classical Retinex decomposition model with convolutional neural networks, learning a reflectance–illumination decomposition for enhancement. Zero-DCE [11] and its variants learn illumination curves in an unsupervised manner, achieving decent performance even without paired training data. Recent advances in Retinex-based methods continue to improve decomposition quality and generalization [12,13].

With the success of Transformers in computer vision [14], attention-based LLIE methods are becoming mainstream. RetinexFormer [2] combines Retinex theory with a single-stage Transformer, using illumination-guided self-attention to model long-range dependencies. Diffusion-based models have also been explored for LLIE, such as GDP/Diff-LL [15], which introduces generative diffusion priors into unsupervised low-light enhancement [16,17]. MambaLLIE [18] further leverages state space models (SSMs) to achieve linear-complexity modeling of long sequences. While these methods significantly improve single-image low-light enhancement, they lack explicit temporal constraints, often resulting in flickering and temporal inconsistency when applied to videos [19].

2.2. Video Restoration and Temporal Alignment

In video restoration and enhancement, effectively using neighboring frames to provide complementary information for the current frame is crucial. Feature alignment is the key technique in this process. EDVR [5] and TDAN [6] use deformable convolutions (DCN) to perform adaptive sampling in feature space, achieving strong performance on various video restoration tasks [20,21]. However, under extremely low SNR, it is difficult for networks to learn accurate offsets from noisy features, which can lead to alignment failure and induce ghosting or distorted textures.

Attention-based alignment methods have gained traction alongside Transformers. FlowFormer [9] introduces Transformer architectures into optical flow estimation, significantly improving flow accuracy under moderate noise [22]. VRT [7] proposes a video restoration Transformer, modeling spatio-temporal features jointly via multi-scale attention [23]. In LLVE, recent methods such as StableLLVE [24] incorporate flow-based alignment to improve enhancement quality in dynamic scenes [25]. However, these methods typically adopt a single temporal sampling rate and do not explicitly exploit complementary information across different time scales.

2.3. Multi-Branch and Multi-Rate Network Architectures

Multi-branch and multi-rate architectures have been successfully used in high-level video understanding tasks. The SlowFast network [8] for action recognition uses a low-frame-rate Slow pathway to capture semantic context and a high-frame-rate Fast pathway to capture fast motion, achieving a good trade-off between accuracy and efficiency. This idea has inspired multi-rate modeling of temporal dynamics in video representation learning [26,27].

Unlike the aforementioned high-level tasks that focus on semantic classification, low-level LLVE requires accurate pixel-wise reconstruction, texture fidelity, and temporal consistency under extremely low SNR and complex noise conditions [28]. Consequently, our design must consider not only multi-rate temporal modeling but also strong geometric priors and robust feature alignment [29,30]. In this

 doi:10.20944/preprints202604.1276.v1

work, we systematically introduce the multi-rate concept into low-light video enhancement, coupled with flow-guided attention, to address the unique challenges of LLVE.

3. Materials and Methods

In this section, we present the proposed Fast–Slow dual-branch and flow-guided attention framework for low-light video enhancement. We first describe the overall architecture and the asymmetric sampling strategy, then detail the Slow branch based on ViT-Base, the Fast branch with embedded flow features, and the Flow-Guided Cross-Attention (FGCA) module. Finally, we introduce the loss functions used for end-to-end training.

3.1. Overall Architecture

Figure 1 illustrates the overall pipeline of our Fast–Slow dual-branch framework with flow-guided cross-attention. The model processes a local video window through asymmetric temporal sampling and multi-scale fusion across three stages, where optical flow explicitly guides the interaction between Fast and Slow branches.

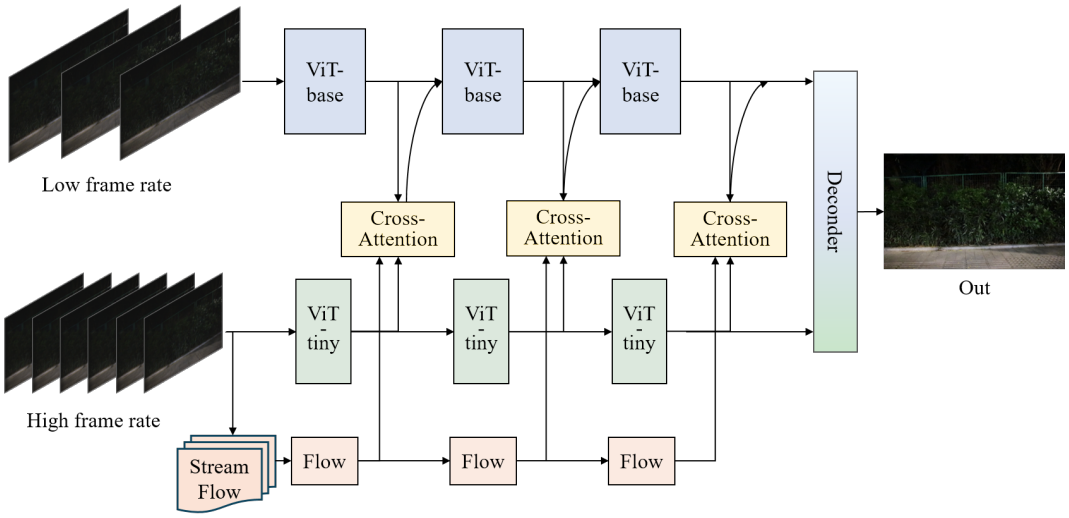

Figure 1. Overview of the proposed Fast–Slow dual-branch architecture with Flow-Guided Cross-Attention (FGCA). The Slow branch (high resolution, sparse frames) captures static details via ViT-Base, while the Fast branch (low resolution, dense frames) models motion dynamics enhanced by optical flow. Multi-stage FGCA modules fuse the two streams in a flow-aware manner to reconstruct the enhanced center frame.

Let $\{I_t\}_{t=1}^{T}$ be a low-light video sequence. For each time index t, we take a local temporal window around it:

$$\mathcal{I}_t = \{I_{t-K}, \ldots, I_t, \ldots, I_{t+K}\}, \tag{1}$$

where K is the half-window size. To achieve multi-rate feature decoupling, we design different temporal sampling and spatial resolution settings for the Fast and Slow branches.

Slow-branch input.

In the temporal dimension, we select a small number of key frames with larger intervals. Specifically, we choose the center frame and two frames with temporal offset $\Delta\tau$:

$$\mathcal{I}_t^{\text{slow}} = \{I_{t-\Delta\tau}, I_t, I_{t+\Delta\tau}\}, \quad \Delta\tau \leq K, \tag{2}$$

and maintain their spatial resolution close to the original. These three frames serve as high-fidelity input for static texture and long-range denoising.

 doi:10.20944/preprints202604.1276.v1

Fast-branch input.

In the temporal dimension, we densely sample all frames within the window:

$$\mathcal{I}_t^{\text{fast}} = \{I_{t-K}, \ldots, I_t, \ldots, I_{t+K}\}, \quad T_{\text{fast}} = 2K+1, \tag{3}$$

and downsample each frame spatially to a lower resolution. These downsampled frames are used to efficiently model large motions and local illumination changes.

Flow-branch input.

The flow branch shares the same temporal sampling set $\mathcal{I}_t^{\text{fast}}$ as the Fast branch. We use a pre-trained StreamFlow model to estimate flow from each neighboring frame to the center frame:

$$\mathbf{V}_{t+\tau\to t} = \text{StreamFlow}(I_{t+\tau}, I_t), \quad \tau \in \{-K, \ldots, K\}, \tag{4}$$

resulting in a flow sequence

$$\mathcal{V}_t = \{\mathbf{V}_{t+\tau\to t}\}_{\tau=-K}^{K}. \tag{5}$$

The overall network is divided into three stages (Stage 1/2/3). At stage l, the Slow, Fast, and flow branches produce multi-scale features:

$$\mathbf{S}^{(l)}, \quad \mathbf{F}^{(l)}, \quad \mathbf{O}^{(l)}. \tag{6}$$

Within each stage, we perform: (1) flow encoding and spatial resizing, (2) Fast+Flow fusion, (3) temporal downsampling and spatial upsampling of Fast features, and (4) FGCA-based pixel-wise fusion with Slow features. After three stages, the updated Slow features are decoded via a multi-scale decoder to reconstruct the enhanced center frame $\hat{Y}_t$.

3.2. Slow Branch: Global Modeling

The Slow branch aims to exploit sparse but long-range temporal context at high spatial resolution for high-fidelity denoising and texture restoration. We adopt a ViT-Base backbone for this branch.

3.2.1. Vision Transformer Overview

Vision Transformers (ViT) first partition an input image into fixed-size 2D patches. For a frame of resolution $H \times W$ and patch size $P \times P$, there are

$$N = \frac{H}{P} \cdot \frac{W}{P} \tag{7}$$

patches per frame. Each patch is flattened and projected into a C-dimensional feature space, and position embeddings are added to obtain a token sequence $\mathbf{E} \in \mathbb{R}^{N\times C}$. The basic attention operation is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{8}$$

where Q, K, V are query, key, and value matrices, and d_k is the key dimension. Stacking multiple Multi-Head Self-Attention (MHSA) and feed-forward (FFN) layers enables ViT to model long-range dependencies effectively.

3.2.2. Multi-Frame Slow-ViT Structure

For the input set $\mathcal{I}_t^{\text{slow}}$, we apply patch partition and linear projection to each of the three frames, obtaining a sequence with $3N$ tokens:

$$\mathbf{E}^{\text{slow}} \in \mathbb{R}^{(3N)\times C}. \tag{9}$$

 doi:10.20944/preprints202604.1276.v1

After passing through the ViT-Base backbone, we obtain multi-scale features at three stages:

$$\mathbf{S}^{(l)} \in \mathbb{R}^{B\times T_{\text{slow}}\times C_s^{(l)}\times H_l\times W_l}, \quad l=1,2,3, \tag{10}$$

where $T_{\text{slow}}=3$, $H_{l+1}<H_l$, $W_{l+1}<W_l$, and $C_s^{(l+1)}>C_s^{(l)}$. To control complexity, we use local or window-based attention in each stage, computing self-attention only within local spatial windows.

In the reconstruction stage, we primarily use the feature channel corresponding to the center time index for decoding, while the other two time channels provide auxiliary denoising and alignment information via FGCA.

3.3. Fast Motion Branch and Flow Branch

3.3.1. Fast Branch with ViT-Tiny

For each frame in $\mathcal{I}_t^{\text{fast}}$, we downsample it to a lower resolution (h,w) using bilinear interpolation or a small CNN, then apply patch partition and projection:

$$\mathbf{E}^{\text{fast}} \in \mathbb{R}^{B\times(T_{\text{fast}}N')\times C_f}, \quad N'=\frac{h}{P}\cdot\frac{w}{P}. \tag{11}$$

A ViT-Tiny backbone with fewer layers and smaller hidden dimensions is used, producing multi-scale Fast features:

$$\mathbf{F}^{(l)} \in \mathbb{R}^{B\times T_{\text{fast}}\times C_f^{(l)}\times h_l\times w_l}, \quad l=1,2,3. \tag{12}$$

Thanks to the low spatial resolution and small model size, the Fast branch can operate at high temporal resolution with manageable computational cost, focusing on capturing motion and illumination dynamics.

3.3.2. Optical Flow and Flow Encoding

Optical flow describes the apparent motion of pixels between frames. For frames $I_{t+\tau}$ and I_t, the flow vector $\mathbf{V}_{t+\tau\to t}(x,y)=(u,v)$ indicates that pixel (x,y) in $I_{t+\tau}$ corresponds to $(x+u,y+v)$ in I_t. In low-light scenarios, classical brightness constancy assumptions are often violated by noise and illumination changes; thus, learning-based optical flow models are preferred.

We use a pre-trained StreamFlow model as a frozen flow estimator, producing a flow sequence:

$$\mathbf{V} \in \mathbb{R}^{B\times T_{\text{fast}}\times 2\times H_{\text{flow}}\times W_{\text{flow}}}. \tag{13}$$

To embed flow into the Fast branch, we design a lightweight flow encoder FlowEnc composed of two 3×3 convolutional layers, mapping the 2-channel flow to higher-dimensional geometric features:

$$\mathbf{O}^{(0)}=\text{FlowEnc}(\mathbf{V}) \in \mathbb{R}^{B\times T_{\text{fast}}\times C_{\text{flow}}\times H_{\text{flow}}\times W_{\text{flow}}}. \tag{14}$$

3.3.3. Multi-Stage Fusion of Fast and Flow

At stage l, we resize $\mathbf{O}^{(0)}$ to match the spatial resolution of $\mathbf{F}^{(l)}$:

$$\mathbf{O}^{(l)}=\text{Resize}(\mathbf{O}^{(0)},h_l,w_l) \in \mathbb{R}^{B\times T_{\text{fast}}\times C_{\text{flow}}\times h_l\times w_l}, \tag{15}$$

and map channels via a 1×1 convolution:

$$\tilde{\mathbf{O}}^{(l)}=\phi^{(l)}(\mathbf{O}^{(l)}), \quad \phi^{(l)}:\mathbb{R}^{C_{\text{flow}}}\to\mathbb{R}^{C_f^{(l)}}. \tag{16}$$

We then inject flow into Fast features by element-wise addition:

$$\tilde{\mathbf{F}}^{(l)}=\mathbf{F}^{(l)}+\tilde{\mathbf{O}}^{(l)}. \tag{17}$$

 doi:10.20944/preprints202604.1276.v1

This treats flow as a geometric modulation signal for the Fast branch, enabling the features to be aware of pixel displacements while keeping the implementation simple and efficient.

3.4. Flow-Guided Cross-Attention (FGCA)

3.4.1. Multi-Rate Feature Alignment

After multi-stage fusion with flow, we obtain flow-enhanced Fast features $\tilde{\mathbf{F}}^{(l)}$ at each stage. The Flow-Guided Cross-Attention (FGCA) module further (i) aligns these features temporally and spatially with the Slow branch, and (ii) performs genuine cross-attention with queries from the Slow branch and keys/values from the concatenation of Fast and flow features. In this way, the attention distribution is explicitly guided by optical flow.

Figure 2 illustrates the detailed mechanism of our pixel-wise flow-guided cross-attention module, highlighting how queries from the Slow branch interact with keys and values derived from the concatenation of Fast and flow features to achieve effective feature fusion under the guidance of optical flow.

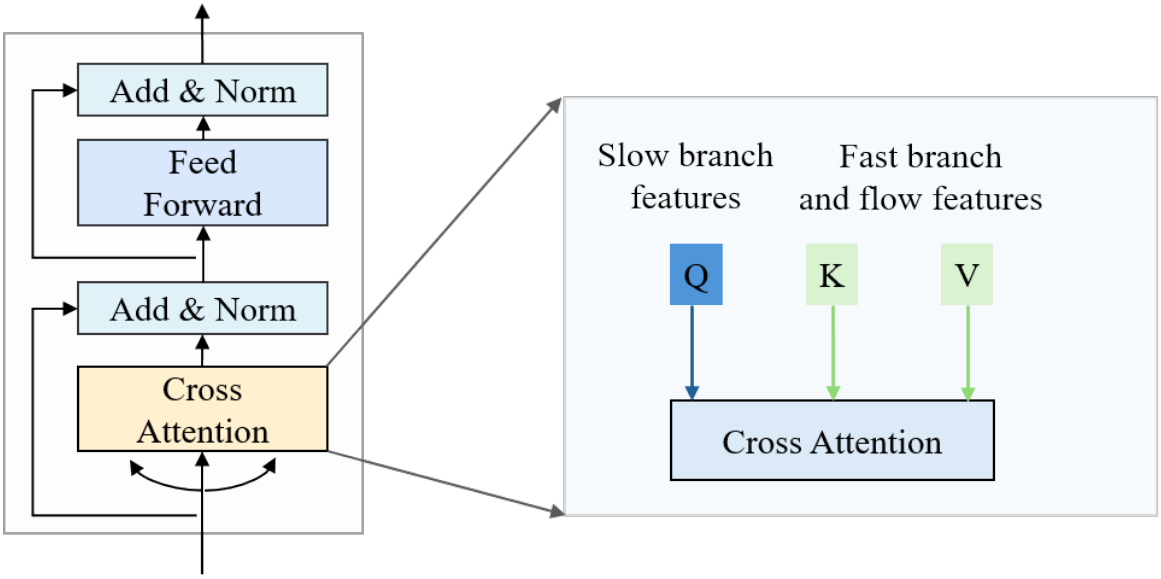

Figure 2. Illustration of the pixel-wise flow-guided cross-attention (FGCA) mechanism. The Slow branch provides queries that focus on high-fidelity static/slow-motion content, whereas keys and values are constructed from concatenated Fast branch and flow features. This setup allows for flow-directed attention, enhancing the interaction and fusion between the Slow and Fast branches.

To match the temporal length $T_{\text{slow}} = 3$ of the Slow branch, we apply temporal convolution with stride on $\tilde{\mathbf{F}}^{(l)}$:

$$\tilde{\mathbf{F}}^{(l)} \xrightarrow{\text{TimeConv}^{(l)}} \bar{\mathbf{F}}^{(l)} \in \mathbb{R}^{B\times T_{\text{slow}}\times C_f^{(l)}\times h_l\times w_l}, \tag{18}$$

where $\text{TimeConv}^{(l)}$ aggregates motion and illumination information from local temporal windows into T_{slow} output steps.

We then upsample $\bar{\mathbf{F}}^{(l)}$ to the spatial resolution of $\mathbf{S}^{(l)}$:

$$\hat{\mathbf{F}}^{(l)} = \text{Upsample}(\bar{\mathbf{F}}^{(l)}, H_l, W_l) \in \mathbb{R}^{B\times T_{\text{slow}}\times C_f^{(l)}\times H_l\times W_l}. \tag{19}$$

Similarly, we temporally aggregate and spatially upsample the encoded flow features $\mathbf{O}^{(0)}$. For stage l, we obtain

$$\hat{\mathbf{O}}^{(l)} = \text{Upsample}(\text{TimeConv}^{(l)}(\mathbf{O}^{(0)}), H_l, W_l) \in \mathbb{R}^{B\times T_{\text{slow}}\times C_{\text{flow}}^{(l)}\times H_l\times W_l}, \tag{20}$$

where $C_{\text{flow}}^{(l)}$ is the flow feature dimension at stage l after temporal aggregation and channel projection.

 doi:10.20944/preprints202604.1276.v1

At each stage l and space-time location (τ, x, y), we therefore have

$$\mathbf{s}_{\tau,x,y}^{(l)} \in \mathbb{R}^{C_s^{(l)}}, \quad \mathbf{f}_{\tau,x,y}^{(l)} \in \mathbb{R}^{C_f^{(l)}}, \quad \mathbf{o}_{\tau,x,y}^{(l)} \in \mathbb{R}^{C_{\text{flow}}^{(l)}}, \tag{21}$$

corresponding to Slow, Fast, and flow features, respectively.

3.4.2. Pixel-Wise Flow-Guided Cross-Attention

FGCA is designed as a pixel-wise cross-attention module in which:

- **Queries** Q come from the Slow branch, encoding high-fidelity static/slow-motion content.
- **Keys** and **values** (K, V) come from the concatenation of Fast and flow features, so that attention is explicitly guided by optical flow.

For each (τ, x, y) at stage l, we first concatenate Fast and flow features:

$$\mathbf{g}_{\tau,x,y}^{(l)} = \left[\, \mathbf{f}_{\tau,x,y}^{(l)} \,\|\, \mathbf{o}_{\tau,x,y}^{(l)} \,\right] \in \mathbb{R}^{C_f^{(l)} + C_{\text{flow}}^{(l)}}. \tag{22}$$

We then project the Slow features to queries and the concatenated Fast+Flow features to keys and values:

$$\mathbf{q}_{\tau,x,y}^{(l)} = W_q^{(l)}\, \mathbf{s}_{\tau,x,y}^{(l)}, \quad \mathbf{k}_{\tau,x,y}^{(l)} = W_k^{(l)}\, \mathbf{g}_{\tau,x,y}^{(l)}, \quad \mathbf{v}_{\tau,x,y}^{(l)} = W_v^{(l)}\, \mathbf{g}_{\tau,x,y}^{(l)}, \tag{23}$$

where $W_q^{(l)} \in \mathbb{R}^{d_l \times C_s^{(l)}}$, $W_k^{(l)}, W_v^{(l)} \in \mathbb{R}^{d_l \times (C_f^{(l)} + C_{\text{flow}}^{(l)})}$, and d_l is the attention embedding dimension at stage l.

In practice, we adopt multi-head cross-attention (MHCA); for clarity, we describe a single-head version. The attention coefficient at (τ, x, y) is computed as

$$\alpha_{\tau,x,y}^{(l)} = \sigma\left(\frac{\langle \mathbf{q}_{\tau,x,y}^{(l)}, \mathbf{k}_{\tau,x,y}^{(l)} \rangle}{\sqrt{d_l}} \right), \tag{24}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product and $\sigma(\cdot)$ is the sigmoid function. This can be viewed as *flow-guided cross-attention*: the query is purely from the Slow branch, while the key/value depends on both Fast and flow features.

We then fuse the attended Fast+Flow representation into the Slow branch:

$$\hat{\mathbf{s}}_{\tau,x,y}^{(l)} = \mathbf{s}_{\tau,x,y}^{(l)} + \alpha_{\tau,x,y}^{(l)} \cdot W_o^{(l)} \mathbf{v}_{\tau,x,y}^{(l)}, \tag{25}$$

where $W_o^{(l)} \in \mathbb{R}^{C_s^{(l)} \times d_l}$ is an output projection. In the multi-head setting, all heads are concatenated before applying $W_o^{(l)}$.

By vectorizing (τ, x, y) along the batch dimension, we can implement FGCA efficiently with overall complexity $O(BT_{\text{slow}} H_l W_l)$. Applying FGCA at three stages yields updated multi-scale Slow features:

$$\hat{\mathbf{S}}^{(l)} = \text{FGCA}^{(l)}\big(\mathbf{S}^{(l)}, \hat{\mathbf{F}}^{(l)}, \hat{\mathbf{O}}^{(l)}\big), \quad l = 1, 2, 3. \tag{26}$$

Finally, the center time-index features are fed into a decoder with multi-scale upsampling and skip connections to reconstruct the enhanced center frame $\hat{Y}_t$.

3.5. Loss Functions

To jointly optimize pixel-level fidelity, perceptual quality, and temporal stability, we use a combination of loss terms. Let $\hat{Y}_t$ denote the enhanced center frame and Y_t the corresponding ground truth.

 doi:10.20944/preprints202604.1276.v1

3.5.1. Reconstruction Loss

We adopt the Charbonnier loss as the base reconstruction loss due to its robustness to outliers:

$$\mathcal{L}_{\text{rec}} = \sqrt{\|\hat{Y}_t - Y_t\|_2^2 + \epsilon^2}, \tag{27}$$

where ϵ is a small constant (e.g., 10^{-3}).

3.5.2. Perceptual Loss

To improve visual quality and texture realism, we use a VGG-based perceptual loss:

$$\mathcal{L}_{\text{per}} = \sum_m \lambda_m \left\| \phi_m(\hat{Y}_t) - \phi_m(Y_t) \right\|_1, \tag{28}$$

where $\phi_m(\cdot)$ denotes the m-th layer of a pre-trained VGG network and λ_m are layer weights.

3.5.3. Overall Loss

The overall training objective is the weighted sum of all terms:

$$\mathcal{L} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{per}} \mathcal{L}_{\text{per}}. \tag{29}$$

where $\lambda_{\text{rec}}, \lambda_{\text{per}}$ are trade-off weights chosen by validation.

4. Results

4.1. Datasets and Implementation Details

4.1.1. Datasets

We evaluate our method on two publicly available real-world low-light video datasets:

- **SDSD-indoor** and **SDSD-outdoor** [31]: These datasets are captured by a high-end camera under diverse low-light conditions, providing paired low-light and normal-light videos. The indoor subset primarily contains indoor scenes with medium motion, while the outdoor subset features nighttime outdoor scenes with larger motions.

4.1.2. Implementation Details

Unless otherwise specified, all models are trained and evaluated under the same optimizer settings, data augmentation strategy, and training protocol on each dataset. The temporal window length is set to $2K+1$ (e.g., $K=3$); the Slow-branch temporal interval $\Delta\tau$ is chosen close to K to cover a large physical time span. We use the Adam optimizer for end-to-end training with an initial learning rate of 4×10^{-4} and cosine annealing. The batch size is set to 4. Standard data augmentation, including random cropping and horizontal flipping, is applied during training. We use PSNR and SSIM to evaluate restoration quality, and additionally report Average Brightness Variance (AB(Var)) and Mean Absolute Brightness Difference (MABD) to measure temporal consistency.

4.2. Comparison with State-of-the-Art Methods

We compare our approach with several representative single-frame and multi-frame baselines, including the single-frame methods SNR [32] and RetinexFormer [2], as well as the multi-frame methods SDSDNet [4], DP3DF [33], StableLLVE [24], and the recently proposed LLVE_STCD. To evaluate temporal consistency, we additionally report Average Brightness Variance (AB(Var)) and Mean Absolute Brightness Difference (MABD), where lower values indicate better temporal stability.

For fair comparison, we follow the official implementations and recommended settings of the compared methods whenever available, and evaluate all methods on the same data splits. Table 1 reports the quantitative results.

 doi:10.20944/preprints202604.1276.v1

Table 1. Quantitative comparison on SDSD-indoor and SDSD-outdoor. We report PSNR, SSIM, AB(Var) and MABD. For AB(Var) and MABD, lower values (↓) indicate better temporal stability. Best results are in bold.

Method	SDSD-indoor				SDSD-outdoor			
	PSNR	SSIM	AB(Var)↓	MABD↓	PSNR	SSIM	AB(Var)↓	MABD↓
SNR	27.10	0.83	0.072	1.481	23.05	0.80	0.088	1.625
RetinexFormer	26.45	0.79	0.065	1.253	22.68	0.77	0.074	1.419
SDSDNet	26.92	0.76	0.015	0.287	23.41	0.72	0.021	0.354
DP3DF	27.54	0.77	0.009	0.194	24.03	0.74	0.012	0.228
StableLLVE	25.63	0.72	0.011	0.210	22.31	0.69	0.018	0.245
LLVE_STCD	28.93	0.88	0.006	0.145	26.32	0.82	0.009	0.176
Ours	**29.54**	**0.91**	**0.003**	**0.092**	**27.15**	**0.86**	**0.004**	**0.118**

As shown in Table 1, our method achieves the highest PSNR and SSIM on both datasets. Furthermore, our method achieves significantly lower scores in AB(Var) and MABD compared to single-frame methods and other video baselines, indicating that our Fast–Slow dual-branch design with FGCA effectively suppresses temporal flickering and maintains high brightness consistency across frames.

4.3. Qualitative Comparison

To visually demonstrate the effectiveness of our method, we present qualitative comparisons on the SDSD-indoor and SDSD-outdoor datasets in Figure 3 and Figure 4, respectively. Comparative methods include the single-image enhancement method RetinexFormer [2] and video-based methods DP3DF [33] and StableLLVE [24].

As shown in Figure 3 (SDSD-indoor), the single-frame method RetinexFormer successfully enhances visibility but suffers from severe noise amplification in dark regions. Since it processes frames independently, it cannot leverage temporal information for denoising, resulting in grainy textures. The video-based baseline DP3DF utilizes multi-frame information to suppress noise; however, it tends to over-smooth fine details, leading to a loss of structural fidelity in complex indoor scenes. StableLLVE achieves better noise removal but introduces artifacts in regions with intricate textures. In contrast, our method effectively balances noise suppression and detail preservation, recovering clear textures and sharp edges that are closest to the ground truth (GT).

Figure 3. Visual comparison on the **SDSD-indoor** dataset. From left to right: Input, RetinexFormer [2], DP3DF [33], StableLLVE [24], Ours, and Ground Truth. Our method (Ours) recovers cleaner textures and effectively suppresses noise compared to other state-of-the-art methods.

Figure 4 presents the results on the SDSD-outdoor dataset, which contains challenging dynamic scenes with nighttime traffic and pedestrians. In these scenarios, maintaining temporal consistency and avoiding motion blur are critical. RetinexFormer, lacking temporal modeling, produces sharp but noisy individual frames. Among the video enhancement methods, DP3DF and StableLLVE struggle with fast-moving objects, often producing visible ghosting artifacts or blurring moving silhouettes due to inaccurate alignment. Thanks to our proposed Fast–Slow dual-branch architecture and Flow-Guided

 doi:10.20944/preprints202604.1276.v1

Cross-Attention (FGCA), our method robustly handles large motions, delivering sharp, ghosting-free results with naturally suppressed background noise.

Figure 4. Visual comparison on the **SDSD-outdoor** dataset. The red boxes indicate zoomed-in patches. Compared with RetinexFormer [2], DP3DF [33], and StableLLVE [24], our method achieves superior performance in restoring details of moving objects and suppressing motion artifacts.

4.4. Temporal Consistency and Error Distribution Analysis

To further evaluate the temporal stability and color restoration accuracy of our method, we visualized the pixel-wise RGB error distribution across consecutive video frames using 3D density plots. In these visualizations, the x-axis represents the pixel error value (difference between the enhanced frame and ground truth), the y-axis denotes the frame sequence and RGB channels, and the z-axis represents the density of the error distribution. A narrower, taller peak centered at zero indicates higher restoration fidelity, while the consistency of these peaks along the time axis reflects temporal stability.

Figure 5 presents the comparison on the **SDSD-indoor** dataset. The baseline method, DP3DF [33], exhibits relatively wide and short distributions for all three RGB channels, implying a higher variance in reconstruction errors and residual noise. Furthermore, the shape of the error distributions for DP3DF fluctuates between frames, suggesting temporal flickering. In contrast, our method consistently produces sharp, leptokurtic distributions tightly centered around zero. This indicates that our Fast–Slow dual-branch architecture effectively suppresses noise and restores accurate colors without introducing inter-frame jitter.

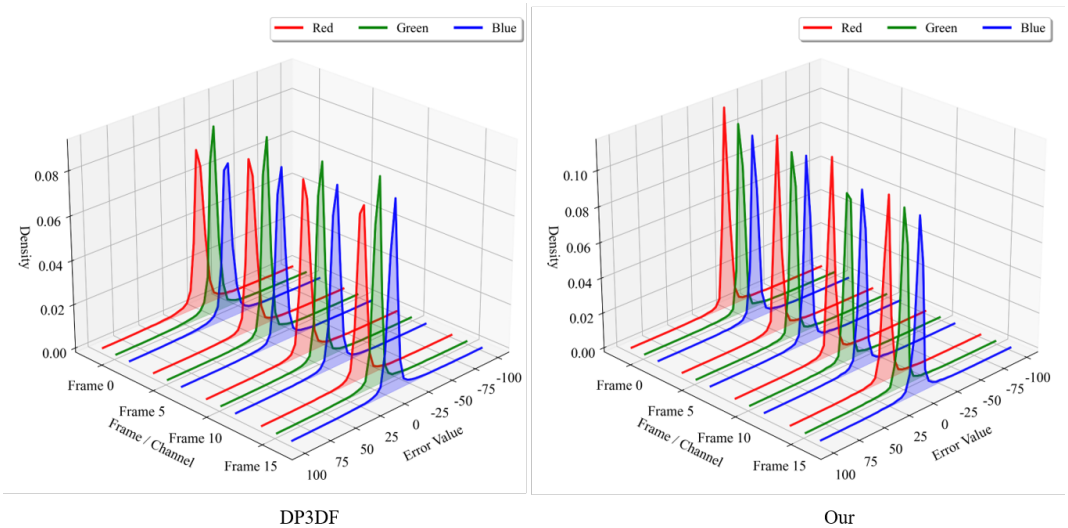

Figure 5. Temporal RGB error distribution on **SDSD-indoor**. Compared to DP3DF [33], our method shows sharper, narrower peaks centered at zero across all frames, indicating superior color fidelity and temporal stability.

 doi:10.20944/preprints202604.1276.v1

The advantage of our method is even more pronounced in the **SDSD-outdoor** dataset (Figure 6), which involves dynamic lighting and fast motion. Under these challenging conditions, DP3DF struggles to maintain consistent features, resulting in spread-out error distributions that vary significantly over time. Conversely, thanks to the explicit geometric guidance from the Flow-Guided Cross-Attention (FGCA) module, our method maintains a stable and concentrated error profile across the entire sequence. The alignment of RGB peaks across frames demonstrates that our method achieves superior temporal coherence and robustness against motion, avoiding the color shifting and ghosting artifacts common in multi-frame processing.

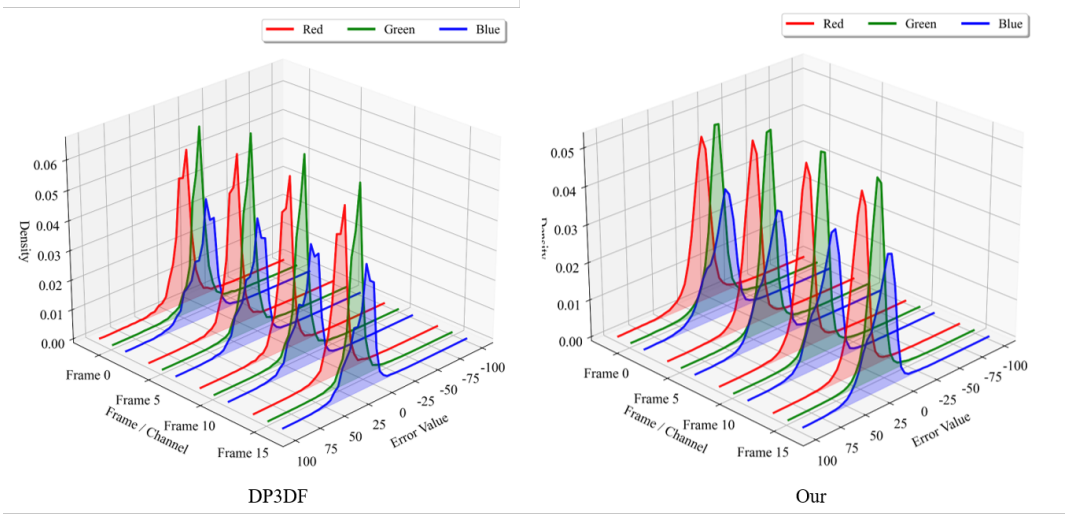

Figure 6. Temporal RGB error distribution on **SDSD-outdoor**. Despite complex motions, our method maintains a consistent and concentrated error distribution over time, whereas DP3DF exhibits wider variance and fluctuations.

4.5. Ablation Study

We conduct ablation studies on SDSD-indoor and SDSD-outdoor to analyze the contributions of individual components. The following variants are considered:

- **w/o Flow**: Remove the flow branch; no flow encoder or Fast+Flow fusion is used.
- **w/o FGCA**: Replace FGCA with simple channel concatenation followed by a 1×1 convolution; Slow and Fast features are fused only by concatenation after alignment.
- **w/o Slow-ViT**: Replace the ViT-Base backbone in the Slow branch with a CNN of comparable depth.
- **w/o Fast branch**: Remove the Fast branch entirely, using only the Slow branch (i.e., a multi-frame ViT-based enhancement).
- **Full**: The full model with Fast–Slow dual branches, flow branch, and FGCA.

From Table 2, removing the flow branch (w/o Flow) or FGCA (w/o FGCA) leads to clear performance drops, highlighting the importance of explicit geometric priors and pixel-wise cross-attention for dynamic scenes. Replacing ViT-Base with a CNN (w/o Slow-ViT) significantly degrades performance, confirming that global self-attention in the Slow branch is particularly beneficial under extremely low SNR. Removing the Fast branch (w/o Fast branch) has a stronger impact on the SDSD-outdoor dataset, which contains larger motions, validating the necessity of a high-frame-rate Fast branch for motion modeling.

 doi:10.20944/preprints202604.1276.v1

Table 2. Ablation study on SDSD-indoor and SDSD-outdoor (PSNR / SSIM).

Variant	SDSD-indoor	SDSD-outdoor
w/o Flow	27.56 / 0.84	24.79 / 0.79
w/o FGCA	27.81 / 0.85	25.01 / 0.80
w/o Slow-ViT	26.98 / 0.82	24.06 / 0.77
w/o Fast branch	27.34 / 0.83	24.52 / 0.78
Full (ours)	**29.54 / 0.91**	**27.15 / 0.86**

5. Discussion

We have presented a Fast–Slow dual-branch framework with flow-guided attention for low-light video enhancement. The experimental results show that:

First, the asymmetric temporal and spatial sampling design in the Fast and Slow branches enables the network to balance long-range temporal denoising and large-motion modeling without excessive computational overhead. The Slow branch, through sparse key frames and a ViT-Base backbone, provides high-fidelity reconstruction for static and slow-motion regions; the Fast branch, with dense temporal sampling and a lightweight ViT-Tiny, effectively captures fast motions and rapid illumination changes.

Second, the flow branch and FGCA play a critical role in cross-branch feature alignment and fusion. By encoding StreamFlow optical flow into geometric features and injecting them into the Fast branch at multiple stages, then performing pixel-wise cross-attention with the Slow branch, FGCA achieves fine spatio-temporal alignment with controlled complexity, effectively suppressing motion artifacts and ghosting.

Nevertheless, there is room for improvement. For ultra-high-resolution videos, ViT-Base can still be computationally expensive; future work may explore more efficient sparse attention or low-rank approximations. Additionally, the current flow is estimated by an external pre-trained model; jointly optimizing flow estimation and enhancement within a single end-to-end framework is an interesting direction for further research.

6. Conclusions

In this paper, we addressed the challenge of jointly performing long-range temporal denoising and fast-motion modeling in low-light video enhancement by proposing a Fast–Slow dual-branch architecture with flow-guided attention. By introducing asymmetric multi-rate sampling strategies in both time and space, and leveraging ViT-Base and ViT-Tiny as the Slow and Fast backbones, respectively, we effectively decoupled static high-frequency texture restoration from dynamic large-displacement modeling and achieved targeted utilization of computational resources. Furthermore, by incorporating pre-estimated optical flow and designing a Flow-Guided Cross-Attention (FGCA) module, we combined physical geometric priors with Transformer-based semantic selection, enabling accurate pixel-level alignment and fusion across branches and mitigating misalignment artifacts caused by multi-rate sampling.

Extensive experiments on SDSD-indoor and SDSD-outdoor demonstrate that our method outperforms several representative baselines in PSNR, SSIM, AB(Var), and MABD, and produces visually pleasing, temporally stable results in challenging low-light dynamic scenes. In future work, we plan to explore more efficient attention mechanisms and adaptive multi-rate scheduling strategies to further reduce computational cost, as well as to extend the proposed framework to other video restoration tasks, such as deraining and desnowing.

Author Contributions: Conceptualization, T.J. and S.W.; methodology, T.J.; software, T.J.; validation, T.J.; formal analysis, T.J.; writing—original draft preparation, T.J.; writing—review and editing, S.W. and Y.Z.; supervision, S.W. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

 doi:10.20944/preprints202604.1276.v1

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. SDSD-indoor and SDSD-outdoor can be accessed from the corresponding public sources cited in the manuscript. The source code and additional evaluation results will be made publicly available upon acceptance of the manuscript.

Acknowledgments: The authors would like to thank Dr. Huaxin Pang for his valuable discussions and early-stage contributions to this work, as well as the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLVE	Low-Light Video Enhancement
LLIE	Low-Light Image Enhancement
ViT	Vision Transformer
FGCA	Flow-Guided Cross-Attention
SNR	Signal-to-Noise Ratio
AB(Var)	Average Brightness Variance
MABD	Mean Absolute Brightness Difference

References

1. Wei, C.; Wang, W.; Yang, W.; Liu, J. Deep Retinex Decomposition for Low-Light Enhancement. In *Proceedings of the British Machine Vision Conference (BMVC)*; 2018. [arXiv]
2. Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; Zhang, Y. RetinexFormer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2023; pp. 12470–12479. [CrossRef]
3. Chan, K.C.K.; Zhou, S.; Xu, X.; Loy, C.C. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022; pp. 5962–5971. [CrossRef]
4. Chen, C.; Chen, Q.; Do, M.N.; Koltun, V. Seeing Motion in the Dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2019; pp. 3184–3193. [CrossRef]
5. Wang, X.; Chan, K.C.K.; Yu, K.; Dong, C.; Loy, C.C. EDVR: Video Restoration with Enhanced Deformable Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; 2019; pp. 1954–1963. [CrossRef]
6. Tian, Y.; Zhang, Y.; Fu, Y.; Xu, C. TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020. [CrossRef]
7. Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; Van Gool, L. VRT: A Video Restoration Transformer. *arXiv* 2022, arXiv:2201.12288. [arXiv]
8. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2019. [CrossRef]
9. Huang, Z.; Shi, X.; Zhang, C.; Wang, Q.; Li, H.; Yang, M.-H. FlowFormer: A Transformer Architecture for Optical Flow. In *Computer Vision–ECCV 2022*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; pp. 668–685. [CrossRef]
10. Zhang, K.; Peng, J.; Fu, J.; Liu, D. Exploiting Optical Flow Guidance for Transformer-Based Video Inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**. [CrossRef]
11. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020; pp. 1777–1786. [CrossRef]
12. Zhang, Y.; Jiang, J.; Wang, Z.; Zhang, Q.; Jiang, Y.; Liu, J.; Hou, Z. Low-Light Image Enhancement Method Based on Retinex Theory and Dual-Tree Complex Wavelet Transform. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 83. [CrossRef]

 doi:10.20944/preprints202604.1276.v1

13. Mou, E.; Wang, H.; Chen, X.; Li, Z.; Cao, E.; Chen, Y.; Wang, Y.; Sun, W. Retinex Theory-Based Nonlinear Luminance Enhancement and Denoising for Low-Light Endoscopic Images. *BMC Med. Imaging* **2024**, *24*, 207. [CrossRef]
14. Hassija, V.; Palanisamy, B.; Chatterjee, A.; Mandal, A. Transformers for Vision: A Survey on Innovative Methods for Computer Vision. *IEEE Access* **2025**, *13*. [CrossRef]
15. Fei, B.; Lyu, W.; Pan, L.; Zhang, J.; Yang, W.; Luo, T.; Zhang, B.; Dai, B. Generative Diffusion Prior for Unified Image Restoration and Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023; pp. 9935–9946. [CrossRef]
16. Li, X.; Ren, Y.; Jin, X.; Lan, C.; Wang, X.; Zeng, W.; Wang, X.; Chen, Z. Diffusion Models for Image Restoration and Enhancement: A Comprehensive Survey. *Int. J. Comput. Vis.* **2025**, *133*, 8078–8108. [CrossRef]
17. Jiang, H.; Luo, A.; Liu, X.; Han, S.; Liu, S. LightenDiffusion: Unsupervised Low-Light Image Enhancement with Latent-Retinex Diffusion Models. In *Computer Vision–ECCV 2024*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2024; pp. 161–179. [CrossRef]
18. Weng, J.; Yan, Z.; Tai, Y.; Qian, J.; Yang, J.; Li, J. MambaLLIE: Implicit Retinex-Aware Low Light Enhancement with Global-then-Local State Space. In *Advances in Neural Information Processing Systems (NeurIPS)*; 2024. [CrossRef]
19. Liu, F.; Fan, L. A Review of Advancements in Low-Light Image Enhancement Using Deep Learning. *Neurocomputing* **2025**, *652*, 131052. [CrossRef]
20. Deng, J.; Dong, S.; Chen, L.; Hu, J.; Zhuo, C. STDF: Spatio-Temporal Deformable Fusion for Video Quality Enhancement on Embedded Platforms. *ACM Trans. Embed. Comput. Syst.* **2024**, *23*, Article 2. [CrossRef]
21. Wang, H.; Chen, Z.; Chen, C.W. Learned Video Compression via Heterogeneous Deformable Compensation Network. *IEEE Trans. Multimed.* **2024**, *26*, 1855–1866. [CrossRef]
22. Lu, Y.; Han, C.; Wang, Q.; Fan, H.; Kong, Z.; Tan, P. Optical Flow as Spatial-Temporal Attention Learners. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11491–11506. [CrossRef]
23. Chen, N.; Xu, T.; Sun, M.; Yao, C.; Yang, D. Understanding Video Transformers: A Review on Key Strategies for Feature Learning and Performance Optimization. *Intell. Comput.* **2025**, Article 0143. [CrossRef]
24. Zhang, F.; Li, Y.; You, S.; Fu, Y. Learning Temporal Consistency for Low Light Video Enhancement from Single Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2021; pp. 4967–4976. [CrossRef]
25. Chen, K.; Liang, G.; Lu, Y.; Li, H.; Wang, L. EvLight++: Low-Light Video Enhancement With an Event Camera: A Large-Scale Real-World Dataset, Novel Method, and More. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 1608–1625. [CrossRef]
26. Zhu, J.; Zhang, X.; Tang, L.; Jiang, J.H. MSNeRV: Neural Video Representation with Multi-Scale Feature Fusion. *arXiv* 2025, arXiv:2506.15276. [arXiv]
27. Sun, W.; Cao, L.; Guo, Y.; Du, K. Multimodal and Multiscale Feature Fusion for Weakly Supervised Video Anomaly Detection. *Sci. Rep.* **2024**, *14*, 22835. [CrossRef]
28. Yakovenko, A.; Chakvetadze, G.; Khrapov, I.; Zhelezov, M.; Vatolin, D.; Timofte, R.; Oh, Y.; Kwon, J.; Park, J.; Cho, N.I.; Xu, S.; Jiang, R.; Peng, L.; Fu, X.; Zha, Z.-J.; Peng, X.; Feng, H.; Tie, Z.; Xia, Z.; Wang, L. AIM 2025 Low-Light RAW Video Denoising Challenge: Dataset, Methods and Results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; 2025. [CrossRef]
29. Yan, W.; Sun, Y.; Yue, G.; Zhou, W.; Liu, H. FVIFormer: Flow-Guided Global-Local Aggregation Transformer Network for Video Inpainting. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2024**, *14*, 235–244. [CrossRef]
30. Merugu, R.; Suhail, M.S.; Sarashetti, A.P.; Reddem, V.B.R.; Bajpai, P.K.; Unde, A.S. JFFRA: Joint Flow and Feature Refinement Using Attention for Video Restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; 2025; pp. 5589–5599. [CrossRef]
31. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2018; pp. 3291–3300. [CrossRef]
32. Xu, X.; Wang, R.; Fu, C.-W.; Jia, J. SNR-Aware Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022; pp. 17693–17703. [CrossRef]
33. Xu, X.; Wang, R.; Fu, C.-W.; Jia, J. Deep Parametric 3D Filters for Joint Video Denoising and Illumination Enhancement in Video Super Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence* **2023**, *37*, 3054–3062. [CrossRef]

 doi:10.20944/preprints202604.1276.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.