Low-Light Video Enhancement via Spatial-Temporal Consistent Decomposition

Xiaogang Xu[1,2], **Kun Zhou**[3], **Tao Hu**[4], **Jiafei Wu**[5*], **Ruixing Wang**[6*], **Hao Peng**[7], **Bei Yu**[1]
[1]The Chinese University of Hong Kong
[2]Zhejiang University
[3]The Chinese University of Hong Kong (Shenzhen)
[4]PICO, Bytedance
[5]The University of Hong Kong
[6]DJI Technology Co., Ltd.
[7]Zhejiang Normal University
xiaogangxu00@gmail.com, kunzhou@link.cuhk.edu.cn, yihouxiang@gmail.com,
jcjiafeiwu@gmail.com, ruixing0406@gmail.com, hpeng@zjnu.edu.cn, byu@cse.cuhk.edu.hk

Abstract

Low-Light Video Enhancement (LLVE) seeks to restore dynamic or static scenes plagued by severe invisibility and noise. In this paper, we present an innovative video decomposition strategy that incorporates view-independent and view-dependent components to enhance the performance of LLVE. We leverage dynamic cross-frame correspondences for the view-independent term (which primarily captures intrinsic appearance) and impose a scene-level continuity constraint on the view-dependent term (which mainly describes the shading condition) to achieve consistent and satisfactory decomposition results. To further ensure consistent decomposition, we introduce a dual-structure enhancement network featuring a cross-frame interaction mechanism. By supervising different frames simultaneously, this network encourages them to exhibit matching decomposition features. This mechanism can seamlessly integrate with encoder-decoder single-frame networks, incurring minimal additional parameter costs. Extensive experiments are conducted on widely recognized LLVE benchmarks, covering diverse scenarios. Our framework consistently outperforms existing methods, establishing a new SOTA performance.

1 Introduction

Low-light enhancement aims to enhance underexposed images and videos captured in low-light conditions [Xu *et al.*, 2022; Wang *et al.*, 2021], improving their visual quality while reducing noise. This technique can be applied in wide-ranging applications, such as portrait photography on mobile devices [Ignatov *et al.*, 2017; Hasinoff *et al.*, 2016], nighttime face recognition [Ma *et al.*, 2022; Wang *et al.*, 2022] and vehicle detection [Fu *et al.*, 2023a; Wu *et al.*, 2022]. The key challenge when enhancing videos is the need for consistent enhancement results across corresponding locations in different frames,

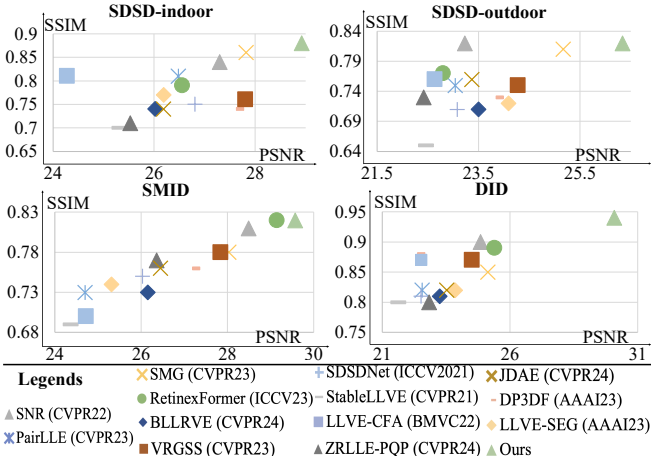

Figure 1: Our proposed LLVE method *consistently* achieves SOTA performance on *different LLVE datasets involving various scenes* with the same network architecture.

which vary both spatially and temporally. Moreover, there is a scarcity of real-world, high-quality, spatially aligned video pairs for dynamic scenes, resulting in limited generalization capabilities for unseen videos with varying conditions.

One solution is to incorporate parameterized enhancement models with physical significance into the enhancement process, reducing the reliance on training data. According to the linear Lambertian model for intrinsic image decomposition [Yuille, 2012; Narihira *et al.*, 2015], an observed image $\boldsymbol{I}$ can be expressed as the element-wise product ($\otimes$) of its albedo component and shading component. The albedo part is normally view-independent, capturing the intrinsic appearance, whereas the shading component is commonly view-dependent, influenced by lighting conditions and the surface normal's direction. Inspired by the Lambertian model, we propose that a frame in the LLVE task can similarly be decomposed as $\boldsymbol{I} = \boldsymbol{L} \otimes \boldsymbol{R}$, where $\boldsymbol{L}$ represents the view-dependent term, and $\boldsymbol{R}$ corresponds to the view-independent component.

In this paper, we introduce an innovative approach to attain

*Corresponding authors.

our decomposition for normal-light video outputs, ensuring consistency in both spatial and temporal dimensions. We leverage cross-frame correspondences and real-world physical continuity constraints to achieve this decomposition without the need for explicit supervision. Furthermore, to enhance the consistency of the decomposition, we've developed an efficient dual-structure enhancement network featuring a novel cross-frame interaction mechanism.

Our approach aims to predict both the $\boldsymbol{R}$ and $\boldsymbol{L}$ of normal-light images from the provided low-light inputs. It is important to note that the decomposition of view-independent and view-dependent terms is not unique, and our goal is to implement a suitable decomposition that enhances the performance of LLVE. In this paper, we aim to design appropriate priors directly from the video data to implement the decomposition. We recognize that $\boldsymbol{R}$ should represent the physical properties of objects, which should remain consistent across various observation perspectives. To achieve this, we first calculate corresponding locations in videos and utilize these computed spatial-temporal correspondences, along with uncertainty values, to link the predictions of $\boldsymbol{R}$ across different frames. Conversely, the terms of $\boldsymbol{L}$ are subject to a spatial smoothness loss to model real-world physical constraints (since the lighting field is continuous in practice). This approach allows us to establish a temporal-spatial consistency constraint in the enhanced videos, without requiring additional supervision. *Note that we cannot guarantee the decomposition results are strictly view-independent or view-dependent, as this is not a 3D method. Our goal is to apply appropriate constraints to approximate desired decomposition properties, while focusing on optimizing the final LLVE performance with our approach.*

Additionally, we introduce a dual-network structure to enforce the consistency of feature representations across frames, facilitating the consistent synthesis of decomposition in turn. Two frames are used as inputs, and their features are propagated within the designed Cross-Frame Interaction Module (CFIM). CFIM differs from other similar architectures in the restoration task [Lv *et al.*, 2023] in terms of the fusion manner. CFIM facilitates interaction through a combination of long-range cross-frame attention computation and short-range channel-spatial fusion, ensuring the accurate and comprehensive merging of features from both global and local perspectives. Moreover, CFIM selectively incorporates knowledge from randomly chosen neighboring frames into the backbone (i.e., the part without feature propagation) during training. In this way, CFIM help the backbone in learning how to adaptively utilize knowledge from various cross-frame features, resulting in a more robust backbone.

We performed experiments on various widely recognized video enhancement datasets. Our results, both quantitative and qualitative, consistently demonstrate the effectiveness and state-of-the-art (SOTA) performance of our framework, as shown in Fig. 1. Furthermore, we conducted a large-scale user study involving 100 participants, which showcased the superiority of our results in terms of human subjective perceptions. In summary, our contributions are three-fold.

- We introduce a novel canonical form for LLVE that predicts spatial-temporal consistent decomposition with view-independent and view-dependent terms for normal-light outputs. This decomposition strategy leverages spatial-temporal correspondences and continuity.
- We design a new LLVE network that facilitates interaction among the features of different frames and fits our consistent decomposition strategy.
- We conduct extensive experiments on public datasets, illustrating the effectiveness of our proposed framework.

2 Related Work

Low-Light Image Enhancement. To enhance the quality of a low-light video, image enhancement methods can be applied on a frame-by-frame basis. In recent years, learning-based Low-Light Image Enhancement (LLIE) techniques [Jiang *et al.*, 2021; Yang *et al.*, 2021] have made significant advancements, with a primary focus on supervised approaches.

Low-Light Video Enhancement. In addition to the need for LLIE, there is a growing demand for video enhancement, considering the widespread use of videos as a popular data format on the internet and in photographic equipment. Various approaches have been proposed in this context [Chen *et al.*, 2019; Triantafyllidou *et al.*, 2020; Ye *et al.*, 2023; Lv *et al.*, 2023; Xu *et al.*, 2023a; Fu *et al.*, 2023a]. Liu et.al. [Liu *et al.*, 2023] and Liang et.al. [Liang *et al.*, 2023] used prior event information to learn enhancement mapping for brightening videos. Xu et.al. [Xu *et al.*, 2023a] designed a parametric 3D filter tailored for enhancing and sharpening low-light videos. Recently, Fu et.al. [Fu *et al.*, 2023a] introduced a video enhancement method called LAN, which iteratively refines illumination and adaptively adjusts it. However, it's important to note that LAN lacks an explicit constraint for maintaining consistent reflectance and illumination decomposition.

Several datasets for video enhancement, encompassing both static [Chen *et al.*, 2019; Jiang and Zheng, 2019; Wang *et al.*, 2019; Triantafyllidou *et al.*, 2020] and dynamic motions [Wang *et al.*, 2021; Fu *et al.*, 2023a], have been introduced. In this paper, we introduce a novel LLVE method, designed to enhance effects on these datasets. Our decomposition method explicitly and consistently models view-dependent and -independent for all frames.

3 Method

3.1 Decomposition Model

Motivation. According to the linear Lambertian model [Narihira *et al.*, 2015], an observed image $\boldsymbol{I}$ can be formulated as the element-wise product of its albedo and shading component. Here, the albedo part is view-independent, capturing the intrinsic appearance, while the shading component is view-dependent. Inspired by this, when presented with an image $\boldsymbol{I}$ (in this paper, we denote the low-light image as $\boldsymbol{I}_d$ and the normal-light image as $\boldsymbol{I}_n$, with d and n serving as subscript abbreviations), we assume its decomposition into view-dependent part $\boldsymbol{L}$ and view-independent part $\boldsymbol{R}$, as

$$\boldsymbol{I} = \boldsymbol{L} \otimes \boldsymbol{R}, \tag{1}$$

where $\otimes$ denotes the element-wise multiplication, $\boldsymbol{L}$ and $\boldsymbol{R}$ can be formulated for different channels, i.e., the channel number is 3 for the sRGB domain. In this context, $\boldsymbol{L}$ mainly

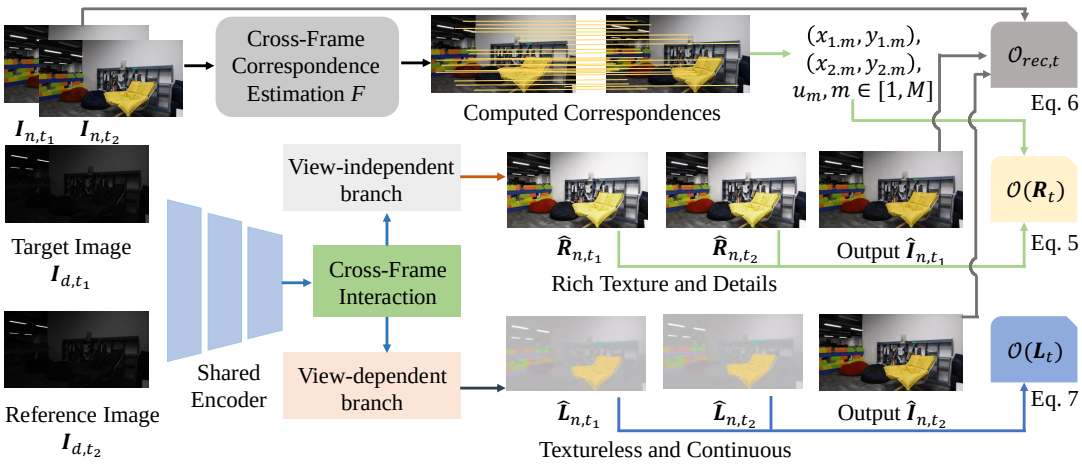

Figure 2: Our framework offers a comprehensive solution that explicitly and consistently models the view-independent and view-dependent decomposition of enhanced normal-light outputs across different frames. To achieve this, we enforce consistent features in the view-independent terms across different frames by leveraging computed correspondences in the temporal dimension of videos ($\mathcal{O}(\boldsymbol{R}_t)$). Simultaneously, we ensure that the view-dependent terms exhibit a spatially continuous distribution ($\mathcal{O}(\boldsymbol{L}_t)$), aligning with real-world scenarios. Furthermore, our network incorporates cross-frame interaction and simultaneous supervision of different frames within a video, encouraging consistent features for these frames derived from one video. For a more detailed visual representation, please refer to Fig. 3.

describes the light intensity of objects, which is expected to exhibit piece-wise continuity. On the other hand, $\boldsymbol{R}$ primarily represents the physical properties of the objects, encompassing textures and details observed in the data $\boldsymbol{I}$. If we obtain $\boldsymbol{L}$ and $\boldsymbol{R}$ in the normal-light conditions for a given input low-light image, the target normal-light image can be acquired accordingly. The common objective can be summarized as

$$\mathcal{O} = \|\boldsymbol{I} - \boldsymbol{L} \otimes \boldsymbol{R}\| + \mathcal{O}(\boldsymbol{L}) + \mathcal{O}(\boldsymbol{R}), \tag{2}$$

where $\mathcal{O}(\boldsymbol{L})$ and $\mathcal{O}(\boldsymbol{R})$ denote the constraints for $\boldsymbol{L}$ and $\boldsymbol{R}$ (we will introduce our designed new constraints in Sec. 3.2). When applying this model to sequential video data, the objective related to the decomposition of video can be written as

$$\begin{aligned} \mathcal{O}_v &= \mathbb{E}_{t=1:T}[\mathcal{O}_{rec,t} + \mathcal{O}(\boldsymbol{L}_t) + \mathcal{O}(\boldsymbol{R}_t)], \\ \mathcal{O}_{rec,t} &= \|\boldsymbol{I}_t - \boldsymbol{L}_t \otimes \boldsymbol{R}_t\|, \end{aligned} \tag{3}$$

where $\mathbb{E}$ is the average operation, T is the number of frame, and $\boldsymbol{L}_t/\boldsymbol{R}_t$ is the decomposition output at the time index of t. Compared with Equation (2), we need to simultaneously guarantee the decomposition quality and the temporal consistency.

Video Data Guide the Decomposition by Themselves. We have discovered that video data itself can play a crucial role in facilitating the achievement of our decomposition predictions for normal-light outputs, thus obviating the need for external priors typically used in low-light enhancement. By establishing correspondences among different frames, we can impose specific constraints: the $\boldsymbol{R}_t$ of each frame should faithfully represent the intrinsic texture of objects within the target scene, remaining consistent regardless of changes in viewpoint. Likewise, the $\boldsymbol{L}_t$ of each frame should exhibit the desired continuity consistent with real-world physical properties. For further elaboration, please refer to Section 3.2.

Problem Formulation. We adopt the supervised setting. Given a clip of low-light data $\boldsymbol{I}_{d,t}, t \in [1, T]$, there is a paired normal-light data $\boldsymbol{I}_{n,t}, t \in [1, T]$. *We aim to directly obtain the decomposition of $\boldsymbol{I}_{n,t}$ from $\boldsymbol{I}_{d,t}$ using network f*, as

$$\hat{\boldsymbol{L}}_{n,t}, \hat{\boldsymbol{R}}_{n,t} = f(\boldsymbol{I}_{d,t}), \hat{\boldsymbol{I}}_{n,t} = \hat{\boldsymbol{L}}_{n,t} \otimes \hat{\boldsymbol{R}}_{n,t}, \tag{4}$$

where $\hat{\boldsymbol{L}}_{n,t}$ and $\hat{\boldsymbol{R}}_{n,t}$ are the estimated targets, and $\hat{\boldsymbol{I}}_{n,t}$ is the predicted enhancement result. Our framework is shown in Fig. 2 that will be introduced in the following sections.

3.2 Spatial-Temporal Consistent Decomposition

Motivation. Video can be conceptualized as static/dynamic 3D multi-view data [Wang *et al.*, 2023b; Wang *et al.*, 2023a]. Prior research has demonstrated that the multi-view data itself can be harnessed to decompose the view-dependent and view-independent elements, which respectively encapsulate the intrinsic texture and view-altering illuminations [Wang *et al.*, 2023a]. Thus, once we establish correspondence relationships among $\boldsymbol{I}_{d,t}$ for all t, we can apply the view-independent constraint to derive $\hat{\boldsymbol{R}}_{n,t}$, as it represents the intrinsic properties of the target scene. Subsequently, obtaining $\hat{\boldsymbol{L}}_{n,t}$ becomes achievable through utilizing the inherent reconstruction loss, in conjunction with adherence to the continuity assumption.

Implementation of Correspondences. Obtaining correspondences for $\boldsymbol{I}_{d,t}$, $t \in [1, T]$, can be a challenging task due to the presence of various degradations, including visibility issues and noise. Fortunately, we have access to corresponding normal-light data, which significantly aids in establishing these correspondences, as illustrated in Fig. 2, where $\boldsymbol{I}_{d,t}$ and $\boldsymbol{I}_{n,t}$ are pixel-wise aligned.

Given two frames $\boldsymbol{I}_{n,t_1}$ and $\boldsymbol{I}_{n,t_2}$, correspondences can be determined using the prediction network F [Edstedt *et al.*, 2023]. We assume the detection of M correspondences, denoted as $c_m = (x_{1,m}, y_{1,m}, x_{2,m}, y_{2,m}), m \in [1, M]$. Each correspondence, represented by c_m, consists of four coordinates: $(x_{1,m}, y_{1,m})$ represents the pixel coordinate in $\boldsymbol{I}_{n,t_1}$,

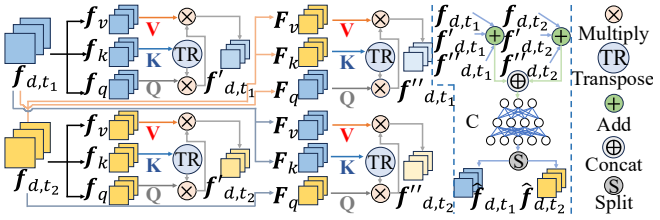

Figure 3: The lightweight cross-frame interaction mechanism (CFIM) propagates different frames' features, along with cross-frame attention and spatial-channel fusion. Cross-frame interaction can be employed in the deep feature space of arbitrary single-image encoder-decoder frameworks, and we choose U-Net here.

while $(x_{2,m}, y_{2,m})$ signifies the coordinate in $\boldsymbol{I}_{n,t_2}$. Furthermore, each correspondence is associated with an uncertainty value u_m, determined through a data-driven approach.

Constraints Formulation. To satisfy the constraint that the corresponding pixels in $\boldsymbol{R}_t$ are matched, we have $\mathcal{O}(\boldsymbol{R}_t)$ as

$$\mathcal{O}(\boldsymbol{R}_{t_1,t_2}) = \mathop{\mathbb{E}}_{m\in[1,M]} u_m \|\hat{\boldsymbol{R}}_{n,t_1}[x_{1,m}, y_{1,m}] - \hat{\boldsymbol{R}}_{n,t_2}[x_{2,m}, y_{2,m}]\|. \tag{5}$$

The reconstruction item in Equation (3) can be denoted as

$$\mathcal{O}_{rec,t} = \|\boldsymbol{I}_{n,t} - \hat{\boldsymbol{L}}_{n,t} \otimes \hat{\boldsymbol{R}}_{n,t}\|. \tag{6}$$

Although the view-dependent part can be obtained via the reconstruction constraint, We further incorporate a continuity constraint to align with real-world properties. For a pixel in the video frame, denoted as p, we introduce a spatial continuity objective on the predicted $\hat{\boldsymbol{L}}_{n,t}$ as the following loss function

$$\mathcal{O}(\boldsymbol{L}_t) = \mathbb{E}_t[v_t^p \times [\partial_x \hat{\boldsymbol{L}}_{n,t}(p)]^2 + u_t^p \times [\partial_y \hat{\boldsymbol{L}}_{n,t}(p)]^2], \tag{7}$$

where ∂_x and ∂_y are partial derivatives in horizontal and vertical directions, respectively. v_t^p and u_t^q are spatially-varying smoothness weights, calculated as

$$v_t^p = (\|\partial_x \boldsymbol{U}_t(p)\| + \Delta)^{-1}, u_t^p = (\|\partial_y \boldsymbol{U}_t(p)\| + \Delta)^{-1}, \tag{8}$$

where $\boldsymbol{U}_t$ represents the logarithmic transformation of $\boldsymbol{I}_{d,t}$, and Δ is a small constant (set to 0.0001) used to avoid division by zero. The diagram is shown in Fig. 2.

3.3 Dual Network for Consistent Decomposition

Overview and Motivation. To enhance the consistency of decomposition, it's crucial for various frames within a video to mutually share features. By incorporating shared features and simultaneous supervision, the features used for synthesizing decomposition parameters across different frames can exhibit greater consistency. When the video is enhanced frame by frame without feature propagation, the decomposition results tend to be suboptimal due to the inherently challenging nature of maintaining consistency. This observation is empirically validated in our ablation study. Current LLVE methods primarily leverage multiple-frame inputs to facilitate propagation [Wang *et al.*, 2021; Fu *et al.*, 2023a]. However, these strategies come with a notable cost in terms of aligning features.

In this paper, we introduce an efficient strategy for propagation using a dual approach. This approach entails loading two frames, specifically target frame $\boldsymbol{I}_{d,t_1}$ and reference frame $\boldsymbol{I}_{d,t_2}$ (where $\boldsymbol{I}_{d,t_2}$ can be randomly selected from the temporal neighbors of $\boldsymbol{I}_{d,t_1}$ during training, and is set as the closest frame of $\boldsymbol{I}_{d,t_1}$ during inference), propagating features at the deepest layer of the network through our cross-frame interaction (as depicted in Fig. 3), and concurrently supervising these dual outputs to guarantee consistency.

Feature Propagation via Spatial-varying Fusion. Suppose the feature of $\boldsymbol{I}_{d,t_1}$ and $\boldsymbol{I}_{d,t_2}$ are denoted as $\boldsymbol{f}_{d,t_1}$ and $\boldsymbol{f}_{d,t_2}$ which is extracted from the same encoder M_E in the network f. To complete the propagation, we initially employ a long-range cross-frame attention operation in the feature space. A traditional attention operation [Vaswani *et al.*, 2017] typically consists of the query vector $\boldsymbol{Q}$, the key vector $\boldsymbol{K}$, and the value vector $\boldsymbol{V}$. The attention relationship is established using $A(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q} \times \boldsymbol{K}^\top) \times \boldsymbol{V}$, where $\times$ represents matrix multiplication. To enable cross-frame attention, we process the feature via dual paths, as

$$\begin{aligned} \boldsymbol{f}'_{d,t_1} &= A(\boldsymbol{f}_{d,t_1}, \boldsymbol{f}_{d,t_1}, \boldsymbol{f}_{d,t_1}), \boldsymbol{f}''_{d,t_1} = A(\boldsymbol{f}_{d,t_1}, \boldsymbol{f}_{d,t_2}, \boldsymbol{f}_{d,t_2}), \\ \boldsymbol{f}'_{d,t_2} &= A(\boldsymbol{f}_{d,t_2}, \boldsymbol{f}_{d,t_2}, \boldsymbol{f}_{d,t_2}), \boldsymbol{f}''_{d,t_2} = A(\boldsymbol{f}_{d,t_2}, \boldsymbol{f}_{d,t_1}, \boldsymbol{f}_{d,t_1}). \end{aligned} \tag{9}$$

Moreover, a short-range fusion operation is set as the refinement operation to the propagation at both spatial and channel levels. This can be written as

$$\begin{aligned} \hat{\boldsymbol{f}}_{d,t_1}, \hat{\boldsymbol{f}}_{d,t_2} = S(C((\boldsymbol{f}_{d,t_1} + \boldsymbol{f}'_{d,t_1} + \boldsymbol{f}''_{d,t_1}) \oplus \\ (\boldsymbol{f}_{d,t_2} + \boldsymbol{f}'_{d,t_2} + \boldsymbol{f}''_{d,t_2}))), \end{aligned} \tag{10}$$

where $\oplus$ represents channel concatenation, C refers to the convolution network, and S signifies channel dimension splitting. The decomposition results for $\boldsymbol{I}_{d,t_1}$ and $\boldsymbol{I}_{d,t_2}$ are obtained by processing $\hat{\boldsymbol{f}}_{d,t_1}$ and $\hat{\boldsymbol{f}}_{d,t_2}$ through the decoder M_D. We have confirmed that our dual network with synchronous supervision for the outputs of $\boldsymbol{I}_{d,t_1}$ and $\boldsymbol{I}_{d,t_2}$ can yield superior decomposition and enhancement results compared to individual enhancement strategies.

The Role of the Cross-attention in CFIM. The cross-attention mechanism facilitates information propagation across different frames, aligning with our training procedure that requires spatial consistency for $\boldsymbol{R}$. Moreover, during training, CFIM brings varying filtered features from randomly selected reference images ($\boldsymbol{I}_{d,t_2}$) into the backbone (i.e., the part without feature propagation), where the input is the image at the current time step $\boldsymbol{I}_{d,t_1}$. This setup enables CFIM to guide the backbone in learning how to adaptively leverage diverse knowledge from reference images, resulting in robust and effective feature spaces within the backbone, without the need for expensive temporal alignment. CFIM also leverages cross-frame information when it complements backbone.

3.4 Overall Objective to Compute Loss Function

During the training, we find it is better to adopt the dual training strategy, i.e., for each $\boldsymbol{I}_{d,t_1}$, we sample its neighboring reference $\boldsymbol{I}_{d,t_2}$ and constrain their outputs simultaneously. Thus, the loss function can be written as

$$\begin{aligned} \mathcal{O}_v = \mathbb{E}_{t_1,t_2}[\mathcal{O}_{rec,t_1} + \mathcal{O}_{rec,t_2} + \\ \lambda_1(\mathcal{O}(\boldsymbol{L}_{t_1}) + \mathcal{O}(\boldsymbol{L}_{t_2})) + \lambda_2 \mathcal{O}(\boldsymbol{R}_{t_1,t_2})], \end{aligned} \tag{11}$$

where λ_1 and λ_2 are the loss weights, and each loss term is defined in Equation (5), Equation (6) and Equation (7).

	SDSD-indoor		SDSD-outdoor		SMID		DID	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SNR	27.30	0.84	23.23	0.82	28.49	0.81	24.85	0.90
SMG	27.82	0.86	25.17	0.81	28.03	0.78	25.14	0.85
PairLLE	23.48	0.71	20.04	0.65	22.70	0.63	22.56	0.82
RetinexFormer	26.56	0.79	22.80	0.77	29.15	0.82	25.40	0.89
MBLLEN	22.17	0.66	21.41	0.63	22.67	0.68	24.22	0.86
SMID	24.84	0.72	23.30	0.67	24.78	0.72	22.28	0.84
SMOID	24.63	0.70	22.25	0.68	23.64	0.71	22.13	0.85
SDSDNet	26.81	0.75	23.08	0.71	26.03	0.75	22.52	0.81
DP3DF	27.63	0.74	23.85	0.73	27.19	0.76	22.39	0.88
StableLLVE	25.32	0.70	22.47	0.65	24.37	0.69	21.64	0.80
LLVE-SEG	26.19	0.77	24.09	0.72	25.31	0.74	23.85	0.82
LLVE-CFA	24.28	0.81	22.64	0.76	24.72	0.70	22.53	0.87
BLLRVE	26.02	0.74	23.50	0.71	26.15	0.73	23.25	0.81
VRGSS	27.81	0.76	24.28	0.75	27.84	0.78	24.51	0.87
ZRLLE-PQP	25.53	0.71	22.42	0.73	26.36	0.77	22.84	0.80
JDAE	26.19	0.74	23.37	0.76	26.45	0.76	23.53	0.82
Ours	**28.93**	**0.88**	**26.32**	**0.82**	**29.60**	**0.82**	**30.10**	**0.93**

Table 1: Quantitative comparison on SDSD, SMID, and DID datasets. Our method performs the best consistently.

	SNR	SMG	PairLLE	RetinexFormer	MBLLEN	SMID
PSNR	20.69	20.18	19.57	21.79	18.63	20.51
SSIM	0.710	0.672	0.667	0.723	0.619	0.686
	SMOID	SDSDNet	DP3DF	StableLLVE	LLVE-SEG	Ours
PSNR	20.66	21.22	22.04	21.48	21.45	**23.38**
SSIM	0.697	0.716	0.749	0.732	0.715	**0.782**

Table 2: Quantitative comparison on the DAVIS dataset.

	DAVIS		SDSD-Indoor		SDSD-Outdoor		DID	
Methods	Short	Long	Short	Long	Short	Long	Short	Long
SNR	0.027	0.070	0.017	0.060	0.022	0.046	0.025	0.068
RetinexFormer	0.029	0.072	0.018	0.061	0.024	0.043	0.024	0.065
SDSDNet	0.024	0.068	0.012	0.053	0.011	0.038	0.017	0.059
DP3DF	0.021	0.065	**0.011**	**0.050**	0.013	0.036	0.019	0.062
StableLLVE	0.023	0.067	0.016	0.055	0.019	0.042	0.023	0.066
LLVE-SEG	0.025	0.071	0.014	0.057	0.016	0.040	0.021	0.064
Ours-R	**0.017**	**0.058**	**0.010**	**0.048**	**0.007**	**0.030**	**0.011**	**0.049**
Ours	**0.020**	**0.063**	**0.012**	**0.051**	**0.010**	**0.034**	**0.015**	**0.054**

Table 3: The quantitative comparison in terms of short-term ("Short") and long-term ("Long") temporal loss. "Ours-R" means the view-independent term produced by ours.

4 Experiments

4.1 Datasets

Our evaluation is conducted on four publicly available datasets, which encompass a wide range of real-world videos with diverse motion patterns and degradations, including SMID [Chen *et al.*, 2019], SDSD [Wang *et al.*, 2021], DID [Fu *et al.*, 2023a], and DAVIS [Pont-Tuset *et al.*, 2017].

4.2 Implementation Details

Experimental Details. We conducted experiments on all datasets using the same network structure. T is set as 5 in the experiment. All modules were trained end-to-end, with the learning rate initialized at $4e^{-4}$ for all layers, adapted by a cosine learning scheduler. The batch size used was 4. Correspondences were computed using the SOTA method DKM [Edstedt *et al.*, 2023]. We used the pre-trained weights of DKM during training because it is designed for general indoor and outdoor scenes. Its generalization ability is supported by its extensive training data and state-of-the-art training strategy, as demonstrated in the original DKM paper and subsequent studies [Zhu and Liu, 2023; Edstedt *et al.*, 2024] (through evaluations on unseen scenes).

4.3 Comparison and Baselines

We conducted a comprehensive comparison with SOTA LLVE methods, including MBLLEN [Lv *et al.*, 2018], SMID [Chen *et al.*, 2019], SMOID [Jiang and Zheng, 2019], SDSDNet [Wang *et al.*, 2021], DP3DF [Xu *et al.*, 2023a], StableLLVE [Zhang *et al.*, 2021], LLVE-SEG [Liu *et al.*, 2023], LLVE-CFA [Chhirolya *et al.*, 2022], BLLRVE [Zhang *et al.*, 2024], VRGSS [Li *et al.*, 2023], ZRLLE-PQP [Wang *et al.*, 2024], and JDAE [Shi *et al.*, 2024]. We also compared our method with SOTA LLIE methods for individual frames, including SNR [Xu *et al.*, 2022], SMG [Xu *et al.*, 2023b], PairLLE [Fu *et al.*, 2023b], RetinexFormer [Cai *et al.*, 2023]. All methods were trained on each dataset using their respective released code and hyper-parameters (e.g., the training epoch that is set to guarantee convergence). *Note that all baselines are trained on our unified data split for a fair comparison. The splits slightly differ from those in baselines' original papers, so scores of baselines may vary from original papers.*

Quantitative Result. In TABLE 1, we present the comparative results with the selected baseline methods across the SDSD, SMID, and DID datasets. The table reveals that our approach *consistently* outperforms all other methods, as indicated by the highest PSNR and SSIM scores on all datasets. Notably, our scores exhibit a substantial lead over all others, particularly on DID, which is a large-scale dynamic video dataset. This superiority underscores the robust capability of our method in enhancing real-world videos.

Furthermore, TABLE 2 provides a summary of the comparative results on DAVIS. In comparison to the degradation synthesis strategy as presented in [Zhang *et al.*, 2021], we have extended our approach to include the degradation of a low-light noise term. As a result, our synthesized data encompasses dynamic scenes with pronounced invisibility and perturbations, posing a considerable challenge. As illustrated in TABLE 2, our method consistently yields the highest PSNR and SSIM values, reaffirming the effectiveness of our approach.

Evaluation for Temporal Consistency. For video processing, the performance of temporal consistency and stability should be evaluated. Thus, we employ the short-term and long-term temporal loss proposed in [Lai *et al.*, 2018] for such temporal evaluation on different datasets. The wrapping operations among frames are computed on the frames with the normal light. Moreover, the long-term loss is computed every 10 frames. The results are shown in TABLE 3 (we normalize the frame values into [0, 1]). Obviously, our results have lower temporal loss than the baselines, e.g., higher temporal consistency and stability. Specifically, we assess the temporal consistency of the view-independent term. The results presented in TABLE 3 further validate the constancy.

Qualitative Result. Besides the quantitative comparison, we present visual comparisons with the selected baselines. Fig. 4 showcases the visual comparisons of SDSD and SMID, while Fig. 5 displays visual cases from DID and DAVIS. In gen-

Figure 4: Visual comparisons on SDSD-indoor (top two rows), SDSD-outdoor (middle two rows), and SMID (bottom two rows). The results of our proposed framework, i.e., "Ours", demonstrate better visual perception with clearer visibility and more enhanced details.

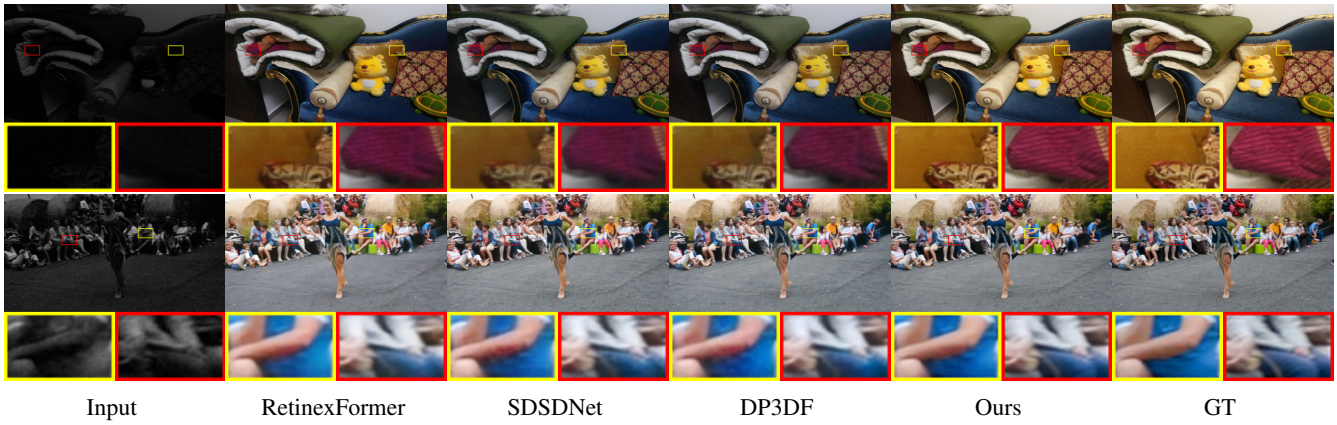

Figure 5: Comparisons on DID (top two rows) and DAVIS (bottom two rows). "Ours" has better visibility and details.

eral, the results enhanced by our approach exhibit a more natural appearance, including accurate color, well-balanced brightness, enhanced contrast, and precise details. Our results show fewer artifacts in regions with complex textures, and they appear cleaner and sharper compared to results produced by other methods. This distinction is particularly notable, as most baselines perform well in simpler areas.

4.4 Ablation Study

We assess the critical components of our framework through five ablation cases. (1) "w/o LR" that removes the constraint on learning the view-independent part. (2) "w/o LL" where the constraint on learning the view-dependent part is removed. (3) "w/o C.F.": we eliminate the cross-frame attention and fusion in the network, resulting in each frame being enhanced individually. (4) "with M.I.": This refers to using multiple neighboring frames as input and employing the temporal alignment of deformable convolution. (5) "w/o Dual": the dual learning strategy is removed, meaning that the loss is applied to only one frame in each iteration.

	SDSD-indoor		SDSD-outdoor		SMID		DID	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o LR	25.78	0.77	23.49	0.75	26.37	0.76	27.54	0.86
w/o LL	26.89	0.80	24.19	0.78	27.74	0.79	27.72	0.89
w/o C.F.	25.18	0.82	24.22	0.77	26.75	0.77	25.60	0.85
with M.I.	26.21	0.84	25.36	0.80	28.42	0.79	28.33	0.88
w/o Dual	27.53	0.84	24.83	0.79	27.43	0.78	26.91	0.86
Full	**28.93**	**0.88**	**26.32**	**0.82**	**29.60**	**0.82**	**30.10**	**0.93**

Table 4: Ablation study on SDSD, SMID, and DID.

The results are summarized in TABLE 4. By comparing "w/o LR" with the full setting ("Full"), we can clearly demon-

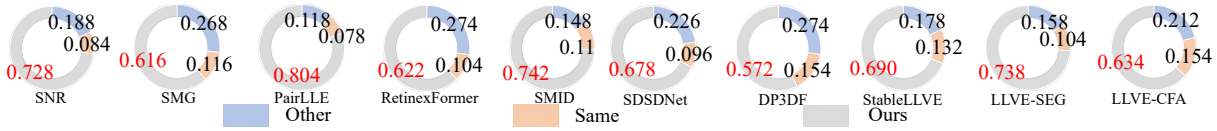

Figure 6: The above pie charts summarize the results of our user study, and ours is preferred by participants.

strate the effectiveness of our proposed constraint for the view-independent part. Similarly, the significant advantage of "Full" over "w/o LL" highlights the significance of the proposed continuity constraint. Furthermore, we can validate the importance of mutually propagating features among different frames to achieve spatial-temporal consistent decomposition by comparing "w/o C.F." with "Full". Our proposed propagation strategy via the dual network structure proves to be more effective than the common method of using multi-frame inputs for propagation in videos, as seen in the comparison between "w/o M.I." and "Full". The dual learning setting, where we simultaneously supervise the two outputs of the dual network, also proves to be valuable, as evidenced by the comparison between "w/o Dual" and "Full".

4.5 User Study

To prove the effectiveness of our proposed framework in terms of human subjective evaluation, we conduct a large-scale user study with 100 participants with varying ages and education backgrounds, and balanced sex distribution. Following most of the existing low-light enhancement works [Xu *et al.*, 2023b; Wang *et al.*, 2023a], we employ the AB-test for the user study. Participants should indicate their preference or select the same option. They should make the decision according to the natural brightness, contrast, and color of each frame; the rich details, fewer artifacts, and the temporal consistency in the video.

Fig. 6 summarizes the user study's results, and we can see that ours gets more selections from participants over all the baselines. This demonstrates that our method's results are more preferred by the human subjective perception.

4.6 The Effects of CFIM Structure

As previously discussed, the use of CFIM yields two key benefits: it enhances the consistency of the image enhancement process and strengthens the robustness of the backbone by providing the cross-attention mechanism with information from random neighboring time steps. In this section, we conduct experiments to show the effects of CFIM. The baseline consists of a network that processes a single frame, while the comparative setup integrates CFIM, where cross-attention is applied between randomly selected pairs of frames during training. The results of this comparison are presented in Table 5, which show that models incorporating the CFIM structure significantly outperform those without it, thereby demonstrating the effectiveness of the CFIM approach. Besides, the ablation results of the mentioned "w/o C.F." (in Sec. 4.4) also support the effects of CFIM in the decomposition.

4.7 Robustness Towards Correspondences

In this section, we explore the robustness of our framework with respect to incorrect and insufficient correspondences. In the first experiment, we introduce random noise perturbations to the computed correspondences, modifying the location values of the correspondences (the perturbation range is sampled from $-20 \sim 20$ pixels, which falls within the normal error range of current correspondence estimation methods [Edstedt *et al.*, 2023]). As shown in Table 6, the results with perturbed correspondences (denoted as "Ours with per.") show a performance drop compared to the unperturbed case. However, the performance remains superior to that of most SOTA baselines, indicating that our framework is robust to a certain of erroneous correspondences, which are likely to occur in areas with complex textures or other challenging scenarios. In the second experiment, we address the issue of insufficient correspondences, which can arise, for instance, in regions with occlusions. To simulate this, we randomly reduce the number of correspondences to 10% of the originally detected ones. The results ("Ours with red.") are presented in Table 6, where we observe that our method continues to outperform almost all baselines in Table 1, albeit with a smaller margin than the original setting. This suggests that even with fewer correspondences, the remaining ones still provide valuable information and they are in accord with the other constraints. This further demonstrates the robustness of our method.

	SDSD-indoor		SDSD-outdoor		SMID		DID	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SNR	27.30	0.84	23.23	0.82	28.49	0.81	24.85	0.90
SNR+CFIM	**28.15**	**0.86**	**24.57**	**0.83**	**29.24**	**0.82**	**26.03**	**0.91**
R.F.	26.56	0.79	22.80	0.77	29.15	0.82	25.40	0.89
R.F.+CFIM	**27.42**	**0.80**	**23.75**	**0.79**	**29.87**	**0.83**	**26.38**	**0.90**

Table 5: The effects of CFIM. "R.F." denotes RetinexFormer.

	SDSD-indoor		SDSD-outdoor		SMID		DID	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Ours with red.	27.97	0.84	25.73	0.80	29.12	0.79	28.86	0.89
Ours with per.	28.10	0.86	26.04	0.79	29.37	0.81	29.08	0.91
Ours	28.93	0.88	26.32	0.82	29.60	0.82	30.10	0.93

Table 6: The robustness of our framework towards correspondences.

5 Conclusion

We introduce a novel LLVE framework that utilizes spatial-temporal consistent decomposition and dual networks for mutual feature propagation. It predicts the decomposition of normal-light outputs without external priors, thanks to cross-frame correspondences and practical constraint of the continuity. Extensive experiments on various datasets showcase the framework's superiority over SOTA approaches and the impact of our designed components.

Ethical Statement

There are no ethical issues.

Acknowledgments

This work is supported by the Natural Science Foundation of Zhejiang Province, China, under No. LD24F020002.

References

[Cai *et al.*, 2023] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023.

[Chen *et al.*, 2019] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *ICCV*, 2019.

[Chhirolya *et al.*, 2022] Shivam Chhirolya, Sameer Malik, and Rajiv Soundararajan. Low light video enhancement by learning on static videos with cross-frame attention. In *BMVC*, 2022.

[Edstedt *et al.*, 2023] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023.

[Edstedt *et al.*, 2024] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, 2024.

[Fu *et al.*, 2023a] Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, and Huadong Ma. Dancing in the dark: A benchmark towards general low-light video enhancement. In *ICCV*, 2023.

[Fu *et al.*, 2023b] Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *CVPR*, 2023.

[Hasinoff *et al.*, 2016] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Barron Jonathan T., Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2016.

[Ignatov *et al.*, 2017] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017.

[Jiang and Zheng, 2019] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *ICCV*, 2019.

[Jiang *et al.*, 2021] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. EnlightenGAN: Deep light enhancement without paired supervision. *IEEE TIP*, 2021.

[Lai *et al.*, 2018] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, 2018.

[Li *et al.*, 2023] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *CVPR*, 2023.

[Liang *et al.*, 2023] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *ICCV*, 2023.

[Liu *et al.*, 2023] Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yan Feng Wang, and Qi Tian. Low-light video enhancement with synthetic event guidance. In *AAAI*, 2023.

[Lv *et al.*, 2018] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. MBLLEN: Low-light image/video enhancement using CNNs. In *BMVC*, 2018.

[Lv *et al.*, 2023] Xiaoqian Lv, Shengping Zhang, Chenyang Wang, Weigang Zhang, Hongxun Yao, and Qingming Huang. Unsupervised low-light video enhancement with spatial-temporal co-attention transformer. *TIP*, 2023.

[Ma *et al.*, 2022] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *CVPR*, 2022.

[Narihira *et al.*, 2015] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *CVPR*, 2015.

[Pont-Tuset *et al.*, 2017] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv*, 2017.

[Shi *et al.*, 2024] Yiqi Shi, Duo Liu, Liguo Zhang, Ye Tian, Xuezhi Xia, and Xiaojing Fu. Zero-ig: Zero-shot illumination-guided joint denoising and adaptive enhancement for low-light images. In *CVPR*, 2024.

[Triantafyllidou *et al.*, 2020] Danai Triantafyllidou, Sean Moran, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Low light video enhancement using synthetic data produced with an intermediate domain mapping. In *ECCV*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2019] Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, and Tao Yue. Enhancing low light videos by exploring high sensitivity camera noise. In *ICCV*, 2019.

[Wang *et al.*, 2021] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: High-quality video dataset with mechatronic alignment. In *ICCV*, 2021.

[Wang *et al.*, 2022] Wenjing Wang, Xinhao Wang, Wenhan Yang, and Jiaying Liu. Unsupervised face detection in the dark. *IEEE TPAMI*, 2022.

[Wang *et al.*, 2023a] Haoyuan Wang, Xiaogang Xu, Ke Xu, and Rynson WH Lau. Lighting up nerf via unsupervised decomposition and enhancement. In *ICCV*, 2023.

[Wang *et al.*, 2023b] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023.

[Wang *et al.*, 2024] Wenjing Wang, Huan Yang, Jianlong Fu, and Jiaying Liu. Zero-reference low-light enhancement via physical quadruple priors. In *CVPR*, 2024.

[Wu *et al.*, 2022] Yirui Wu, Haifeng Guo, Chinmay Chakraborty, Mohammad Khosravi, Stefano Berretti, and Shaohua Wan. Edge computing driven low-light image dynamic enhancement for object detection. *IEEE Transactions on Network Science and Engineering*, 2022.

[Xu *et al.*, 2022] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-Aware low-light image enhancement. In *CVPR*, 2022.

[Xu *et al.*, 2023a] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Deep parametric 3d filters for joint video denoising and illumination enhancement in video super resolution. In *AAAI*, 2023.

[Xu *et al.*, 2023b] Xiaogang Xu, Ruixing Wang, and Jiangbo Lu. Low-light image enhancement via structure modeling and guidance. In *CVPR*, 2023.

[Yang *et al.*, 2021] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality. *IEEE TIP*, 2021.

[Ye *et al.*, 2023] Jing Ye, Changzhen Qiu, and Zhiyong Zhang. Spatio-temporal propagation and reconstruction for low-light video enhancement. *Digital Signal Processing*, 2023.

[Yuille, 2012] A.L. Yuille. The lambertian reflectance model. https://www.cs.jhu.edu/ ayuille/courses/Stat238-Winter12/lecture10.pdf, 2012.

[Zhang *et al.*, 2021] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, 2021.

[Zhang *et al.*, 2024] Gengchen Zhang, Yulun Zhang, Xin Yuan, and Ying Fu. Binarized low-light raw video enhancement. In *CVPR*, 2024.

[Zhu and Liu, 2023] Shengjie Zhu and Xiaoming Liu. Pmatch: Paired masked image modeling for dense geometric matching. In *CVPR*, 2023.