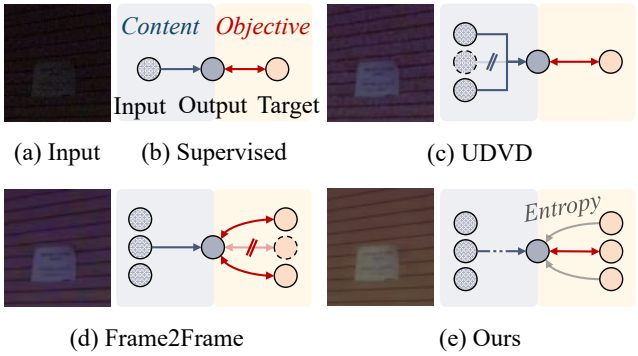
Seeing Dark Videos via Self-Learned Bottleneck Neural Representation

Haofeng Huang[1], Wenhan Yang[2], Ling-Yu Duan[1], Jiaying Liu[1]*

[1]Peking University, Beijing, China,
[2]Peng Cheng Laboratory, Beijing, China
hhf@pku.edu.cn, yangwh@pcl.ac.cn, lingyu@pku.edu.cn, liujiaying@pku.edu.cn

Abstract

Enhancing low-light videos in a supervised style presents a set of challenges, including limited data diversity, misalignment, and the domain gap introduced through the dataset construction pipeline. Our paper tackles these challenges by constructing a self-learned enhancement approach that gets rid of the reliance on any external training data. The challenge of self-supervised learning lies in fitting high-quality signal representations solely from input signals. Our work designs a bottleneck neural representation mechanism that extracts those signals. More in detail, we encode the frame-wise representation with a compact deep embedding and utilize a neural network to parameterize the video-level manifold consistently. Then, an entropy constraint is applied to the enhanced results based on the adjacent spatial-temporal context to filter out the degraded visual signals, *e.g.* noise and frame inconsistency. Last, a novel Chromatic Retinex decomposition is proposed to effectively align the reflectance distribution temporally. It benefits the entropy control on different components of each frame and facilitates noise-to-noise training, successfully suppressing the temporal flicker. Extensive experiments demonstrate the robustness and superior effectiveness of our proposed method. Our project is publicly available at: https://huangerbai.github.io/SLBNR/.



(a) Input (b) Supervised (c) UDVD

(d) Frame2Frame (e) Ours

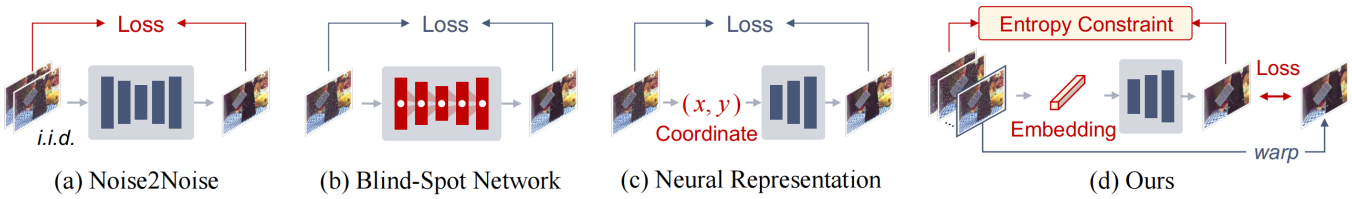
Figure 1: Results and frameworks of different methods. (b) The learning paradigm of supervised methods; (c) Even bounded by the content bottleneck, UDVD (Sheth et al. 2021)'s results still suffer from remaining noise due to the overfitting caused by self-supervision; (d) Frame2Frame (Ehret et al. 2019) introduces artifacts as solely the objective bottleneck might incur misalignment during model optimization; (e) Our method constructs the bottleneck in both content and objective views, which leads to superior results.

Introduction

Videos captured in a low-light environment suffer from severe visual degradation. Common intuitive choices (*e.g.* long exposure or high ISO) are not satisfactory for videos. It is difficult to apply the former for video, while the latter leads to heavy noise. Therefore, low-light video enhancement from the software perspective is highly demanded. This task is challenging because of the complex noise distribution and flicker (Wei et al. 2022; Jiang et al. 2022), which damages the spatial structure and the temporal consistency.

In the deep-leaning era, many learning-based low light enhancement methods (Chen et al. 2019; Jiang and Zheng 2019; Wang et al. 2021) are proposed. Most of them adopt full-supervision to pursue better restoration performance, which learns the mapping strategy from low-light images/videos to normal-light ones with an end-to-end learned neural network. To serve that, several paired datasets (Chen et al. 2019; Jiang and Zheng 2019; Wang et al. 2021) are collected for training and evaluation. However, it is resource-intensive to construct a real paired low light video dataset, which demands professional equipment *e.g.* an electric slide rail (Wang et al. 2021) or a split optic (Jiang and Zheng 2019), and careful registration. Furthermore, this means of collection assumes that the directly captured videos are normal-light ground truth and processed videos processed

*Corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 62332010 and Grant 62088102, in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation.



(a) Noise2Noise (b) Blind-Spot Network (c) Neural Representation (d) Ours

Figure 2: The overview of the self-supervised methodology for restoration. Red lines/modules denote where the bottleneck is employed to screen out the visual degradation, aiding in the reconstruction of the intrinsic signal from the input.

by physics devices, *e.g.* neutral density filtered videos, are low-light ones. In this way, the scene is actually normally lit and low-light videos are simulated by decreasing the input photon, whose illumination distribution is notably different from authentic dark ones. Some researchers endeavour to construct unpaired/self-supervised learning methods to enhance low-light images (Guo et al. 2020; Liu et al. 2021; Jiang et al. 2021; Ma et al. 2022) without the need of paired datasets. Typically, these approaches exploit the intrinsic prior of normal images, *e.g.* illumination distribution or histogram, to facilitate the enhancement of low-light images and improve their visual quality. However, owing to the lack of robust supervision, the majority of these methods struggle to deal with intensive noise.

Some restoration methods focus on learning to suppress noise, *e.g.* Noise2Noise (Lehtinen et al. 2018) and Blind-Spot Network (Krull, Buchholz, and Jug 2019), with the pixel-level self-supervision that takes the input image itself as the ground truth. Noise2Noise paradigm (Fig. 2 (a)) assumes different noisy versions of the input follows *i.i.d.* and deploys the **bottleneck on the objective**. Blind-Spot Network (Fig. 2 (b)) limits the information flow from the input to the output by controlling the receptive field of the network, called **content bottleneck**. However, these two routes face challenges when enhancing a dark video. The former relies on the *i.i.d.* assumption, which is seldom guaranteed in low-light videos. The latter follows certain noise models (Sheth et al. 2021), which fails to describe the real noise in low-light conditions. Fig. 1 provides results with the implementation of both methodology on self-learned video enhancement, which still show severe visual distortions.

In essence, the critical aspect of self-supervised methods resides in their capacity to extract high-quality signal representations from noisy inputs with a bottleneck mechanism to effectively filter out the visual degradation signal. In this work, we focus on designing novel bottleneck mechanisms from both objective and content perspectives, which offers an enhanced means to regulate the information flow for self-supervised learning. In detail, we develop a self-learned low-light video enhancement method based on bottleneck neural representation. First, to limit the information from the input content, besides taking the coordinates as input as shown in Fig. 2 (c), the proposed neural representation (Fig. 2 (d)) additionally takes a compact deep embedding as input, which limits the information from the input content while still offering richer content. The dimension of this deep embedding is kept low. Being trained over the whole video, it implicitly integrates temporal information into the network parameters, which naturally leads to temporal consistency. For the objective bottleneck, we design a learnable entropy estimation and constraint to suppress intensive noise and control the illumination distribution. Furthermore, instead of directly predicting the signal, we propose to generate a layer-wise representation of given frames. Compared with the original Retinex model (Fu et al. 2016), the novel layer-wise representation fully considers the characteristics of low-light videos and naturally disentangles the visual degradation, *i.e.* noise and color bias. It also resolves the distribution bias of adjacent frames, facilitating noise-to-noise training and benefits the entropy control on different components of the signal. With experiments of no-reference evaluation, the superiority of our method over the state-of-the-art is verified and the effectiveness of our design is demonstrated. Note that, our method does not rely on any external data.

Our contributions are summarized as follows:

- We make the pioneering effort in devising a self-supervised deep approach for low-light video enhancement. It integrates the bottleneck mechanisms from both the content and objective perspectives as well as a chromatic Retinex model, obtaining satisfactory visual quality without using any external data.
- For the content bottleneck, a hybrid neural representation is introduced. A learnable low-dimension deep embedding provides richer content information. The model is trained over the whole video, whose parameters implicitly integrate temporal information and naturally lead to temporal consistency.
- For the objective bottleneck, an entropy constraint is applied to the predicted results. Intensive noise and biased illumination are suppressed with the objective to reduce the entropy of the signal.
- A novel Chromatic Retinex model is proposed to transform the signal into layer-wise representation. It benefits explicit entropy control on different components of results and better aligns distributions for noise-to-noise training.

Related Works

Low Light Enhancement

Enhanced imaging in a photon-limited scene is a long-standing demand because of intensive hardware noise and inaccurate white balance (Huang et al. 2022; Li et al. 2022). In the deep learning era, focusing on image enhancement,

researchers began to collect real paired datasets (Wei et al. 2018; Chen et al. 2018) in a static scene with twice capturing which facilitates supervised learning. Following researchers came up with more diverse architectures (Wang et al. 2022; Ren et al. 2019) and various loss functions (Cai, Gu, and Zhang 2018; Wang et al. 2019a) for better illumination adjustment and details reconstruction. In many works (Wei et al. 2018; Zhang, Zhang, and Guo 2019; Zhang et al. 2021; Wang et al. 2019b), Retinex theory is integrated into the network, providing a physical model-based decomposition to satisfy the human visual system.

For a video shot in a dark dynamic scene, such methodology does not work because long exposure is unavailable and limited to the frame rate. Some researchers (Chen et al. 2019; Jiang and Zheng 2019; Wang et al. 2021) made valuable efforts to collect paired datasets. However, in the need of capturing bright videos as ground truth, the scene is actually well-lit. It brings an inevitable domain gap. Therefore, efficient unsupervised low-light video enhancement methods are needed to save the effort of dataset collection and resolve the issue of domain gap with less dependence on the training set.

Self-supervised Restoration

With similar motivation for the denoising task, self-supervised restoration methods (Lehtinen et al. 2018; Krull, Buchholz, and Jug 2019; Huang et al. 2021) are proposed. These methods are based on the self-regression and integrate prior (Lehtinen et al. 2018; Huang et al. 2021) or regularization (Krull, Buchholz, and Jug 2019) into the network and training dynamics. However, the noise model in a dark scene is hugely disturbed and hybrid (Wei et al. 2022; Monakhova et al. 2022), which is usually in contradiction with the assumption of unsupervised methods. Therefore, trivially enhancing dark images/videos using a denoising module concatenated with another lightening module can not guarantee a promising performance.

Implicit neural representation (Chen and Zhang 2019) parameterizes a signal with continuous functions via neural network. It only takes the coordinate as the input and outputs the corresponding value, *e.g.* frame number for videos (Chen et al. 2021). Based on the deep prior (Ulyanov, Vedaldi, and Lempitsky 2018), the structure of the neural network acts as a regularization to limit the transmutation of the signal. It has been widely used in 3D vision (Mescheder et al. 2019; Mildenhall et al. 2020), providing a way to encode high-dimension signals. However, generating detailed texture is difficult for implicit neural representation which requires a long-time training without the guidance of local information. Some researchers also attempt to inject local information with a learned embedding (Peng et al. 2020; Yu et al. 2021; Chibane, Alldieck, and Pons-Moll 2020; Chen et al. 2023), showing a performance gain.

Bottleneck Neural Representation

Motivation

As mentioned in Sec. , challenges to low-light video enhancement exist in three aspects:

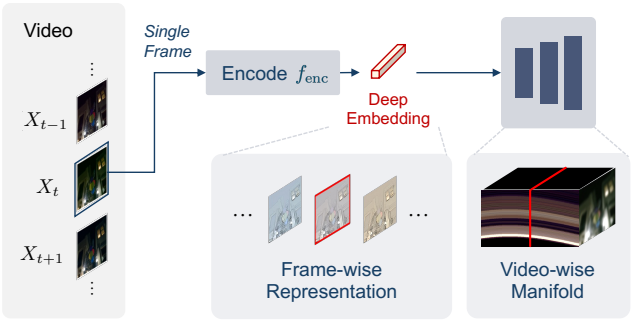


Figure 3: The hybrid neural representation utilizes the content-agnostic coordinate of the implicit neural representation with a content-adaptive compact deep embedding, which provides richer intrinsic information for reconstructing normal-light images.

- Collecting a paired real video dataset is challenging and leads to domain gaps.
- There are severe hybrid distortions in low-light videos, which hardly are well handled by existing bottleneck mechanisms or their simple combinations.
- An unstable white balance in videos causes severe temporal flicker, which leads to the violation of the assumptions in existing self-supervised restoration methods.

Therefore, we propose to construct a **self-learned bottleneck neural representation** for low-light video enhancement **without using any external data**. The **joint bottleneck of content and objective** is employed via *Hybrid Neural Representation* and *Entropy Minimization Model*, respectively, to obtain temporal consistent, noise-suppressed and well-lit results. The proposed *Chromatic Retinex model* decomposes the signal for better enhancement, which alleviates inconsistent distributions and facilitates more effective self-supervision.

Hybrid Neural Representation

Hybrid neural representation (Chen et al. 2023) is in between the explicit (embedding-centric) and implicit (network-centric) representation. As mentioned in Sec. , implicit neural representation encodes the information into network parameters, where the input is a content-agnostic coordinate, *e.g.* (x, y) for 2D images, (x, y, z) for 3D representation or videos and (x, y, z, σ, ϕ) for Neural Radiance Field (NeRF) (Mildenhall et al. 2020). While this formulation forces the network to learn a continuous manifold space, it does so at the cost of neglecting local information, making it challenging for the network to generate fine-grained details (Yu et al. 2021; Peng et al. 2020). To tackle this, Chen et al. (2023) proposes to replace the coordinate with an extracted deep embedding, which includes rich semantic information. Here we extend the network for video restoration.

Content Bottleneck on Embeddings. Neural representation of videos (Chen et al. 2021) set content bottleneck on the input of the network for enhancement, *i.e.* using only the

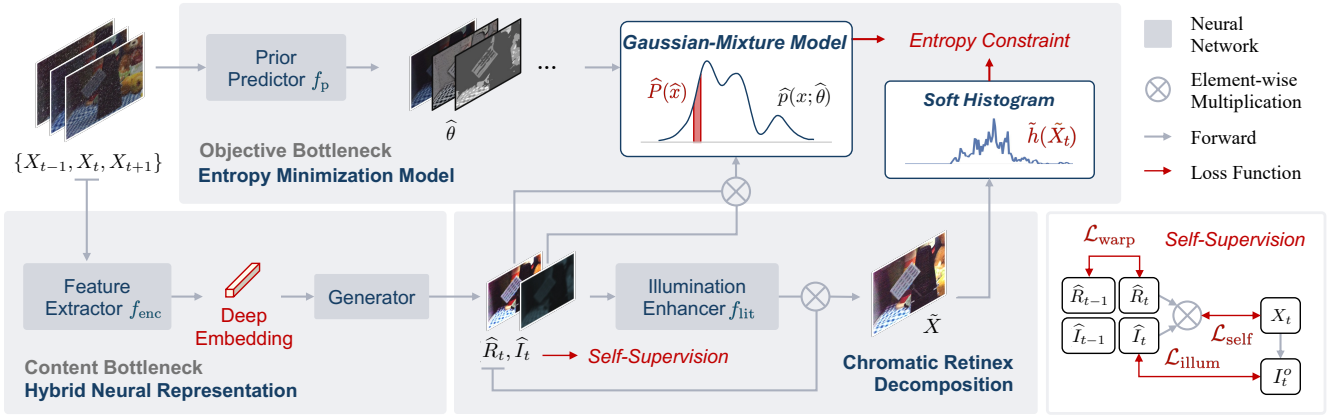


Figure 4: The framework of the proposed bottleneck neural representation. A constrained deep embedding is first extracted and then transformed into enhanced Retinex-based layer-wise representations. Hybrid neural representation provides richer intrinsic information but still set bottlenecks from the perspective of content. Entropy minimization applies the bottleneck constraint in the objective view to suppress noise and correct illumination. A chromatic Retinex representation helps align layer-wise frames, which facilitates self-supervised learning.

index t of the frame X_t as the image-wise representation:

$$f_{\text{dec}}(t) = \widehat{X}_t, \tag{1}$$

where $f_{\text{dec}}(\cdot)$ denotes a neural network as decoder. Mildenhall et al. (2020) demonstrates its rendering capacity of the neural network with only t to reconstruct a sequence of high-quality photos. But when applied to low-light videos, the presence of corrupted frames significantly complicates the learning process of this mapping, as highlighted in (Mildenhall et al. 2022). Following the design of Yu et al. (2021) and Chen et al. (2023), to improve the modeling capacity to regress details, we replace the coordinate with a compact learned embedding that brings richer information:

$$f_{\text{dec}}(z_t) = \widehat{X}_t, \quad z_t = f_{\text{enc}}(X_t), \tag{2}$$

where $f_{\text{enc}}(\cdot)$ denotes the encoder network and z_t is the compact embedding. The content bottleneck is guaranteed by limiting the dimension of z. As shown in Fig 3, the embedding additionally retains content-adaptive information and alleviates the learning burden of the network compared with only taking coordinates as the input. On the other hand, z is so compact that most information is derived from the network parameters, which implicitly forces z only to record intrinsic signals instead of noise. As z is compact, this hybrid way naturally leads to a more intrinsic and temporally consistent manifold with sufficient details. It is optimized on the given sequence with self-regressed Mean-Square-Error (MSE) loss. Naturally, this self-regression needs to consider getting rid of fitting degradation, *i.e.* avoiding the network over-fitting the noise, which are explored in the following.

Implicit Multi-Frame Fusion. One of the most common ways is to fuse the information from multiple frames and utilize temporal consistency to suppress noise. However, in low-light conditions, severely degraded frames cannot be robustly pre-aligned.

Our neural representation can perform an implicit multi-frame fusion that utilizes temporal information effectively. Decoder's parameters are shared across frames, after training on the input videos, the model naturally learns to generate temporal consistent results. Besides the neural representation that injects the temporal information into network parameters, we further regularize the implicit fusion with a warping loss $\mathcal{L}_{\text{warp}}$:

$$\mathcal{L}_{\text{warp}} = d\Big(\widehat{X}_t, \texttt{warp}\big(\widehat{X}_{t-1}, o(\widehat{X}_t, \widehat{X}_{t-1})\big)\Big), \tag{3}$$

where $\texttt{warp}(\cdot,\cdot)$ takes the former prediction $\widehat{X}_{t-1}$ and optical flow between them $o(\widehat{X}_t, \widehat{X}_{t-1})$, then predicts the warped result. Optical flow is calculated with TV-L1 algorithm (Sánchez Púrez, Meinhardt-Llopis, and Facciolo 2013) based on predicted frames. Note that Eqn. (3) derives from Noise2Noise (Lehtinen et al. 2018) paradigm, which is invalid without the *i.i.d.* assumption because of flicker. It will be further discussed in Sec. , where we design a chromatic Retinex model to alleviate this issue.

Entropy Minimization Model

In addition to the above-mentioned neural representations that restrict information flow from the input, we also explore creating an entropy minimization model that incorporates a bottleneck from an objective perspective to suppress noise and correct the illumination distribution.

Objective Bottleneck to Suppress Noise. Based on the high-entropy nature of noise, we propose to model the distribution of the noisy signal using a Gaussian-mixture model and then suppress noise by minimizing the corresponding entropy. Conventional losses for the objective bottleneck can be derived from certain statistic models. For example, the Mean Square Error (MSE) loss is based on the maximum likelihood of the Gaussian distribution. However, the distribution of noisy signals in low-light videos is complex and hybrid (Wei et al. 2022). Therefore, we adopt the Gaussian-mixture model and use a deep network to predict the prior

distribution in a variational manner:

$$\begin{aligned}\widehat{\theta} &= \{\mu_t^i, \sigma_t^i, w_t^i\} \\ &= f_{\mathrm{p}}(\{X_{t-1}, X_t, X_{t+1}\}),\end{aligned} \tag{4}$$

where $\widehat{\theta}$ is the parameters for Gaussian-mixture model and f_{p} denotes the prior predictor network. As shown in Fig. 4, $\widehat{\theta}$ is used to construct probability density function $\widehat{p}$:

$$\widehat{p}(x;\widehat{\theta}) = \sum_{i=1}^{M} w_t^i \, \mathcal{N}(x;\mu_t^i,\sigma_t^i), \tag{5}$$

where $\mathcal{N}(\cdot;\cdot)$ denotes Gaussian distribution. Given a pixel $\widehat{x}$ from frame $\widehat{X}_t$, the probability $\widehat{P}(\widehat{x})$ is:

$$\widehat{P}(\widehat{x}) = \int_{x-\delta/2}^{x+\delta/2} \widehat{p}(x;\widehat{\theta})dx, \tag{6}$$

where δ is set as the bin size $1/255$ for $\widehat{x} \in [0,1]$ considering the quantization for output. Therefore, the mean entropy of frame $\widehat{X}_t$ is:

$$\mathcal{L}_{\mathrm{gmm}} = E(\widehat{X}_t) = \frac{1}{HW}\sum_{\widehat{x}} -\log_2 \widehat{P}(\widehat{x}). \tag{7}$$

It is also the maximum likelihood estimation of the given Gaussian-mixture model. To bottleneck the objective, when predicting the distribution of pixel value of $\widehat{X}[i]$, center pixel $X[i]$ is masked using the blind-spot strategy.

Objective Bottleneck to Correct Illumination. Low-light frames usually show biased illumination distribution and exhibit color flicker. To prevent overfitting to this distortion, we regulate the entropy of the soft histogram as outlined in (Liang et al. 2022), along with the cross-entropy of the conditioned soft histogram based on the average frame. The former objective draws inspiration from the principles of histogram equalization (Abdullah-Al-Wadud et al. 2007), while the latter assesses the deviation of the current frame from the averaged distribution.

Let $\mathcal{S}(\cdot)$ denote the sigmoid function. Given a predicted result $\widehat{X} \in [0,1]$, the soft histogram $\tilde{h}(\widehat{X})$ is defined by:

$$\begin{aligned}\tilde{h}(\widehat{X})[j] = \sum \mathcal{S}(\widehat{X} - \frac{j}{255} + \frac{\delta}{2}) - \mathcal{S}(\widehat{X} - \frac{j}{255} - \frac{\delta}{2}), \\ j \in \{0,1,\cdots,255\}.\end{aligned} \tag{8}$$

It is a relaxation of histogram $h(\widehat{X})$ to enable backpropagation. The entropy of the soft histogram is given by:

$$\mathcal{L}_{\mathrm{hist}} = -E(\tilde{h}(\widehat{X})) = \sum_j \tilde{h}(\widehat{X})[j]\log_2 \tilde{h}(\widehat{X})[j]. \tag{9}$$

With $\bar{h}$ defined as the soft histogram of the averaged predicted frames, cross-entropy is calculated as:

$$\mathcal{L}_{\mathrm{ce}} = -\sum_j \tilde{h}(\widehat{X})[j]\log_2 \bar{h}[j]. \tag{10}$$

Chromatic Retinex Decomposition

Instead of directly predicting the signal via neural representation, we propose to separately generate layer-wise representations with a Chromatic Retinex decomposition. The decomposition well decouples the coarse/fine-grained distortions, benefiting the enhancement of different components.

Chromatic Retinex. The traditional Retinex model assumes that the ambient lighting is monochromatic. During image capture, accurate and stable white balance is the key to maintaining this assumption. It adjusts the relative value of RGB channels according to light temperature, making the image appear as if captured under white light. However, as mentioned in Sec. , the white balance in low-light videos is unstable among frames, causing severe color flicker. It destroys the temporal consistency and disenables the *i.i.d.* condition for frame-to-frame training in Eqn. 3).

With a biased white balance, the color of ambient lighting is recorded. Therefore, we propose to extend the original monochromatic illumination layer into a chromatic one:

$$X^c = I^c \otimes R^c, \forall c \in \{r,g,b\}, \tag{11}$$

where X^c is one of RGB channels of the original color image, $\otimes$ means element-wise multiplication, I^c represents the illumination layer with color, and R^c represents the reflectance layer. With unbalanced data d^c and accurate white balance weights w^c, there is:

$$\begin{aligned} & X^c = \sigma(w^c \cdot d^c), \\ \Rightarrow \quad & I^c \otimes R^c = \sigma(w^c \cdot d^c), \\ \Rightarrow \quad & \frac{\sigma^{-1}(I^c \otimes R^c)}{w^c} = d^c, \\ \Rightarrow \quad & \frac{\sigma^{-1}(I^c) \otimes \sigma^{-1}(R^c)}{w^c} = d^c, \end{aligned} \tag{12, 13}$$

where $\sigma(\cdot)$ is the gamma correction which we simplify as $\sigma(x) = x^{(1/2.2)}$. With inaccurate white balance weights $\tilde{w}^c$, we maintain the reflectance layer unchanged:

$$\frac{\sigma^{-1}(\tilde{I}^c) \otimes \sigma^{-1}(R^c)}{\tilde{w}^c} = d^c. \tag{14}$$

We follow the grey-world assumption and the white balance design in the image signal processor (ISP) to set the green channel as the denominator. I^c is formulated as:

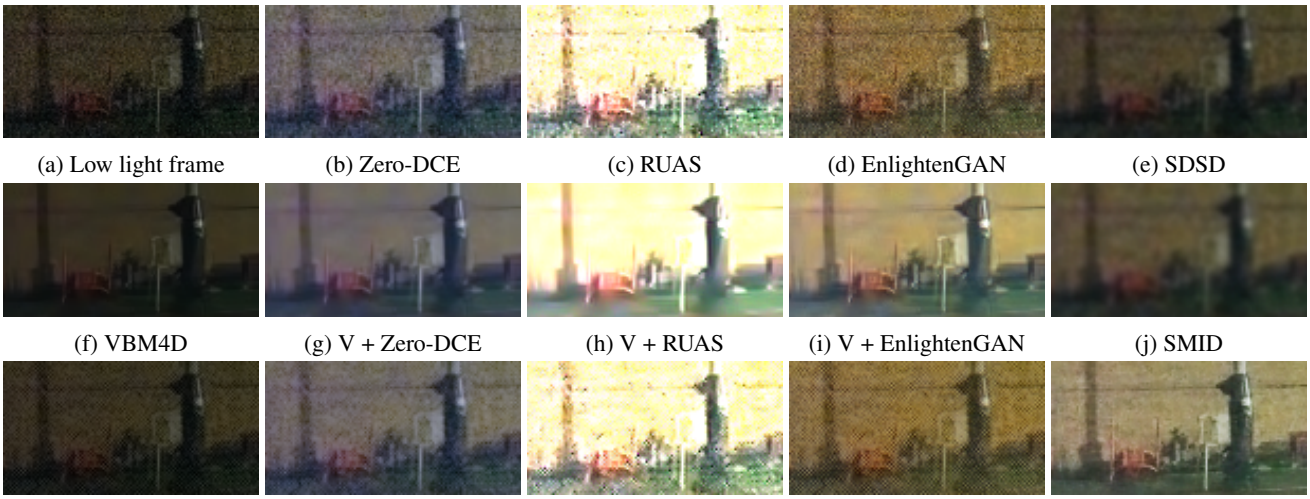
$$I^c = \sigma(\frac{\mathtt{mean}(\sigma^{-1}(X^c))}{\mathtt{mean}(\sigma^{-1}(X^g))})I, \tag{15}$$

where I denotes original illumination layer in Retinex model and `mean` can be calculated locally or globally. In the implementation, we use global mean and optimize I with the loss function:

$$\mathcal{L}_{\mathrm{illum}} = \| \downarrow(\widehat{I}_t) - \downarrow(I_t^o) \|_2^2, \tag{16}$$

where we replace the regularization term in Eqn. (16) with a downsampler $\downarrow(\cdot)$ following Liang et al. (2022).

Combined with this Chromatic Retinex decomposition, the output of hybrid neural representation changes from



(a) Low light frame (b) Zero-DCE (c) RUAS (d) EnlightenGAN (e) SDSD

(f) VBM4D (g) V + Zero-DCE (h) V + RUAS (i) V + EnlightenGAN (j) SMID

(k) UDVD (l) U + Zero-DCE (m) U + RUAS (n) U + EnlightenGAN (o) Ours

Figure 5: Comparison results on the evaluation datasets. V represents VBM4D and U denotes UDVD. **Zoom in for best view**.

Eqn. (2) into 3-channel reflectance and 3-channel illumination layers:

$$f_{\text{dec}}(z_t) = \{\widehat{R}_t, \widehat{I}_t\}, \quad \widehat{X}_t = \widehat{R}_t \otimes \widehat{I}_t. \tag{17}$$

Considering optimization in terms of illumination, the self-regression loss $\mathcal{L}_{\text{self}}$ is defined as:

$$\mathcal{L}_{\text{self}} = \|\widehat{R}_t \otimes \widehat{I}_t - X_t\|_2^2. \tag{18}$$

Because the Chromatic Retinex decouples distortions where color flickers mainly exist in the illumination layer, we can assume the temporal consistency of the reflectance layers. Therefore, the warping loss $\mathcal{L}_{\text{warp}}$ changes from Eqn. (3) into:

$$\mathcal{L}_{\text{warp}} = d\Big(\widehat{R}_t, \texttt{warp}\big(\widehat{R}_{t-1}, o(\widehat{R}_t, \widehat{R}_{t-1})\big)\Big). \tag{19}$$

After Chromatic Retinex decomposition, the noise in the reflectance is naturally suppressed via hybrid neural representation and entropy minimization as discussed before. Then we propose a Channel-wise Gamma Estimation to enhance and calibrate the illumination.

Channel-wise Gamma Estimation. For brightening and white re-balance, a mapping from original illumination to a satisfactory distribution is needed. Specifically, for chromatic illumination $\widehat{I}_t^c$, we predict a channel-wise Gamma curve for brightening and color refinement. For simplicity, we omit frame number t and color channel c in the following. It predicts a channel-wise global parameter to control the curve:

$$\tilde{I} = f_{\text{lit}}(\widehat{I}) = \widehat{I}^{\gamma}, \tag{20}$$

where γ acts as the index for Gamma function, predicted by a convolution network with a similar architecture as (Guo et al. 2020).

Then we obtain the final prediction $\tilde{X}$. Added the enhancement module, the target of entropy constraints on the soft histogram in Eqn. (9) and (10) changes from $\widehat{X}$ into $\tilde{X}$.

Loss Function

The proposed bottleneck neural representation is optimized by the following loss function:

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{illum}} + \lambda_1\mathcal{L}_{\text{self}} + \lambda_2\mathcal{L}_{\text{warp}} \\ + \lambda_3\mathcal{L}_{\text{gmm}} + \lambda_4\mathcal{L}_{\text{hist}} + \lambda_5\mathcal{L}_{\text{ce}}.\end{aligned} \tag{21}$$

Experimental Results

Implementation Details

The training starts with a 300-epochs self-regression then continue with a fully-equipped loss for another 300 epochs. We choose λ_1=100, λ_2=10^{-4}, λ_3=10^{-3}, λ_4=1, λ_5=1. To evaluate the performance of the proposed method, we compare it with 1) *unsupervised low-light image enhancement methods*: MF (Fu et al. 2016), LIME (Guo, Li, and Ling 2017), Zero-DCE (Guo et al. 2020), RUAS (Liu et al. 2021) and EnlightenGAN (Jiang et al. 2021); 2) *supervised low-light video enhancement methods*: SMID (Chen et al. 2019) and SDSD (Wang et al. 2021); 3) *unsupervised denoising methods*: VBM4D (Maggioni et al. 2012) and UDVD (Sheth et al. 2021); 4) *combined unsupervised denoising and low-light methods*. All deep methods use their own pretrained checkpoints. We have attempted to retrain unsupervised methods on the given sequence from scratch but there is no obvious gain. It may be because the scale of training samples is small.

The evaluation dataset is commonly used DRV (Chen et al. 2019) which provides dynamic videos of a real dark scene. After omitting videos captured in the same scene or with too many or few frames, we randomly choose an evaluation set where each video has 100-120 frames. No-reference metrics NIQE (Mittal, Soundararajan, and Bovik 2013), ILNIQE (Zhang, Zhang, and Bovik 2015), and NIQMC (Gu et al. 2017) are chosen as metrics.

Methods	MF	LIME	Zero-DCE	RUAS	EnGAN	SMID	SDSD	VBM4D	UDVD	Ours
NIQE ↓	8.8771	8.7152	5.1637	6.8804	8.9096	11.1896	13.0713	<u>6.8583</u>	13.5302	**4.7616**
ILNIQE ↓	58.4152	41.5132	44.0923	<u>33.0497</u>	37.7286	51.6142	59.5753	49.0038	51.8735	**31.1489**
NIQMC ↑	3.5850	4.5716	3.9014	**4.9217**	4.4574	4.3616	3.9162	2.5386	2.6228	<u>4.8123</u>

Table 1: Quantitative results of different methods. The best scores are bold and the second ones are underlined.

	VBM4D+				UDVD+				
Methods	-	Zero-DCE	RUAS	EnGAN	-	Zero-DCE	RUAS	EnGAN	Ours
NIQE ↓	6.8583	5.5158	5.6597	<u>4.8043</u>	13.5302	7.7535	7.809	12.7212	**4.7616**
ILNIQE ↓	49.0038	42.1871	39.2185	35.3030	51.8735	41.7951	<u>33.5089</u>	41.9427	**31.1489**
NIQMC ↑	2.5386	3.7030	4.6428	4.4398	2.6228	3.8060	<u>4.7444</u>	4.5547	**4.8123**

Table 2: Comparison with cascaded denoising and enhancement. The best scores bold and the second ones are underlined.

Quantitative Evaluation

We provide quantitative results in Table 1 and Table 2. It is clearly observed that, our method achieves significantly superior performance compared with previous conventional, unpaired learning, and self-supervised learning methods. In addition, our method also outperforms the cascaded version of low-light enhancement and denoising methods.

Qualitative Evaluation

We also provide qualitative results in Fig. 5. As shown, most unsupervised low-light methods cannot handle the intensive noise in frames captured in the real low-light scene. On the contrary, amplifying the originally hidden noise hugely affects the subjective quality of the frames. The performance of supervised low-light video methods heavily relies on the quality of the training set and lacks generalization to real dark videos because of the domain gap. Unsupervised denoising methods show relatively promising results but do not consider distortions of color. Besides, because the noise model is hugely biased and hybrid in dark frames, predictions from these methods are blurred.

Furthermore, we attempt to combine unsupervised denoising and unsupervised low-light image enhancement. However, such a cascading does not guarantee a promising performance as well. In fact, pre-denoising may output an over-smoothing result which causes loss of information for the following enhancement. As a result, our proposed method offers more visually promising results.

Ablation Studies

We conduct ablation studies as shown in Table 3.

Hybrid Neural Representation (HNR). Adopting hybrid neural representation or not introduces a huge performance gap. Without this design, the input is replaced by the frame number. During the same training time, there is still an obvious blur in predictions.

Gaussian-Mixture Model (GMM). Without the Gaussian-mixture model as an objective bottleneck, the method easily generates noise which originates from the powerful learning capacity of the neural network.

HNR	GMM	SH	CR	CGE	NIQE↓
✓	✓	✓	✓	✓	4.7616
	✓	✓	✓	✓	8.5161
✓		✓	✓	✓	5.1642
✓	✓		✓	✓	5.6177
✓	✓	✓		✓	4.9132
✓	✓	✓	✓		5.7324

Table 3: Ablation studies on the proposed designs. The meaning of abbreviations can be found in Sec. .

Soft Histogram (SH). We attempt to replace the entropy constraint on the soft histogram with the loss function proposed by (Guo et al. 2020). The enhanced frames show an unsatisfactory restoration of color.

Chromatic Retinex (CR). Replacing Chromatic Retinex with a traditional one, the color bias is introduced in the reflectance layer. The temporal consistency of reflectance layers can not be guaranteed because of color bias.

Channel-wise Gamma Estimation (CGE). Instead of estimating a channel-wise Gamma function, we attempt to use a pixel-wise or global Gamma function. The former shows unnatural illumination.

Conclusion

In this paper, we develop a self-learned enhancement approach that gets rid of the reliance on external data. We adopt a bottleneck neural representation mechanism to squeeze out only the high-quality signals. Compact deep embeddings are used to describe frame-wise information, which forms a consistent manifold. An entropy constraint is then applied to use spatial-temporal context to filter out degraded visual signals such as noise. At last, a novel Chromatic Retinex decomposition is built for effective temporal alignment, which facilitates self-supervised learning. Comprehensive experiments demonstrate our method's effectiveness and robustness in both spatial and temporal qualities.

References

Abdullah-Al-Wadud, M.; Kabir, M.; Dewan, M.; and Chae, O. 2007. A Dynamic Histogram Equalization for Image Contrast Enhancement. *IEEE Trans. Consum. Electron.*, 53(2): 593–600.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Trans. Image Process.*, 27(4): 2049–2062.

Chen, C.; Chen, Q.; Do, M. N.; and Koltun, V. 2019. Seeing Motion in the Dark. In *Proc. Int'l Conf. Comput. Vision*.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to See in the Dark. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Chen, H.; Gwilliam, M.; Lim, S.-N.; and Shrivastava, A. 2023. HNeRV: A Hybrid Neural Representation for Videos. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Chen, H.; He, B.; Wang, H.; Ren, Y.; Lim, S.-N.; and Shrivastava, A. 2021. NeRV: Neural Representations for Videos. In *Proc. Annu. Conf. Neural Inf. Process. Systems*.

Chen, Z.; and Zhang, H. 2019. Learning Implicit Fields for Generative Shape Modeling. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Chibane, J.; Alldieck, T.; and Pons-Moll, G. 2020. Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Ehret, T.; Davy, A.; Morel, J.-M.; Facciolo, G.; and Arias, P. 2019. Model-blind Video Denoising Via Frame-to-frame Training. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Fu, X.; Zeng, D.; Huang, Y.; Liao, Y.; Ding, X.; and Paisley, J. 2016. A Fusion-Based Enhancing Method for Weakly Illuminated Images. *Signal Process.*, 129: 82–96.

Gu, K.; Lin, W.; Zhai, G.; Yang, X.; Zhang, W.; and Chen, C. W. 2017. No-Reference Quality Metric of Contrast-Distorted Images Based on Information Maximization. *IEEE Trans. Cybern.*, 47(12): 4559–4565.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Guo, X.; Li, Y.; and Ling, H. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.*, 26(2): 982–993.

Huang, H.; Yang, W.; Hu, Y.; Liu, J.; and Duan, L.-Y. 2022. Towards Low Light Enhancement with RAW Images. *IEEE Trans. Image Process.*, 31: 1391–1405.

Huang, T.; Li, S.; Jia, X.; Lu, H.; and Liu, J. 2021. Neighbor2Neighbor: Self-supervised Denoising from Single Noisy Images. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Jiang, H.; and Zheng, Y. 2019. Learning to See Moving Objects in the Dark. In *Proc. Int'l Conf. Comput. Vision*.

Jiang, K.; Wang, Z.; Wang, Z.; Chen, C.; Yi, P.; Lu, T.; and Lin, C. 2022. Degrade Is Upgrade: Learning Degradation for Low-Light Image Enhancement. In *Proc. AAAI Conf. on Artif. Intell.*

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. EnlightenGAN: Deep Light Enhancement without Paired Supervision. *IEEE Trans. Image Process.*, 30: 2340–2349.

Krull, A.; Buchholz, T.-O.; and Jug, F. 2019. Noise2Void-Learning Denoising from Single Noisy Images. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning Image Restoration without Clean Data. In *Proc. Int'l Conf. Mach. Learn.*

Li, C.; Guo, C.; Han, L.; Jiang, J.; Cheng, M.-M.; Gu, J.; and Loy, C. C. 2022. Low-Light Image and Video Enhancement Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 9396–9416.

Liang, J.; Xu, Y.; Quan, Y.; Shi, B.; and Ji, H. 2022. Self-Supervised Low-Light Image Enhancement Using Discrepant Untrained Network Priors. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11): 7332–7345.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward Fast, Flexible, and Robust Low-light Image Enhancement. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Maggioni, M.; Boracchi, G.; Foi, A.; and Egiazarian, K. 2012. Video Denoising, Deblocking, and Enhancement Through Separable 4-D Nonlocal Spatiotemporal Transforms. *IEEE Trans. Image Process.*, 21(9): 3952–3966.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3D reconstruction in function space. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. NeRF in the Dark: High Dynamic Range View Synthesis from Noisy Raw Images. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. Eur. Conf. Comput. Vision*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Process. Lett.*, 20(3): 209–212.

Monakhova, K.; Richter, S. R.; Waller, L.; and Koltun, V. 2022. Dancing Under the Stars: Video Denoising in Starlight. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional Occupancy Networks. In *Proc. Eur. Conf. Comput. Vision*.

Ren, Y.; Ying, Z.; Li, T. H.; and Li, G. 2019. LECARM: Low-Light Image Enhancement Using the Camera Response Model. *IEEE Trans. Circuits Syst. Video Technol.*, 29(4): 968–981.

Sánchez Púrez, J.; Meinhardt-Llopis, E.; and Facciolo, G. 2013. TV-L1 Optical Flow Estimation. *Image Processing On Line*, 3: 137–150.

Sheth, D. Y.; Mohan, S.; Vincent, J. L.; Manzorro, R.; Crozier, P. A.; Khapra, M. M.; Simoncelli, E. P.; and Fernandez-Granda, C. 2021. Unsupervised Deep Video Denoising. In *Proc. Int'l Conf. Comput. Vision*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep Image Prior. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Wang, R.; Xu, X.; Chi-Wing Fu, J. L.; Yu, B.; and Jia, J. 2021. Seeing Dynamic Scene in the Dark: High-Quality Video Dataset with Mechatronic Alignment. In *Proc. Int'l Conf. Comput. Vision*.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019a. Underexposed Photo Enhancement Using Deep Illumination Estimation. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Wang, Y.; Cao, Y.; Zha, Z.-J.; Zhang, J.; Xiong, Z.; Zhang, W.; and Wu, F. 2019b. Progressive Retinex: Mutually Reinforced Illumination-Noise Perception Network for Low Light Image Enhancement. In *Proc. ACM Int'l Conf. Multimedia*.

Wang, Y.; Wan, R.; Yang, W.; Li, H.; Chau, L.-P.; and Kot, A. C. 2022. Low-Light Image Enhancement with Normalizing Flow. In *Proc. AAAI Conf. on Artif. Intell.*

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *Proc. Brit. Mach. Vision Conf.*

Wei, K.; Fu, Y.; Zheng, Y.; and Yang, J. 2022. Physics-Based Noise Modeling for Extreme Low-Light Photography. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11): 8520–8537.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *Proc. IEEE/CVF Int'l Conf. Comput. Vision and Pattern Recognit.*

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.*, 24(8): 2579–2591.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond Brightening Low-Light Images. *Int'l J. of Comput. Vision*, 129(2): 1013–1037.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the Darkness: A Practical Low-Light Image Enhancer. In *Proc. ACM Int'l Conf. Multimedia*.