SNR-Aware Low-light Image Enhancement

Xiaogang Xu[1] Ruixing Wang[1,2] Chi-Wing Fu[1] Jiaya Jia[1]
[1] The Chinese University of Hong Kong [2] SmartMore
{xgxu, rxwang, cwfu, leojia}@cse.cuhk.edu.hk

Abstract

This paper presents a new solution for low-light image enhancement by collectively exploiting Signal-to-Noise-Ratio-aware transformers and convolutional models to dynamically enhance pixels with spatial-varying operations. They are long-range operations for image regions of extremely low Signal-to-Noise-Ratio (SNR) and short-range operations for other regions. We propose to take an SNR prior to guide the feature fusion and formulate the SNR-aware transformer with a new self-attention model to avoid tokens from noisy image regions of very low SNR. Extensive experiments show that our framework consistently achieves better performance than SOTA approaches on seven representative benchmarks with the same structure. Also, we conducted a large-scale user study with 100 participants to verify the superior perceptual quality of our results. The code is available at https://github.com/dvlab-research/SNR-Aware-Low-Light-Enhance.

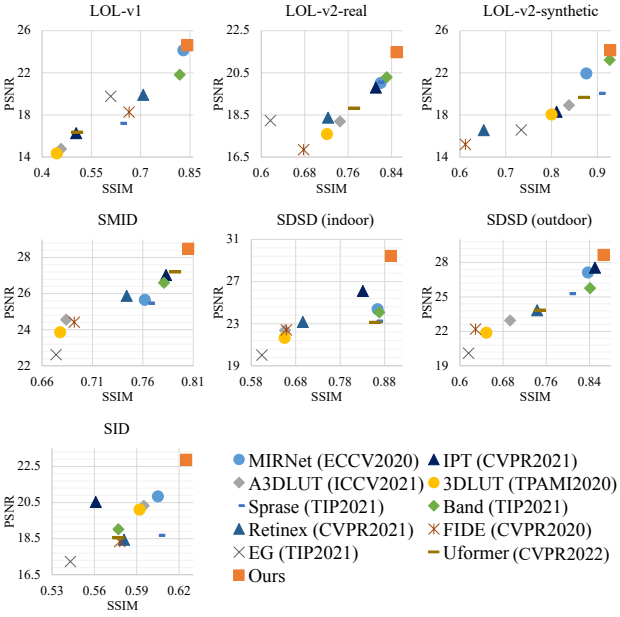


Figure 1. Our approach consistently achieves better performance in terms of PSNR/SSIM over 10 SOTA methods on 7 different benchmarks with the same network structure. Each plot is for comparison on one benchmark dataset.

1. Introduction

Low-light imaging is critical for many tasks, such as object and action recognition at night [18, 27]. Low-light images are generally with poor visibility for human perception. Similarly, downstream vision tasks can be affected when taking low-light images directly as the input.

Several methods have been proposed to enhance low-light images. The de facto approach nowadays is to develop neural networks that learn to manipulate color, tone, and contrast to enhance low-light images [12, 15, 41, 56], while some recent works account for noise in images [29, 48]. In this paper, our key insight is that different regions in a low-light image can have different characteristics of lightness, noise, visibility, etc. Regions of extremely low lightness are heavily corrupted by noise, while other regions in the same image can still have reasonable visibility and contrast. For better overall image enhancement, we should adaptively consider different regions in the low-light images.

To this end, we study the relation between signal and noise in image space by exploring Signal-to-Noise-Ratio (SNR) [3, 54] for achieving spatial-varying enhancement. In particular, regions of lower SNR are typically unclear. So we exploit non-local image information in longer spatial range for image enhancement. On the other hand, regions of relatively higher SNR typically have higher visibility and less noise. Thus local image information is typically sufficient. Fig. 2 shows a low-light image example for illustration. Further discussion is presented in Sec. 3.1.

Our solution to low-light image enhancement in the RGB domain is to collectively exploit long- and short-range operations. In the deepest hidden layer, we design two branches. The long-range branch with a transformer structure [38] is to capture non-local information and the short-range one with convolutional residual blocks [17] captures local information. When enhancing each pixel, we determine the contribution of local (short-range) and non-local (long-range) information dynamically based on the

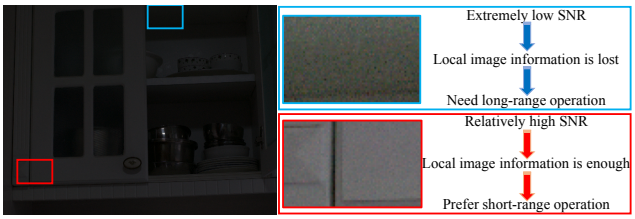


Figure 2. Low-light image enhancement requires spatial-varying operations. The blue (or red) region has extremely low (or relatively high) SNR. It offers inadequate (or sufficient) local image information for image enhancement. In operations, we use long-range image information for the blue region, since it has been heavily corrupted by noise. We linearly scale up brightness on the right for visualizing noise in different image regions.

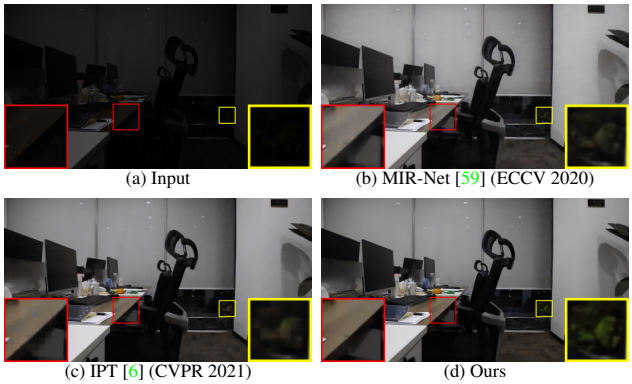
pixel's signal-to-noise ratio. Hence, in regions of high SNR, local information plays a vital role during the enhancement, whereas in regions of very low SNR, non-local messages are effective. To achieve such spatial-varying operations, we construct an SNR prior and use it to guide feature fusion. Also, we modify the attention mechanism in the transformer structure and propose the SNR-aware transformer. Unlike existing transformer structure, not all tokens contribute to the attention computation. We consider only the tokens with sufficient SNR values to avoid noise influence from the very-low-SNR regions.

Our framework effectively enhances low-light images of dynamic noise levels. Extensive experiments were conducted on 7 representative datasets: LOL (v1 [45], v2-real [53], & v2-synthetic [53]), SID [5], SMID [4], and SDSD (indoor & outdoor) [39]. As shown in Fig. 1, our framework outperforms 10 SOTA approaches on all datasets with the same structure. Further, we conduct large-scale user study with 100 participants to verify the effectiveness of our approach. Qualitative comparison is presented in Fig. 3. Overall, our contribution is threefold.

- We propose a new signal-to-noise-aware framework that simultaneously adopts a transformer structure and a convolutional model for achieving spatial-varying low-light image enhancement with an SNR prior.
- We design an SNR-aware transformer with a new self-attention module for low-light image enhancement.
- We conducted extensive experiments on seven representative datasets, manifesting that our framework consistently outperforms a rich set of SOTA methods.

2. Related Work

No-learning-based Low-light Image Enhancement. To enhance low-light images, histogram equalization and gamma correction are fundamental tools to expand the dynamic range and increase the image contrast. These



(a) Input (b) MIR-Net [59] (ECCV 2020) (c) IPT [6] (CVPR 2021) (d) Ours

Figure 3. A challenging low-light frame (a) enhanced by a SOTA method with a convolutional structure (b), a SOTA transformer structure (c), and our method (d). Ours exhibits clearer details, more distinct contrast, and less noise (best view by zoom-in).

primary approaches tend to produce undesirable artifacts in the enhanced real-world images. Retinex-based methods, which treat the reflectance component as plausible approximation for image enhancement, are able to produce more realistic and natural results [28, 35]. However, when enhancing complicated real-world images, this line of methods often distort colors locally [40].

Learning-based Low-light Image Enhancement. Many learning-based low-light image enhancement approaches were proposed in recent years [2, 14, 20, 22, 29, 31, 42, 48–50, 52, 53, 59, 60, 62, 63]. Wang *et al*. [40] proposed to predict the illumination map for enhancing underexposed photos. Sean *et al*. [33] designed a strategy to learn three different types of spatially local filters for enhancement. Yang *et al*. [51] presented a semi-supervised approach to recover a linear-band representation of the low-light image. Also, there are unsupervised methods [7, 14, 19]. For instance, Guo *et al*. [14] built a lightweight network to estimate pixel-wise and high-order curves for dynamic range adjustment.

Unlike previous work, our new approach enhances low-light image based on a signal-to-noise-aware framework consisting of a new SNR-aware transformer design and a convolutional model to adaptively enhance low-light images in a spatial-varying manner. As shown in Fig. 1, our framework achieves better performance consistently on seven different benchmarks with the same structure.

3. Our Method

Fig. 4 shows the overview of our framework. The input is a low-light image, from which we first obtain an SNR map using a simple and yet effective strategy (see Sec. 3.2 for details). We propose to take SNR to guide our framework to learn different enhancement operations adaptively for image regions of varying signal-to-noise ratios.

In the deepest hidden layer of our framework, we design

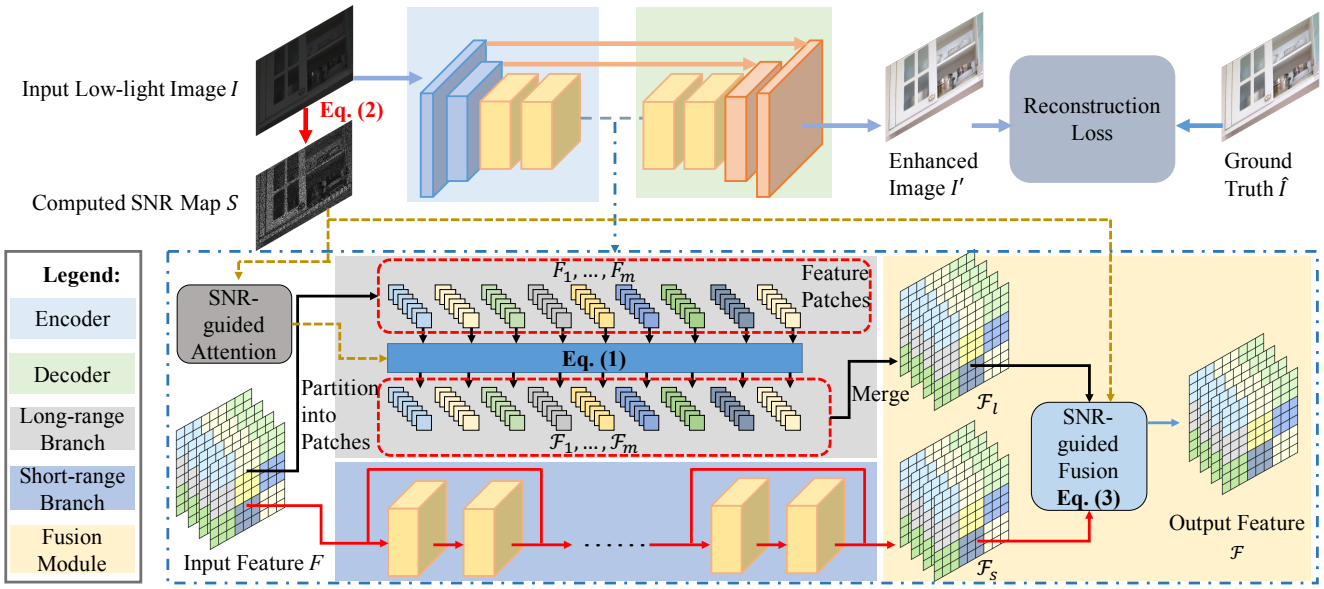


Figure 4. Our framework for low-light image enhancement starts by estimating an SNR map for guiding pixel enhancement in different image regions. We formulate an SNR-guided attention (Fig. 5) to guide how our patch-wise SNR-aware transformer processes long-range image information, particularly for enhancing image regions of very low SNR. Further, we develop the SNR-guided fusion to combine resulting long-range feature $\mathcal{F}_l$ with short-range feature $\mathcal{F}_s$ to produce the final image feature $\mathcal{F}$.

two different branches for long- and short-range. They are specifically formulated for achieving efficient operations, implemented by transformer [38] and convolution structures, respectively. To achieve the long-range operations while avoiding the influence of noise in extremely low-light regions, we guide the attention mechanism in transformer with an SNR map. To adopt different operations, we develop an SNR-based fusion strategy to obtain a combined representation from the long- and short-range features. Also, we use skip connections from the encoder to decoder to enhance image details.

3.1. Long- and Short-range Branches

Necessity of spatial-varying operations. Traditional networks for low-light image enhancement adopt convolution structures in the deepest hidden layer. These operations have short-range to capture local information mostly. Local information may suffice to recover image regions that are not extremely dark, since these pixels still contain an amount of visible image content (or signals). But for extremely dark regions, local information is inadequate for enhancing pixels, since adjacent local regions are also weak in terms of visibility and are mostly dominated by noise.

To address this key issue, we dynamically enhance pixels in different regions with varying local and non-local communication. Local and non-local information is complementary. The effect can be determined based on the SNR distribution over the image. On the one hand, for image regions of high SNR, local information should play the major role, since local information is adequate for enhancement. It is generally more accurate than long-distance non-local information.

On the other hand, for image regions of very low SNR, we pay more attention to non-local information, since local regions likely have very little image information while being dominated by noise. Unlike previous methods, we explicitly formulate a long-range branch for image regions of very low SNR and a short-range branch for other regions in the deepest hidden layer of our framework (see Fig. 4).

Implementation of two branches. The short-range branch is implemented based on the structure of convolutional residual blocks for capturing local information, whereas the long-range branch is implemented based on the structure of transformer [38] since transformers are good at capturing long-range dependency via the global self-attention mechanism, as demonstrated in many high-level [10, 16, 21, 30, 46, 57, 58] and low-level tasks [6, 44].

In the long-range branch, we first partition feature map F (extracted by the encoder from input image $I \in \mathbb{R}^{H\times W\times 3}$) into m feature patches, *i.e.*, $F_i \in \mathbb{R}^{p\times p\times C}, i = \{1, ..., m\}$. Suppose the size of feature map F is $h \times w \times C$ and the patch size is $p \times p$. There are $m = \frac{h}{p} \times \frac{w}{p}$ feature patches for covering the entire feature map.

As shown in Fig. 4, our SNR-aware transformer is patch-based. It consists of the multi-head self-attention (MSA) modules [38] and the feed-forward networks (FFN) [38], both composed of two fully connected layers. The output features $\mathcal{F}_1, ..., \mathcal{F}_m$ from the transformer have the same size as the input feature patches. We flatten $F_1, ..., F_m$ into 1D

features and perform the computation of

$$\begin{aligned} y_0 &= [F_1, F_2, ..., F_m], \\ q_i &= k_i = v_i = LN(y_{i-1}), \\ \widehat{y}_i &= MSA(q_i, k_i, v_i) + y_{i-1}, \\ y_i &= FFN(LN(\widehat{y}_i)) + \widehat{y}_i, \\ \text{and } [\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_m] &= y_l, i = \{1, ..., l\}, \end{aligned} \tag{1}$$

where LN denotes layer normalization; y_i denotes the output of the i-th transformer block; MSA denotes our SNR-aware multi-head self-attention module (see Fig. 5), which will be detailed in Sec. 3.3; q_i, k_i, and v_i denote the query, key, and value vectors, respectively, in the i-th multi-head self-attention module; and l denotes the number of layers in the transformer. The transformed features $\mathcal{F}_1, ..., \mathcal{F}_m$ can be merged to form the 2D feature map $\mathcal{F}_l$ (see Figure 4).

3.2. SNR-based Spatially-varying Feature Fusion

The SNR map. As shown in Fig. 4, our framework starts by estimating an SNR map. It is difficult and tedious to estimate the amount of noise in input image I and prepare a clean version of I to determine the SNR value of each pixel, given only a single input image. Similar to previous no-learning-based denoising approaches [1,8], we treat noise as discontinuous transition between adjacent pixels in spatial domain. The noise component can be modeled as the distance between the noisy image and an associated clean image. In this work, we use it to estimate the SNR map of I and make it an effective prior for our spatial-varying feature fusion. Given $I \in \mathbb{R}^{H\times W\times 3}$, we first compute the gray-scale of I, *i.e.*, $I_g \in \mathbb{R}^{H\times W}$, followed by computing the SNR map $S \in \mathbb{R}^{H\times W}$ as

$$\widehat{I}_g = denoise(I_g), \quad N = abs(I_g - \widehat{I}_g), \quad S = \widehat{I}_g / N, \tag{2}$$

where $denoise$ is a no-learning-based denoising operation (with experiments in Sections 4.2 and 4.3), *e.g.*, averaging local groups of pixels; abs denotes the absolute value; and $N \in \mathbb{R}^{H\times W}$ is the estimated noise map. Although the resulting SNR values are approximates given the extracted noise not accurate, our framework with such SNR map is still effective as verified by our extensive experiments.

Spatial-varying feature fusion with the SNR map. As shown in Fig. 4, we employ encoder $\mathcal{E}$ to extract feature F from input image I. The feature is then processed separately by the long- and short-range branches, which produce long-range feature $\mathcal{F}_l \in \mathbb{R}^{h\times w\times C}$ and short-range feature $\mathcal{F}_s \in \mathbb{R}^{h\times w\times C}$. To adaptively combine the two features, we resize the SNR map into $h \times w$, normalize its values to range $[0, 1]$, and take the normalized SNR map S' as interpolation weights to fuse $\mathcal{F}_l$ and $\mathcal{F}_s$ as

$$\mathcal{F} = \mathcal{F}_s \times S' + \mathcal{F}_l \times (1 - S'), \tag{3}$$

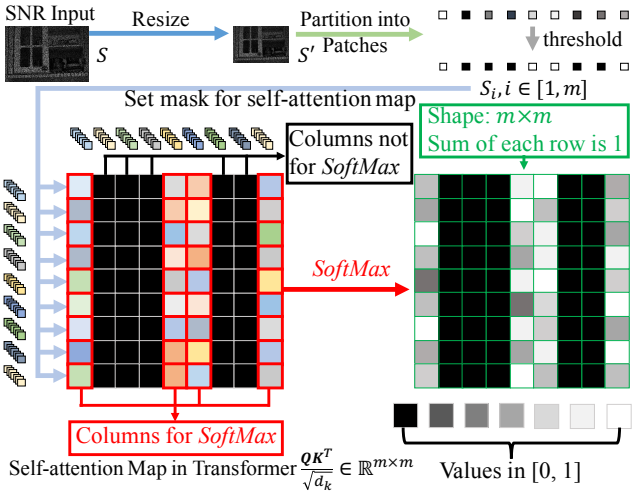


Figure 5. Illustration: SNR-guided attention in transformer. Black squares are elements ignored by SoftMax; colored squares reveal the similarity between feature tokens. They are used in SoftMax.

where $\mathcal{F} \in \mathbb{R}^{h\times w\times C}$ is the output feature to be passed to the decoder for producing the final output image. Since values in the SNR map dynamically reveal the level of noise in different regions of the input image, the fusion can adaptively combine local (short-range) and non-local (long-range) image information for production of $\mathcal{F}$.

3.3. SNR-guided Attention in Transformer

Limitation of traditional transformer structures. Although traditional transformers can capture non-local information for enhancing images, they have critical issues. In original structures, the attention is computed among all patches. To enhance a pixel, the long-range attention may come from any image region, regardless of the signal and noise levels. In fact, regions of very low SNR are dominated by noise. Thus, their information is inaccurate, severely disrupting image enhancement. Here, we propose SNR-guided attention to improve transformer in this special task.

SNR-aware transformer. Fig. 5 shows our SNR-aware transformer with the new self-attention module. Given input image $I \in \mathbb{R}^{H\times W\times 3}$ and associated SNR map $S \in \mathbb{R}^{H\times W}$, we first resize S to $S' \in \mathbb{R}^{h\times w}$ to match the size of feature map F. We then partition S' into m patches following the way we partition F into patches, and compute the average value in each patch, *i.e.*, $S_i \in \mathbb{R}^1, i = \{1, ..., m\}$. We pack these values into vector $\mathcal{S} \in \mathbb{R}^m$. It acts as a mask in the attention computation of the transformer, which can avoid the message propagation from image regions of extremely low SNR (see Fig. 5) in the transformer. The mask value of the i-th element of $\mathcal{S}$ is

expressed as

$$\mathcal{S}_i = \begin{cases} 0, & S_i < s \\ 1, & S_i \geq s \end{cases}, i = \{1, ..., m\}, \quad (4)$$

where s is a threshold value. Next, we stack m copies of $\mathcal{S}$ to form matrix $\mathcal{S}' \in \mathbb{R}^{m \times m}$. Suppose the head number is B for the multi-head self-attention (MSA) module (Eq. (1)), the b-th head self-attention calculation $Attention_{i,b}$ in the transformer's i-th layer is formulated as

$$\mathbf{Q}_{i,b} = q_i W_b^q, \mathbf{K}_{i,b} = k_i W_b^k, \mathbf{V}_{i,b} = v_i W_b^v, \quad \text{and} \quad (5)$$

$$Attention_{i,b}(\mathbf{Q}_{i,b}, \mathbf{K}_{i,b}, \mathbf{V}_{i,b}) = \text{Softmax}(\frac{\mathbf{Q}_{i,b}\mathbf{K}_{i,b}^T}{\sqrt{d_b}} + (1 - \mathcal{S}')\sigma)\mathbf{V}_{i,b}, \quad (6)$$

where q_i, k_i, $v_i \in \mathbb{R}^{m \times (p \times p \times C)}$ are input 2D features in Eq. (1); W_k^q, W_k^k, $W_k^v \in \mathbb{R}^{(p \times p \times C) \times C_k}$ represent the projection matrices for the k-th head; $\mathbf{Q}_{i,b}$, $\mathbf{K}_{i,b}$, $\mathbf{V}_{i,b} \in \mathbb{R}^{m \times C_k}$ are the query, key, and value features, respectively, in attention computation.

The output shape of Softmax() and $Attention_{i,b}()$ is $m \times m$ and $m \times C_k$, respectively, where C_k is the channel number in the self-attention computation. Also, $\sqrt{d_k}$ is for normalization and σ is a small negative scalar $-1e9$. The output of all B heads are concatenated. All values are linearly projected to produce the final output of MSA in the transformer's i-th layer. Thus, we ensure that long-range attentions are from the image regions with sufficient SNR.

3.4. Loss Function

Data flow. Given the input image I, we first apply an encoder with convolutional layers to extract feature F. Each stage in the encoder contains a stack of the convolutional layer and LeakyReLU [47]. Residual convolutional blocks are used after the encoder. Then, we forward F to the long- and short-range branches to produce features $\mathcal{F}_l$ and $\mathcal{F}_s$. Lastly, we fuse $\mathcal{F}_l$ and $\mathcal{F}_s$ into $\mathcal{F}$ and use the decoder (symmetrical to the encoder) to transfer $\mathcal{F}$ into the residual R. The final output image I' is $I' = I + R$.

Loss terms. There are two reconstruction loss terms to train our framework, *i.e.*, the Charbonnier loss [25] and the perceptual loss. The Charbonnier loss is written as

$$L_r = \sqrt{\|I' - \widehat{I}\|_2 + \epsilon^2}, \quad (7)$$

where $\widehat{I}$ is the ground truth and ϵ is set as 10^{-3} in all experiments. The perceptual loss compares the VGG feature distances between $\widehat{I}$ and I' using an L_1 loss as

$$L_{vgg} = \|\Phi(I') - \Phi(\widehat{I})\|_1, \quad (8)$$

where $\Phi()$ is the operation of extracting features from the VGG network [37]. The overall loss function is

$$L = L_r + \lambda L_{vgg}, \quad (9)$$

where λ is a hyper parameter.

4. Experiments

4.1. Datasets and Implementation Details

We evaluate our framework on several datasets with noise observable in low-light image regions. They are LOL (v1 & v2) [45, 53], SID [5], SMID [4], and SDSD [39].

LOL has noticeable noise in both v1 and v2 versions. LOL-v1 [45] contains 485 pairs of low-/normal-light images for training and 15 pairs for testing. Each pair includes a low-light input image and an associated well-exposed reference image. LOL-v2 [53] is divided into LOL-v2-real and LOL-v2-synthetic. LOL-v2-real contains 689 low-/normal-light image pairs for training and 100 pairs for testing. Most low-light images were collected by changing the exposure time and ISO with other camera parameters fixed. LOL-v2-synthetic was created by analyzing the illumination distribution in the RAW format.

For SID and SMID, each input sample is a pair of short- and long-exposure images. Both SID and SMID have heavy noise, as the low-light images were captured in extreme dark environments. For SID, we use the subset captured by the Sony camera and follow the script provided by SID to transfer the low-light images from RAW to RGB using rawpy's default ISP. For SMID, we use its full images and also transfer the RAW data to RGB, since our work explores low-light image enhancement in the RGB domain. We set the training and testing split according to that of [4].

Lastly, we adopt the SDSD dataset [39] (the static version) for evaluation. It contains an indoor subset and an outdoor subset, both providing low- and normal-light pairs.

We implement our framework in PyTorch [34], and train and test it on a PC with a 2080Ti GPU. We train our method from scratch with network parameters randomly initialized with Gaussian distribution and adopt standard augmentation, *e.g.*, vertical and horizontal flips. The encoder of our framework has three convolution layers (i.e., strides 1, 2, and 2) with one residual block after the encoder. The decoder is symmetric to the encoder with the upsampling mechanism implemented using the pixel shuffle layer [36]. For the loss minimization, we adopt the Adam [23] optimizer with momentum set to 0.9.

4.2. Comparison with Current Methods

We compare our method with a rich collection of SOTA methods for low-light image enhancement, including Dong [9], LIME [15], MF [11], SRIE [12], BIMEF [55], DRD [45], RRM [28], SID [5], DeepUPE [40], KIND [61], DeepLPF [33], FIDE [48], LPNet [26], MIR-Net [59], RF [24], 3DLUT [60], A3DLUT [42], Band [52], EG [20], Retinex [29], and Sparse [53]. Also, we compared our framework with two recent transformer structures for low-level tasks, *i.e.*, IPT [6] and Uformer [44].

Quantitative analysis. We adopt Peak Signal-to-Noise

Figure 6. Visual comparison on LOLv1, LOL-v2-real, and LOL-v2-synthetic (top to bottom). Our method yields less noise and higher visibility.

Methods	Dong [9]	LIME [15]	MF [11]	SRIE [12]	BIMEF [55]	DRD [45]	RRM [28]	SID [5]	DeepUPE [40]	KIND [61]	DeepLPF [33]	FIDE [48]
PSNR	16.72	16.76	18.79	11.86	13.86	16.77	13.88	14.35	14.38	20.87	15.28	18.27
SSIM	0.580	0.560	0.640	0.500	0.580	0.560	0.660	0.436	0.446	0.800	0.473	0.665
Methods	LPNet [26]	MIR-Net [59]	RF [24]	3DLUT [60]	A3DLUT [42]	Band [52]	EG [20]	Retinex [29]	Sparse [53]	IPT [6]	Uformer [44]	Ours
PSNR	21.46	24.14	15.23	14.35	14.77	20.13	17.48	18.23	17.20	16.27	16.36	**24.61**
SSIM	0.802	0.830	0.452	0.445	0.458	0.830	0.650	0.720	0.640	0.504	0.507	**0.842**

Table 1. Quantitative comparison on LOL-v1.

Methods	Dong [9]	LIME [15]	MF [11]	SRIE [12]	BIMEF [55]	DRD [45]	RRM [28]	SID [5]	DeepUPE [40]	KIND [61]	DeepLPF [33]	FIDE [48]
PSNR	17.26	15.24	18.73	17.34	17.85	15.47	17.34	13.24	13.27	14.74	14.10	16.85
SSIM	0.527	0.470	0.559	0.686	0.653	0.567	0.686	0.442	0.452	0.641	0.480	0.678
Methods	LPNet [26]	MIR-Net [59]	RF [24]	3DLUT [60]	A3DLUT [42]	Band [52]	EG [20]	Retinex [29]	Sparse [53]	IPT [6]	Uformer [44]	Ours
PSNR	17.80	20.02	14.05	17.59	18.19	20.29	18.23	18.37	20.06	19.80	18.82	**21.48**
SSIM	0.792	0.820	0.458	0.721	0.745	0.831	0.617	0.723	0.816	0.813	0.771	**0.849**

Table 2. Quantitative comparison on LOL-v2-real.

Methods	Dong [9]	LIME [15]	MF [11]	SRIE [12]	BIMEF [55]	DRD [45]	RRM [28]	SID [5]	DeepUPE [40]	KIND [61]	DeepLPF [33]	FIDE [48]
PSNR	16.90	16.88	17.50	14.50	17.20	17.13	17.15	15.04	15.08	13.29	16.02	15.20
SSIM	0.749	0.776	0.751	0.616	0.713	0.798	0.727	0.610	0.623	0.578	0.587	0.612
Methods	LPNet [26]	MIR-Net [59]	RF [24]	3DLUT [60]	A3DLUT [42]	Band [52]	EG [20]	Retinex [29]	Sparse [53]	IPT [6]	Uformer [44]	Ours
PSNR	19.51	21.94	15.97	18.04	18.92	23.22	16.57	16.55	22.05	18.30	19.66	**24.14**
SSIM	0.846	0.876	0.632	0.800	0.838	0.927	0.734	0.652	0.905	0.811	0.871	**0.928**

Table 3. Quantitative comparison on LOL-v2-synthetic.

Ratio (PSNR) and Structural Similarity Index (SSIM) [43] for evaluation. In general, a higher SSIM means more high-frequency details and structures in results. Tables 1-3 show the comparisons on LOL-v1, LOL-v2-real, and LOL-v2-synthetic. Our method surpasses all the baselines. Note that we obtain these numbers either from the respective papers or by running the respective public code. Out method (24.61/0.842) also outperforms those of [22] (22.81/0.827) and [62] (21.71/0.834) on LOL-v1. Table 4 compares methods on SID, SMID, SDSD-indoor, and SDSD-outdoor. Ours yields the best performance.

Qualitative analysis. First, we present visual samples in Fig. 6 (top row) for comparing our method with baselines that achieve the best performance (in terms of PSNR) on LOL-v1. Our result shows better visual quality with higher contrast, more precise details, color consistency, and better brightness. Fig. 6 also shows visual comparison on LOL-v2-real and LOL-v2-synthetic. While the original images in these datasets have apparent noise and weak illumination, our method can still produce more realistic results. Also, in regions with complex textures, our output exhibits less visual artifacts.

Fig. 7 (top row) shows visual comparison on SID, demonstrating that our method can effectively deal with

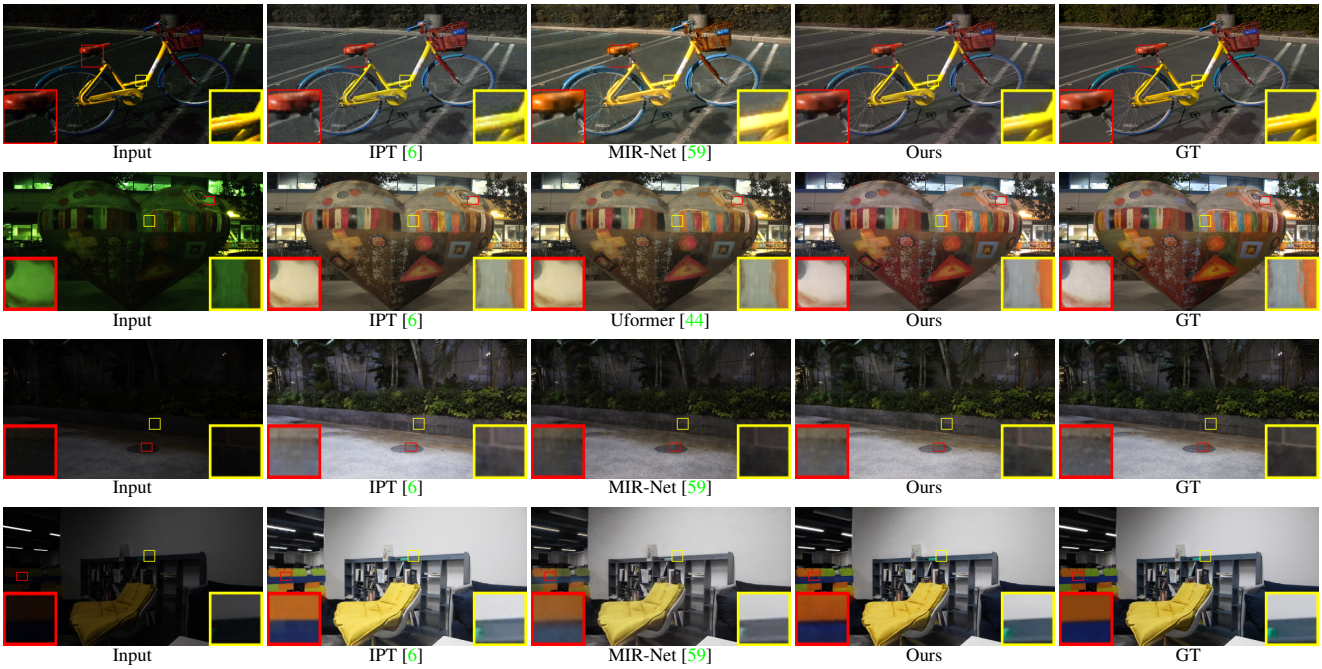


Figure 7. Qualitative comparison on SID (top row), SMID (2nd row), SDSD-indoor (3rd row) and SDSD-outdoor (4th row).

	SID		SMID		SDSD-indoor		SDSD-outdoor	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DRD [45]	16.48	0.578	22.83	0.684	20.84	0.617	20.96	0.629
SID [5]	16.97	0.591	24.78	0.718	23.29	0.703	24.90	0.693
DeepUPE [40]	17.01	0.604	23.91	0.690	21.70	0.662	21.94	0.698
KIND [61]	18.02	0.583	22.18	0.634	21.95	0.672	21.97	0.654
DeepLPF [33]	18.07	0.600	24.36	0.688	22.21	0.664	22.76	0.658
FIDE [48]	18.34	0.578	24.42	0.692	22.41	0.659	22.20	0.629
LPNet [26]	20.08	0.598	26.55	0.772	23.87	0.841	22.09	0.629
MIR-Net [59]	20.84	0.605	25.66	0.762	24.38	0.864	27.13	0.837
RF [24]	16.44	0.596	23.11	0.681	20.97	0.655	21.21	0.689
3DLUT [60]	20.11	0.592	23.86	0.678	21.66	0.655	21.89	0.649
A3DLUT [42]	20.32	0.595	24.56	0.684	22.39	0.656	22.95	0.692
Band [52]	19.02	0.577	26.60	0.781	24.08	0.868	25.77	0.841
EG [20]	17.23	0.543	22.62	0.674	20.02	0.604	20.10	0.616
Retinex [29]	18.44	0.581	25.88	0.744	23.17	0.696	23.84	0.743
Sparse [53]	18.68	0.606	25.48	0.766	23.25	0.863	25.28	0.804
IPT [6]	20.53	0.561	27.03	0.783	26.11	0.831	27.55	0.850
Uformer [44]	18.54	0.577	27.20	0.792	23.17	0.859	23.85	0.748
Ours	**22.87**	**0.625**	**28.49**	**0.805**	**29.44**	**0.894**	**28.66**	**0.866**

Table 4. Quantitative comparison on SID, SMID, SDSD-indoor, and SDSD-outdoor. Our method performs the best consistently.

very noisy low-light images. Fig. 7 also shows visual results on SMID, SDSD-indoor, and SDSD-outdoor. These results also manifest that our method is effective to enhance image lightness and reveal details, while suppressing noise.

User study. We further conduct large-scale user study with 100 participants to evaluate the human perception of our method and five strongest baselines (chosen by mean PSNR on SID, SMID, and SDSD) on enhancing low-light photos captured by an iPhone X or Huawei P30. Altogether, 30 low-light photos were captured in various environments, including roads, parks, libraries, schools, portraits, etc., and the intensity of 50% image pixels are lower than 30%.

Following the setting in [40], we evaluate results via user ratings on the six questions shown in Fig. 8 using a Likert scale of 1 (worst) to 5 (best). All methods were trained on SDSD-outdoor, since [39] shows that the trained models can effectively enhance low-light images captured by mobile phones. Fig. 8 reports the rating distributions of different methods, among which our approach receives more "red" and less "blue" ratings. Also, we performed a statistical analysis on the ratings using a paired t-test (using the T-Test function in MS Excel) between our approach and each of the other methods. With a significant level of 0.001, all the t-test results are statistically significant, since all p-values are smaller than 0.001.

4.3. Ablation Study

We consider four ablation settings by removing different components from our framework individually.

- "Ours w/o L" removes the long-range branch, so the framework has only convolutional operations.
- "Ours w/o S" removes the short-range branch, keeping the full long-range branch and SNR-guided attention.
- "Ours w/o SA" further removes the SNR-guided attention from "Ours w/o S", keeping only the basic transformer structure in the deepest layer.
- "Ours w/o A" removes the SNR-guided attention.

We performed ablation studies on all seven datasets. Table 5 summarizes the results. Compared with all ablation

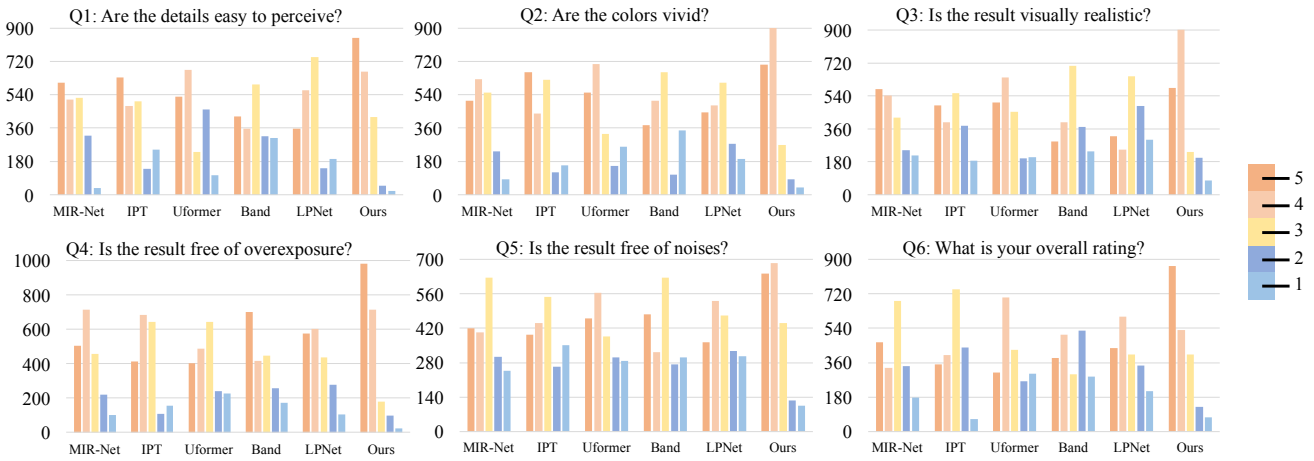


Figure 8. Rating distributions for different methods on the six questions in the user study. The ordinate axis records the rating frequency received from the 100 participants. Obviously, our approach receives more "red" (score 5) and less "blue" (score 1).

	LOL-v1		LOL-v2-real		LOL-v2-synthetic		SID		SMID		SDSD-indoor		SDSD-outdoor	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Ours w/o L	16.27	0.638	16.98	0.687	20.81	0.881	19.10	0.593	26.20	0.776	22.24	0.818	20.03	0.713
Ours w/o S	23.06	0.828	18.98	0.790	23.47	0.919	22.30	0.604	27.00	0.768	28.13	0.884	25.43	0.823
Ours w/o SA	20.67	0.752	18.85	0.765	21.88	0.842	21.02	0.544	27.01	0.774	25.78	0.839	24.57	0.832
Ours w/o A	21.86	0.760	19.40	0.782	22.23	0.866	21.19	0.550	26.87	0.769	27.36	0.874	26.62	0.857
Ours	**24.61**	**0.842**	**21.48**	**0.849**	**24.14**	**0.928**	**22.87**	**0.625**	**28.49**	**0.805**	**29.44**	**0.894**	**28.37**	**0.862**

Table 5. Results of the ablation study in Sec. 4.3.

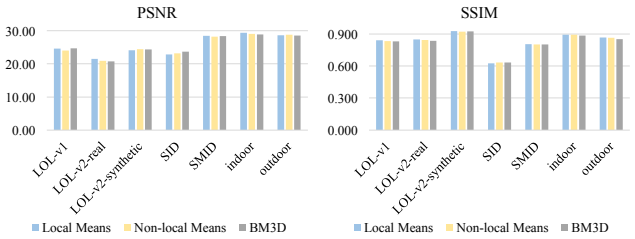


Figure 9. Our framework yields consistent performance on datasets when incorporated with different denoising operations for obtaining the input SNR prior.

settings, our full setting yields the highest PSNR and SSIM. "Ours w/o L," "Ours w/o S," and "Ours w/o SA" show the shortcomings of solely using convolution operations or transformer structures, thus manifesting effectiveness of collectively exploiting short-range (convolution models) and long-range (transformer structures) operations. The results also show effects of "SNR-guided attention" ("Ours w/o A" vs. "Ours") and "SNR-guided fusion" ("Ours w/o S" vs. "Ours").

4.4. Influence of SNR Prior

The SNR input to our framework is obtained by applying a no-learning-based denoising operation to the input frame (Eq. (2)). In all experiments, we adopt the local means as the denoising operation considering its fast speed. In this section, we analyze the impact when embracing other operations, including non-local means [1] and BM3D [8]. Fig. 9 shows the results, revealing that our framework is not sensitive to strategies of obtaining the SNR input. All these results are better than those of baselines.

5. Conclusion

We have presented a novel SNR-aware framework that collectively exploits short- and long-range operations to dynamically enhance pixels in a spatial-varying manner. An SNR prior is adopted to guide feature fusion. The SNR-aware transformer is formulated with a new self-attention module. Extensive experiments, including user study, demonstrate that our framework consistently achieves the best performance on representative benchmark using the same network structure.

Our future work is to explore other semantics to enhance the spatial-varying mechanism. Also, we plan to extend our method to handling low-light videos by simultaneously considering temporal- and spatial-varying operations. Another direction is to explore generative methods [13, 32] for nearly black areas in low-light images.

Acknowledgements We would like to thank Prof. Ying-Cong Chen in HKUST(GZ) for discussions.

References

[1] Antoni Buades, Bartomeu Coll, and J-M. Morel. A non-local algorithm for image denoising. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2005. 4, 8

[2] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Trans. Image Process.*, 2018. 2

[3] Damon M. Chandler and Sheila S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Trans. Image Process.*, 2007. 1

[4] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *Int. Conf. Comput. Vis.*, 2019. 2, 5

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 5, 6, 7

[6] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3, 5, 6, 7

[7] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep Photo Enhancer: Unpaired learning for image enhancement from photographs with GANs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3D filtering. In *Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning*, 2006. 4, 8

[9] Xuan Dong, Guan Wang, Yi Pang, Weixin Li, Jiangtao Wen, Wei Meng, and Yao Lu. Fast efficient algorithm for enhancement of low lighting video. In *Int. Conf. Multimedia and Expo*, 2011. 5, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021. 3

[11] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 2016. 5, 6

[12] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1, 5, 6

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Adv. Neural Inform. Process. Syst.*, 2014. 8

[14] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[15] Xiaojie Guo, Yu Li, and Haibin Ling. LIME: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.*, 2016. 1, 5, 6

[16] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 1

[18] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[19] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. WESPE: Weakly supervised photo enhancer for digital cameras. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[20] 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 Trans. Image Process.*, 2021. 2, 5, 6, 7

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2021. 3

[22] Hanul Kim, Su-Min Choi, Chang-Su Kim, and Yeong Jun Koh. Representative color transform for image enhancement. In *Int. Conf. Comput. Vis.*, 2021. 2, 6

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 5

[24] Satoshi Kosugi and Toshihiko Yamasaki. Unpaired image enhancement featuring reinforcement-learning-controlled image editing software. In *AAAI*, 2020. 5, 6, 7

[25] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018. 5

[26] Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang. Luminance-aware pyramid network for low-light image enhancement. *IEEE Trans. Multimedia*, 2020. 5, 6, 7

[27] Jing Li, Stan Z. Li, Quan Pan, and Tao Yang. Illumination and motion-based video enhancement for night surveillance. In *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005. 1

[28] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust Retinex model. *IEEE Trans. Image Process.*, 2018. 2, 5, 6

[29] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1, 2, 5, 6, 7

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, 2021. 3

[31] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 2017. 2

[32] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv:1411.1784*, 2014. 8

[33] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. DeepLPF: Deep local parametric filters for image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2, 5, 6, 7

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst.*, 2019. 5

[35] Zia-ur Rahman, Daniel J. Jobson, and Glenn A. Woodell. Retinex processing for automatic image enhancement. *IEEE Trans. Image Process.*, 2004. 2

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 5

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Int. Conf. Learn. Represent.*, 2014. 5

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1, 3

[39] 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 *Int. Conf. Comput. Vis.*, 2021. 2, 5, 7

[40] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2, 5, 6, 7

[41] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Trans. Image Process.*, 2013. 1

[42] Tao Wang, Yong Li, Jingyang Peng, Yipeng Ma, Xian Wang, Fenglong Song, and Youliang Yan. Real-time image enhancer via learnable spatial-aware 3D lookup tables. In *Int. Conf. Comput. Vis.*, 2021. 2, 5, 6, 7

[43] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 2004. 6

[44] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general U-Shaped transformer for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3, 5, 6, 7

[45] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep Retinex decomposition for low-light enhancement. In *Brit. Mach. Vis. Conf.*, 2018. 2, 5, 6, 7

[46] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. In *Int. Conf. Comput. Vis.*, 2021. 3

[47] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. In *ICML*, 2015. 5

[48] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1, 2, 5, 6, 7

[49] Jianzhou Yan, Stephen Lin, Bing Kang Sing, and Xiaoou Tang. A learning-to-rank approach for image color enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 2

[50] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 2016. 2

[51] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 2

[52] 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 Trans. Image Process.*, 2021. 2, 5, 6, 7

[53] Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep Retinex network for robust low-light image enhancement. *IEEE Trans. Image Process.*, 2021. 2, 5, 6, 7

[54] Susu Yao, Weisi Lin, EePing Ong, and Zhongkang Lu. Contrast Signal-to-Noise Ratio for image quality assessment. In *IEEE Int. Conf. Image Process.*, 2005. 1

[55] Zhenqiang Ying, Ge Li, and Wen Gao. A bio-inspired multi-exposure fusion framework for low-light image enhancement. *arXiv:1711.00591*, 2017. 5, 6

[56] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. In *Int. Conf. Comput. Vis.*, 2017. 1

[57] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Int. Conf. Comput. Vis.*, 2021. 3

[58] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet. In *Int. Conf. Comput. Vis.*, 2021. 3

[59] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Eur. Conf. Comput. Vis.*, 2020. 2, 5, 6, 7

[60] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 2, 5, 6, 7

[61] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM Int. Conf. Multimedia*, 2019. 5, 6, 7

[62] Lin Zhao, Shao-Ping Lu, Tao Chen, Zhenglu Yang, and Ariel Shamir. Deep symmetric network for underexposed image enhancement with recurrent attentional learning. In *Int. Conf. Comput. Vis.*, 2021. 2, 6

[63] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. In *Int. Conf. Comput. Vis.*, 2021. 2