Computer Vision Faculty Publications Scholarly Works

1-4-2021

Low light image enhancement via global and local context modeling

Aditya Arora
Inception Institute of Artificial Intelligence

Muhammad Haris
Mohamed bin Zayed University of Artificial Intelligence

Syed Waqas Zamir
Inception Institute of Artificial Intelligence

Munawar Hayat
Monash University. Australia

Fahad Shahbaz Khan
Mohamed bin Zayed University of Artificial Intelligence

See next page for additional authors
Follow this and additional works at: https://dclibrary.mbzuai.ac.ae/cvfp

Part of the Artificial Intelligence and Robotics Commons
Preprint: arXiv
Archived with thanks to arXiv
Preprint License: CC BY-NC-SA 4.0
Uploaded 25 March 2022

Recommended Citation

Authors

Aditya Arora, Muhammad Haris, Syed Waqas Zamir, Munawar Hayat, Fahad Shahbaz Khan, Ling Shao, and Ming-Hsuan Yang

Low Light Image Enhancement via Global and Local Context Modeling

Aditya Arora[1]   Muhammad Haris[2]   Syed Waqas Zamir[1]   Munawar Hayat[3]
Fahad Shahbaz Khan[2]   Ling Shao[1]   Ming-Hsuan Yang[4,5]
[1]Inception Institute of AI, UAE   [2]Mohamed bin Zayed University of AI, UAE
[3]Monash University, Australia   [4]University of California, Merced   [5]Google Research

Abstract

Images captured under low-light conditions manifest poor visibility, lack contrast and color vividness. Compared to conventional approaches, deep convolutional neural networks (CNNs) perform well in enhancing images. However, being solely reliant on confined fixed primitives to model dependencies, existing data-driven deep models do not exploit the contexts at various spatial scales to address low-light image enhancement. These contexts can be crucial towards inferring several image enhancement tasks, e.g., local and global contrast, brightness and color corrections; which requires cues from both local and global spatial extent. To this end, we introduce a context-aware deep network for low-light image enhancement. First, it features a global context module that models spatial correlations to find complementary cues over full spatial domain. Second, it introduces a dense residual block that captures local context with a relatively large receptive field. We evaluate the proposed approach using three challenging datasets: MIT-Adobe FiveK, LoL, and SID. On all these datasets, our method performs favorably against the state-of-the-arts in terms of standard image fidelity metrics. In particular, compared to the best performing method on the MIT-Adobe FiveK dataset, our algorithm improves PSNR from 23.04 dB to 24.45 dB.

1. Introduction

In well-lit conditions, consumer DSLR and smartphone cameras capture reasonably good quality images. However, in dimly-lit scenes, they often yield images that are noisy, without scene details, and manifest poor color and contrast. These degraded images pose challenges to various fundamental tasks in computer vision, such as semantic segmentation, object detection and tracking. Thus, it is of great importance to develop effective image enhancement methods to generate higher-quality images from the degraded inputs.

To capture well-exposed images in low light conditions, one may use high ISO, long exposures, and camera flash. However, such techniques are limited in various aspects [54]. For instance, increasing ISO can improve brightness but at the expense of noise amplification. Long exposure is limited to stationary scenes, and slight camera and/or object motion can potentially result in a blurry image. On the other hand, flashes can light up the scenes but usually introduces unwanted highlights and unbalanced illumination. Above all, it is likely for a user with constrained photographic devices *e.g.*, typical mobile phone cameras, does not have access to these options. Interactive methods for image enhancement allow users to adjust photos. They are, however, difficult and tedious for ordinary users. These methods require simultaneous manipulation of different parameters relating to color and contrast [41]. Semi-automatic methods alleviate this problem to an extent by reducing the number of adjustable parameters. As these methods are sensitive to parameters and developed based on heuristic rules, the quality of the enhanced images is likely low [8].

Several algorithmic advances have been made to address this problem. Initial attempts [3, 12, 42, 51, 56] mainly focus on enhancing contrast, and are limited in recovering color and local details. Recently, deep CNNs yield promising results by learning adjustments of color, contrast, brightness and saturation simultaneously. Towards this end, Chen *et al*. [5] develop a low-light image dataset and a data-driven camera processing pipeline, and Ran *et al*. [37] propose a two-stream architecture through combining encoder-decoder network and a Recurrent Neural Network (RNN). Based on the Retinex model, Zhang *et al*. [54] decompose an image into illumination and reflectance components, whereas the method by Wang *et al*. [41] learn image-to-illumination mapping. Despite significant progress, these methods may not generate optimal images in terms of brightness, color and contrast see Figures 1, 4, 5, and 6).

A plausible explanation for this can be attributed to the reliance on local operators to model dependency across different regions of the (degraded) input. Local operators, typically in the form of a convolution kernel, have a fixed, local receptive field. It cannot potentially infer several adjustments and remove artifacts that potentially demand involvement of correlated information, termed *context* hereafter, scattered at various spatial scales. Contextual information
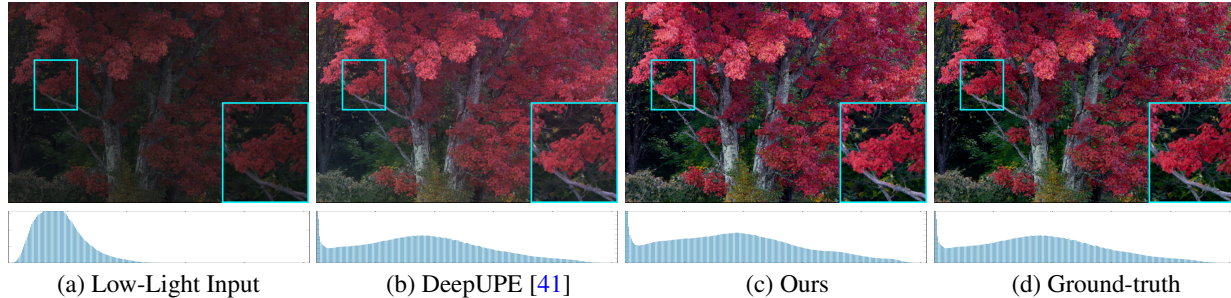
Figure 1: Visual comparison of our approach with the state-of-the-art method DeepUPE [41]. The proposed algorithm exhibits better local and global contrast (see histograms) with more faithful details to the ground-truth.

(a) Low-Light Input   (b) DeepUPE [41]   (c) Ours   (d) Ground-truth

is known for playing a pivotal role towards improving the performance of numerous high-level computer vision tasks [6, 10, 27, 47]. However, its potential is not fully harnessed for the data-driven based low-light image enhancement methods. We show that modeling context flexibly at various scales can provide complementary cues, which are not typically available with local receptive fields.

Global context leverages full spatial extent to search for correlated cues in the hope of enhancing a given feature. For example, to be able to infer global contrast for a given pixel on some foreground object, we may need to refer distant background regions. Further, the flexibility of querying full spatial extent and adjusting response accordingly allows effective local adjustments based on overall lighting conditions and scene settings. Aside from global context, information surrounding immediate vicinity of a certain spatial location can facilitate the model to better handle difficult regions. For instance, an underexposed image may have some regions (*e.g.*, dark) that are difficult to reconstruct reliably relative to others (*e.g.*, bright). To recover color in a dimly-lit car frontal region, we may need to also encode visual cues from the well-exposed similar appearing local regions.

Contributions. We propose a context-aware network for low-light image enhancement. First, the proposed method exploits a *global context* module for adjusting to the overall scene settings via querying full spatial extent and recalibrating accordingly for a given feature. Second, the proposed method introduces *local context* modeling; it offers a relatively larger field-of-view with dense feature sampling. The local context modeling improves context diversity by complementing its global context counterpart. The local context modeling encodes relatively large contextual representation in a dense manner, thereby better recovering local details in difficult regions of a degraded image for reliable reconstruction. We validate our method on three challenging image enhancement benchmarks: MIT-Adobe FiveK [2], LoL [46], and SID [5]. We perform a thorough ablation to demonstrate the effectiveness of harnessing learnable context towards low-light enhancement problem. Experimental results demonstrate that our approach performs favorably against baselines without using contextual information.

When compared to the state-of-the-art method [41] on the MIT-Adobe FiveK dataset, our algorithm achieves performance gain from 23.04 dB to 24.45 dB in terms of PSNR. Finally, the novel incorporation of context into an image enhancement network is likely to be adopted by future work and will serve as a step to develop new context-aware mechanisms for better and effective image enhancement.

2. Related Work

Image Processing Methods. An intuitive approach to enhance the visibility in a low-light image is via increasing brightness. Albeit simple, this approach may amplify noise and generate strong color casts. Histogram equalization based methods aim to make the image histogram uniform. However, histogram equalization method often produces results that are either under-enhanced or over-enhanced [1].

Instead of globally adjusting contrast, more informed approaches make use of illumination concept. The enhancement methods based on the Retinex theory [24] decompose an image into a pixel-wise product of an illumination and reflectance components, *e.g.*, Single-scale Retinex (SSR) [21] and Multi-scale Retinex (MSR) [20]. However, these methods often generate unnatural outputs. Wang *et al.* [43] propose to jointly enhance the contrast and conserve illumination, and Guo *et al.* [15] recover a structured illumination map from an initial one. These methods do not explicitly deal with noise and color distortions while enhancing images. In [12], Fu *et al.* propose a weighted variational model for simultaneous estimation of illumination and reflectance. Li *et al.* [26] extend [15] by introducing an additional term to account for extra noise. These two approaches [12, 15] are able to deal with noise to some extent, but may not effectively correct color distortions.

Deep Learning Methods. Numerous CNN models have been developed for low-level vision tasks, e.g., denoising [25, 35], demosaicking [23], restoration [53], and dehazing [4]. Specifically for low-light imaging, Lore *et al.* [29] propose a stacked-sparse denoising autoencoder for both contrast enhancement and denoising. Shen *et al.* [39] show that the multi-scale Retinex model is equivalent to a feedforward

CNN with different Gaussian kernels and propose the MSR-Net to learn the mapping from dark to bright images. Based on the Retinex model, Wei *et al*. [46] develop the Retinex-Net that performs decomposition followed by illumination adjustment for image enhancement. Recently, generative adversarial networks (GAN) have been used for image enhancement. Chen *et al*. [8] present an unpaired learning model in a two-way GAN framework, while Ignatov *et al*. [19] design a weakly-supervised GAN-based model. In addition, Deng *et al*. [9] develop an aesthetic-driven enhancement method via adversarial learning.

With the aim of learning camera imaging pipeline, Chen *et al*. [5] train an encoder-decoder network. Ren *et al*. [37] develop an architecture by combining an encoder-decoder and a RNN for low-light enhancement. Inspired by the Retinex theory, Zhang *et al*. [54] decompose an image into an illumination component for light adjustment and a reflectance component for degradation removal. Wang *et al*. [41] pose the recovery of the reflectance component of a degraded image as the enhanced image and propose to learn image-to-illumination mapping. Despite significant progress, the aforementioned approaches may not perform well in cases that require complementary cues from the full spatial extent. To this end, we introduce a context-aware CNN that imposes global consistency with the capability of encoding richer contextual features densely.

Contextual cues in vision tasks. Contextual cues have been widely used for vision problems, such as detection [6, 27] and semantic segmentation [10, 34]. Recently, *learnable context* has been shown to be effective for numerous high-level vision tasks [14, 27, 34, 50]. Here, we show that contexts (both global and local) can be better exploited in data-driven models for low-light image enhancement.

3. Method

Figure 2 shows the overall architecture of our context-aware deep network. It takes a degraded image (an under-exposed photo) as an input, and reproduces a high quality, well-exposed image as output. Fundamentally, it is an encoder-decoder formulation built upon small convolutions (*e.g.*, 3×3) and skip connections. We briefly describe encoder-decoder architecture in the following section.

Encoder-Decoder Architecture. Let x be the short-exposure (sRGB) image captured in low-light and $\hat{x}$ be the enhanced output image. At the encoding stage, the input is downsampled steadily after generating features eventually bringing it to the lowest possible resolution. We can represent encoder latent output z as:

$$z = e(x), \tag{1}$$

where e is the encoding function. It can be expressed as:

$$e(x) = s_{e,m}(s_{e,m-1}(s_{e,0})), \tag{2}$$

where $s_{e,m}$ denotes the mth stage of the encoder. It comprises a basic block, denoted as bb, for computing new features followed by a maxpool operation for downsampling.

$$s_{e,m} = pool_{e,m}(bb_{e,m}(x)). \tag{3}$$

After computing new features $bb_{e,m}$ and before downsampling, the signal is bypassed with a skip connection, *i.e.*, $skip_{e,m} = bb_{e,m}$ to the decoder side to compensate for the missing details while spatial decoding.

When decoding, the encoder output is gradually upsampled after concatenating skip features to reconstruct the output with the same resolution as input. The decoder output can be written as:

$$\hat{x} = d(z), \tag{4}$$

where d is the decoding function, further expanded as:

$$d(z) = s_{d,m}(s_{d,m-1}(s_{d,0})), \tag{5}$$

where $s_{d,m}$ denotes the mth stage of the decoder. It first upsamples the input, concatenates it with the bypass (skip) output from the encoder side, and finally generates new features using the basic block:

$$s_{d,m} = bb_{d,m}([up_{d,m}(z), skip_{e,m}]), \tag{6}$$

where $[\cdot]$ is a concatenation operator. We use the encoder-decoder based architecture, including skip connections, as our baseline framework for low-light image enhancement.

3.1. Proposed Context-aware Network

Encoder-decoder architectures [28, 32, 38] extract deep features at various resolutions to encode the input x into a latent representation z, before performing spatial decoding to generate the output $\hat{x}$. When applied to low-light image enhancement problems the (expected) output is a well-exposed image. Even after successive downsampling, these models rely on convolutional operators to improve the view of the input, to generate z. As such, these models have limited capacity in exploiting the full spatial domain. To this end, we introduce the global context (GC) model. Furthermore, we show that the basic block in these network formulations, responsible for extracting new features, is limited in terms of receptive field size and rigorous input sampling. Notably, this block is repeated multiple times (in each stage of the encoder as well as the decoder). Therefore, addressing this issue is important for the problem studied in this work. For instance, it can facilitate enhancing the color and contrast of local regions in the input image. To accomplish this goal, we propose a local context (LC) model.

Global Context (GC). Deep CNNs for low-light image enhancement are mainly built upon convolutional operators [5, 41, 54]. Being local in nature and fixed-shaped, convolutional operators are limited in modeling dependencies
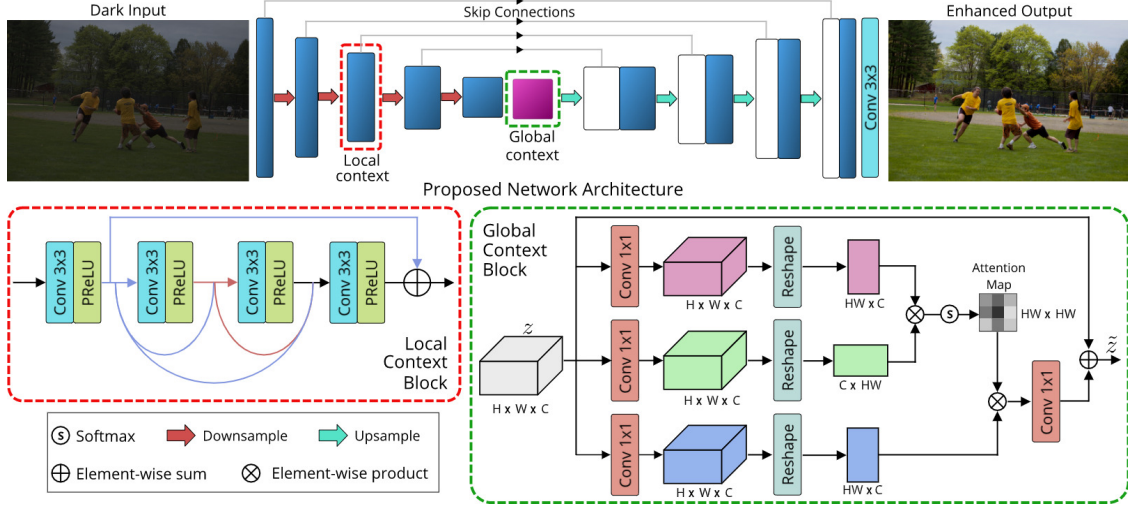
Figure 2: Overall network architecture for low-light image enhancement. It takes a degraded image (an underexposed photo) as input, and reproduces a well-exposed image as output. At its core, it is an encoder-decoder formulation. First, we introduce global context module, highlighted with dotted green box, to resolve inconsistencies requiring distant complementary cues, thereby utilizing full spatial domain. Second, we introduce a dense residual block as an effective replacement to the basic block, shown with dotted red, to capture local context at various resolutions in an encoder-decoder architecture.

across features. However, numerous vision problems entail involvement of non-local features as complementary cues for accurate processing. For instance, inferring global contrast of a foreground pixel requires considering certain distant background regions. Notably, these operations cannot be achieved with a limited view of the input image.

Although stacking several convolutional layers may alleviate this issue and capture greater context, it also brings a number of different challenges. First, it makes the model unnecessarily deep and large, thereby requiring higher computational load and bigger memory in addition to the increased over-fitting risks. Second, information of features that are distant from a location need to be passed through several layers before affecting the location for both forward propagation and backward propagation, thereby making the optimization more complex [7].

The integration of contextual information, within deep CNNs, has been investigated in different vision tasks [17, 16, 55]. Most of these approaches, such as [17] summarize the input information across spatial dimensions and rescale channel features to acquire global context. While simple and efficient, they assume the same mask for each spatial position. To this end, we propose to capture global context via non-local blocks (see Figure 2) which are designed to model long-range dependencies in video/image classification [45]. To our knowledge, we are the first to propose the utilization of non-local blocks to capture global context for low-light image enhancement. A non-local block strengthens a given feature by aggregating features from the full spatial extent via a pairwise affinity measure.

We denote $z = \{z_i\}_{i=1}^{N_p}$ as the feature map of an input

tensor where z is the latent representation (Eq. 1), and N_p is the total number of positions in a feature map, *i.e.*, the full spatial extent of a feature map ($N_p = H \times W$). With z and $\tilde{z}$ (of the same dimension) representing the input and output of the non-local block, we have:

$$\tilde{z}_i = z_i + W_u \sum_{j=1}^{N_p} \frac{f(z_i, z_j)}{C(z)} (W_v.z_j), \qquad (7)$$

where i corresponds to the index of a query position and j iterates through all possible positions. In Eq. 7, $f(z_i, z_j)$ finds the affinity between position i and j, and has a normalization factor $C(z)$. In addition, W_u and W_v denote linear projection matrices (implemented as 1x1 conv.). For notation brevity, we denote $d_{ij} = \frac{f(z_i, z_j)}{C(z)}$ as normalized pairwise affinity between position i and j. The difference lies in how pairwise affinity d_{ij} is computed. In our global context module, we utilize the embedded Gaussian to instantiate non-local block. Next, we describe our dense residual block for modeling local context.

Local Context (LC). As discussed earlier, the basic block, designated with the generation of new features, in an encoder-decoder based architecture has limited representational capability due to small receptive-fields along both resolution and scale axis w.r.t the input. Typically, a basic block comprises of two stacked convolutions of size 3×3 to generate features. Let f be the input feature maps, we can describe a basic block by:

$$bb(f) = \sigma_2(W_2 \sigma_1(W_1 f)), \qquad (8)$$

where W_1 and W_2 are the weight matrices for convolutions

Table 1: Ablation study. Both global and local context modeling contribute to the overall improvement. Our method achieves an absolute gain of 1.96 dB over the baseline. In addition, it achieves an absolute gain of 1.65 dB over either module.

Add-on				Ours
No-add (Baseline)	✓	✓	✓	✓
Global Context		✓		✓
Local Context			✓	✓
PSNR	22.49	22.69	22.80	**24.45**

and $\sigma(\cdot)$ is the PReLU activation.

A straightforward approach to increase the receptive-field size is to downsample the input (say ×2) and then use convolutions (3×3) followed by upsampling operation. This scheme, however, causes loss of spatial information (due to downsampling and upsampling) that is important for image restoration. To this end, we introduce a dense residual block (DRB) to model local context (see Figure 2) by allowing faster growth of receptive field in a dense fashion. It consists of 3 cascaded convolutions of kernel size 3×3. The output of each convolution is concatenated with the input features and all the outputs from previous convolutions. The concatenated features are then fed to the next convolution. Finally, we employ a skip connection from the input to the output of this block. Formally, we formulate each convolution layer output y_l in a DRB as:

$$y_l = M_{k,l}([y_{l-1}, y_{l-2}, \ldots, y_0]), \tag{9}$$

where $M_{k,l}$ is the convolutional operator with kernel size of k at layer position l, and $[\cdot]$ is a concatenation operator that combines outputs from all preceding layers. The output of a DRB is comptued by cascading convolutional layers y_l:

$$DRB(f) = y_l(y_{l-1}(y_0)). \tag{10}$$

Note that, the DRB output (Eq. 10) captures larger field-of-view and manifests denser features via dense connections (Eq. 9) than the bb output (Eq. 8). Furthermore, this operation is efficiently achieved by reusing features without incurring much computational overhead.

4. Experiments and Analysis

In this section, we first describe experimental settings and then we present ablation studies and main results.

Datasets. (1) **MIT-Adobe FiveK** [2] contains 5000 images captured with DSLR cameras. All photographs feature a broad range of scenes, subjects, and lighting conditions. Five photography students (A/B/C/D/E) manually adjust photos to generate expert-retouched images. Similar to [18, 33, 41], we use only the output of Expert C as the ground-truth. The first 4500 images are used for training and the last 500 for testing. (2) **LoL** [46] contains 500 real camera image pairs, out of which 485 images are for

training and 15 for testing. Each pair consists of a low-light source image and its corresponding well-exposed ground-truth. (3) **See-in-the-dark (SID)** [5] dataset is captured in extreme low-light with two different cameras: a) Sony α 7S II with Bayer filter array, and b) Fujifilm X-T2 with X-Trans array. We use the Sony subset which contains 424 images for training and 88 for testing.

Evaluated Methods. We evaluate our method by comparing it with low-light image enhancement methods. The LIME method [15] estimates illumination map by attending to each pixel in color channels and exploits the structure of illumination map. The KinD scheme [54] decomposes image into illumination and reflectance components via two separate networks. The DPE method [8] uses a two-way GAN for learning the mapping from dark to bright images. The DeepUPE model [41] learns a image-to-illumination transformation function by applying constraints on the recovered illumination map.

Training Details. We implement the proposed network using PyTorch with a single Titan-V100 GPU. The entire network is optimized using the Adam optimizer with an initial learning rate of 10^{-4}. For data augmentation, we extract random crops of size 512×512 followed by random flipping and rotation. The initial learning rate is decayed by a factor of 2 after every 128K iterations, and we train for a total number of 640K iterations. We employ the L_1 loss for network optimization. The source code and trained models will be released publicly.

4.1. Ablation Study

We validate the performance impact of integrating the proposed global context modeling and the local context modeling into the non-contextual (baseline) framework. Ablation experiments are performed on the MIT-Adobe FiveK [2] dataset. Table 1 shows the results from ablation studies. The baseline architecture (based on U-Net) achieves PSNR of 22.49 dB. The introduction of global context modeling improves the performance from 22.49 dB to 22.69 dB. We observe almost similar improvement (0.31 dB) after introducing local context modeling into baseline. The proposed algorithm, with both global and local context modeling, achieves an absolute performance gain of 1.96 dB in terms of PSNR in comparison to baseline. Note that the overall performance gain is substantially higher (by 1.65 dB) than that of using each module, individually. It indicates that both modules make complementary contributions for image enhancement.

In addition to the quantitative ablation results, we show enhanced images using these models in Figure 3. The enhanced images by the baseline method contain artifacts: for instance, on the bumper and around the lower left side of the car. When we introduce the global context modeling

Table 2: Quantitative comparisons of the evaluated methods on LoL dataset [46]. The proposed algorithm performs favorably against the state-of-the-art method with an absolute performance gain of 2.14 dB (PSNR). Bold indicates the best results.

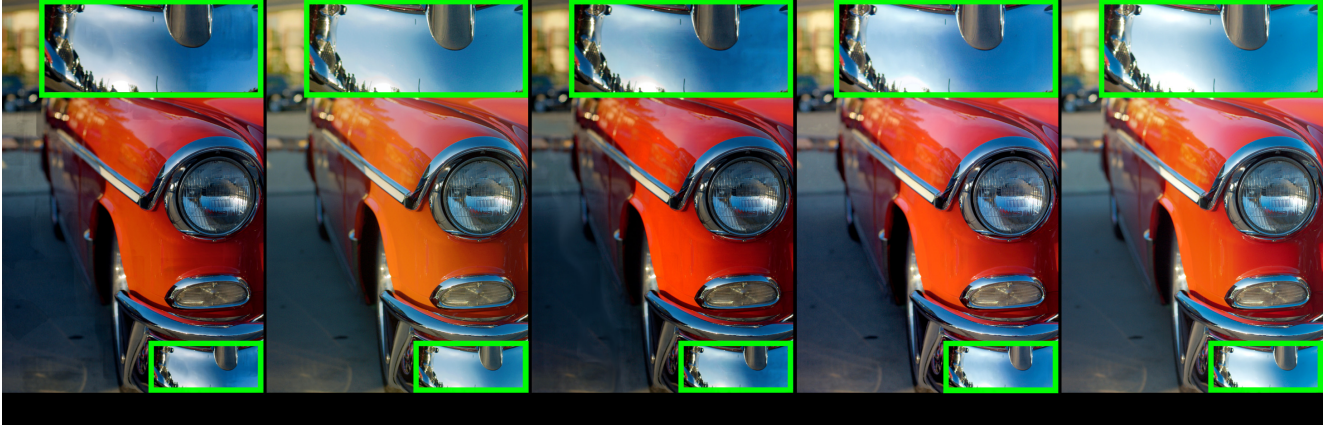| Method	BIMEF [48]	CRM [49]	Dong [11]	LIME [15]	MF [12]	RRM [27]	SRIE [12]	Retinex-Net [46]	MSR [20]	NPE [43]	GLAD [44]	KinD [54]	Ours
PSNR	13.86	17.20	16.72	16.76	18.79	13.88	11.86	16.77	13.17	16.97	19.72	20.87	**23.01**
SSIM	0.58	0.64	0.58	0.56	0.64	0.66	0.50	0.56	0.48	0.59	0.70	0.80	**0.89**

Figure 3: Illustration of how the proposed context-aware framework. From left, the first image is the baseline output, revealing artifacts on the bumper and around the lower left side of the car. Second image: with the inclusion of global context modeling these artifacts are removed. However, it cannot faithfully recover colors at different locations *e.g.*, behind the two lights. Third image: encoding local context helps decoding of colors (closer to the target) in the spatially homogeneous regions at various car locations. However, being local in nature, it does not perform adjustments at locations requiring access to global scene settings and generates artifacts. Fourth image: our approach based on both global and local contexts is able to recover vivid colors at different car locations without artifacts.

Table 3: Quantitative comparisons of evaluated methods on the MIT-Adobe FiveK dataset [2].

Method	HDRNet [13]	W-Box [18]	DR [33]	DPE [8]	DeepUPE [41]	Ours
PSNR	21.96	18.57	20.97	22.15	23.04	**24.45**
SSIM	0.866	0.701	0.841	0.850	0.893	**0.929**

Table 4: Quantitative results on the SID dataset. Proposed approach shows improved results compared to the method of [5].

Method	Chen *et al.* [5]	Maharjan *et al.* [30]	Zamir *et al.* [52]	Karadeniz *et al.* [22]	Ours
PSNR	28.96	29.17	28.84	29.29	**29.70**
SSIM	0.896	0.886	0.876	0.882	**0.902**

in the baseline, the enhanced images no longer have artifacts; however, the color appearance do not match faithfully to the well-exposed reference images (for example, see the red color of the car). Encoding local context helps reproducing colors in the spatially homogeneous regions around various car locations. However, being local in nature, it is not effective for doing adjustments requiring global view, thereby generating some artifacts. Our framework, based on both GC and LC models, removes artifacts and recovers vivid colors at two different car locations, generating an enhanced image perceptually-faithful to the ground-truth.

4.2. Quantitative Evaluation

First, we compare the performance of our method with several other algorithms on the LoL dataset. The results are shown in Table 2. The proposed method achieves 2.14 dB improvement over the recent best method KinD [54]. Next, we evaluate our method against five approaches on the MIT-Adobe FiveK dataset. Table 3 shows that our method performs well against all competing algorithms. Compared to the recent best method [41], our method achieves 1.41 dB PSNR gain).

Finally, we evaluate the proposed method on the RAW data from the SID dataset. It is a more challenging case as the input to the network is now a RAW image and the output is an enhanced sRGB image. It implies that the network has to learn the complete camera imaging pipeline (which applies a series of complex operations on RAW data in order to generate sRGB images). We use the Sony subset from the SID dataset [5] for evaluation. Table 4 shows a comparison of results obtained with our method and with those of the state-the-art [5, 30, 22, 52]. Our method achieves 0.41 dB improvement in PSNR over [22].

4.3. Qualitative Evaluation

We show visual comparisons of our algorithm and the state-of-the-art on three datasets. Figure 4 show the images produced by our method and the state-of-the-art methods on the MIT-Adobe FiveK dataset. It is noticeable that our algorithm generates images that are visually more pleasant and have better local and global contrast as compared to other competing approaches. Next, we present in Figure 5

(a) Low-Light Input — (b) DPE [8] — (c) HDRNet [13]

(d) DeepUPE [41] — (e) Ours — (f) Ground-Truth

(a) Low-Light Input — (b) DPE [8] — (c) HDRNet [13]

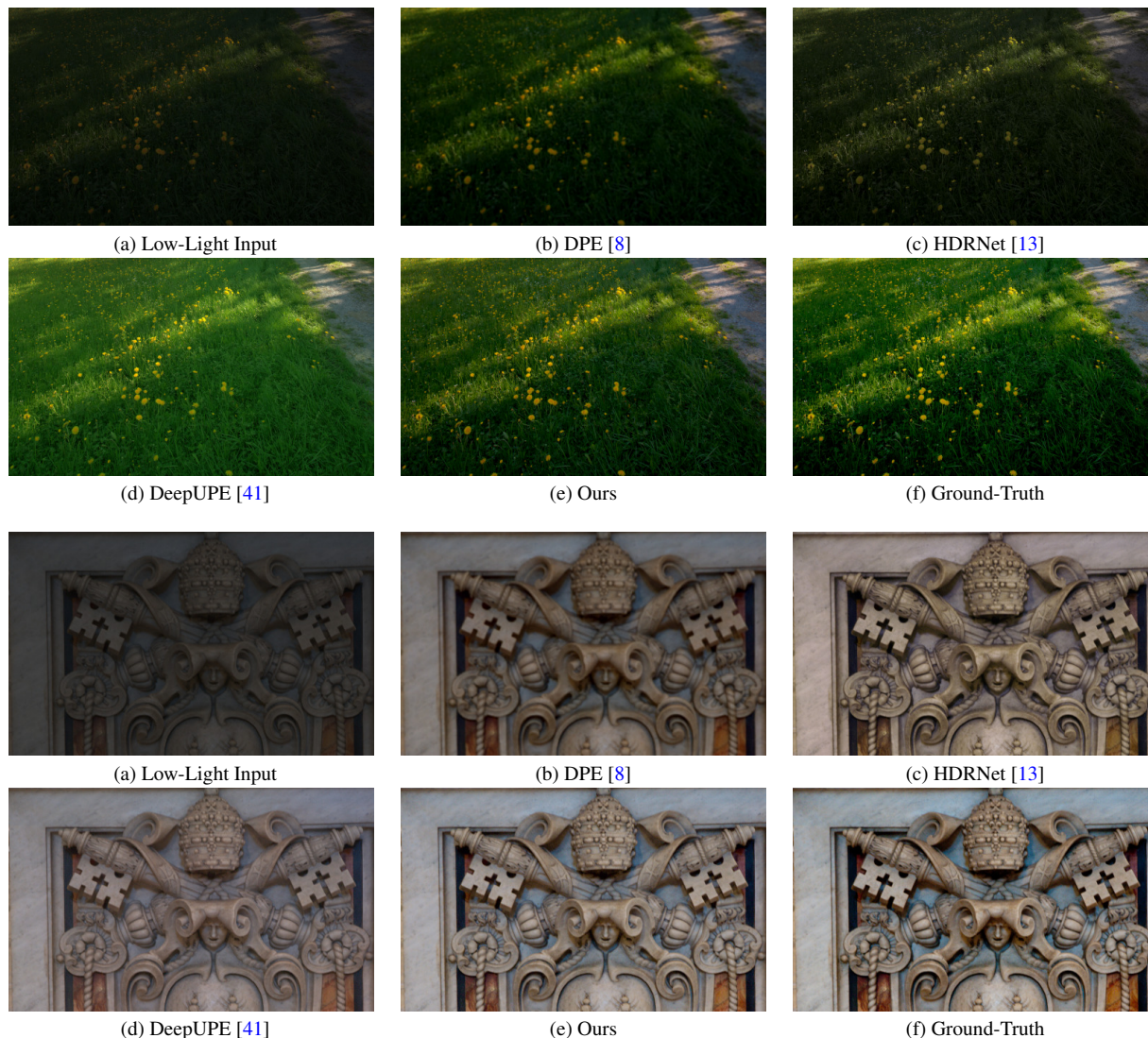(d) DeepUPE [41] — (e) Ours — (f) Ground-Truth

Figure 4: Qualitative example from MIT-Adobe Fivek [2]. The result produced by our method is vivid, and have better local and global contrast than other algorithms.

the image enhancement results on LoL dataset. It shows that our method produces perceptually more faithful results than other algorithms. For example, compare the overall brightness, contrast and object level details.

Finally, we show enhanced results of the proposed method under extreme low-light conditions. Figure 6 shows the results generated by the proposed algorithm and the state-of-the-art method of Chen *et al.* [5] on two challenging examples from the SID dataset. The images generated by the state-of-the-art method contain noticeable artifacts (all three zoomed-in views marked with green boxes). In contrast, the proposed algorithm generates images that are natural in appearance with sharper details. We note there is no existing work that achieves consistently better results on all three datasets. It corroborates the impact of harnessing global context and local context into a single architecture

that is effective for image enhancement.

4.4. Pyschophysical Evaluation

To further assess the effectiveness of our approach, we conduct user studies. We create a subset of 25 images by taking every 20th image from the test set of MIT-Adobe FiveK dataset. Each of these image is enhanced with our algorithm and the following four state-of-the-art approaches: W-Box [18], HDRNet [13], DPE [8], DeepUPE [41]. In the experiments, 20 observers with normal color vision participated. We used pair-wise comparison technique and each observer had to judge 250 pair of images. For analyzing psychophysical data, we use the method of [31], that is based on Thurstone's law of comparative judgement [40].

We ran two different psychophysical experiments, where each observer was shown a pair of corresponding enhanced

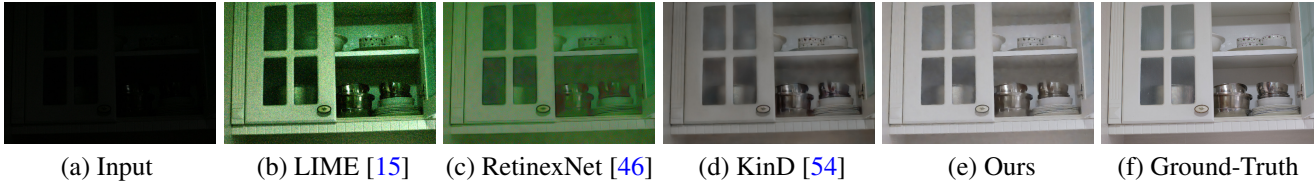(a) Input   (b) LIME [15]   (c) RetinexNet [46]   (d) KinD [54]   (e) Ours   (f) Ground-Truth

Figure 5: Qualitative comparison of our method with state-of-the-art approaches on the LoL dataset. Our method reconstructs image with distinct contrast and natural brightness. The object-level details in our results are more apparent than those of the other approaches.

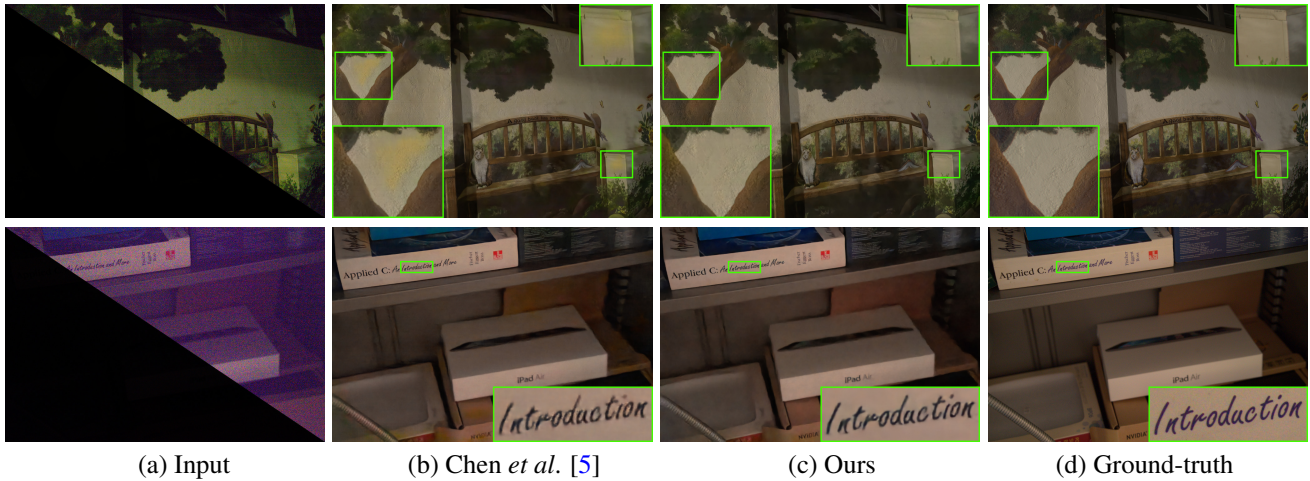(a) Input   (b) Chen *et al.* [5]   (c) Ours   (d) Ground-truth

Figure 6: Visual comparisons of the proposed algorithm with the conventional camera ISP and state-of-the-art method of [5] on the SID dataset. (a) Input image. For better visualization, the RAW image is passed through a conventional camera pipeline [36]. Note that the brightness of top right half is increased to reveal image details. (b) The enhanced images by Chen *et al.* [5] contain artifacts (highlighted with green boxes). (c) Our method is able to remove these artifacts and recover natural color and image details.
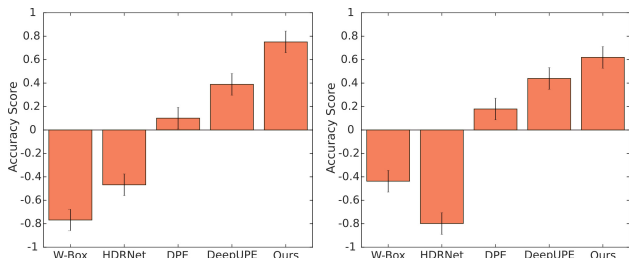
(a) Results showing ground-truth to observers.

(b) Results *without* showing ground-truth to observers.

Figure 7: Psychophysical Evaluation. Two experiments are performed: (a) aiming for the accurate image reproduction with respect to ground-truth, and (b) with the pleasant reproduction intent (without showing ground-truth to observers). Our method ranks first in both experiments according to the observers' choices.

images (1) in isolation, and (2) alongside the ground-truth. The first psychophysical test is to investigate which algorithm produces images that the observers find most pleasant, aesthetically (color, brightness, contrast, etc.). And the second psychophysical test is to evaluate which algorithm produces images that are perceptually most faithful to the ground-truth. Figure 7b shows the results for the first psychophysical experiment, where it can be seen that the observers preferred our method over other competing approaches. Furthermore, the HDRNet [13] ranks last in the

users' preference. In Figure 7a, we present the accuracy scores for the second psychophysical experiment. According to the observers, our method generates images that are perceptually more closer to the ground-truth than those of the other algorithms. Moreover, it is worth noting that, for this experiment, the overall ranking trend of the competing methods is same as obtained with the image fidelity metrics.

5. Conclusion

We propose a context-aware hierarchical network for low-light image enhancement. The proposed approach adapts to global scene settings via modeling spatial correlations across the whole spatial extent. This is realized by introducing global context module. We further improve context diversity by encoding relatively greater range of contextual representation in a dense manner at each resolution. This facilitates reliable recovery of local color and contrast while enhancing short-exposure images captured under low-light conditions. We perform experiments on three challenging datasets: MIT-Adobe FiveK, LoL, and SID. Results show that the proposed algorithm performs favorably against the state-of-the-art, both quantitatively and qualitatively. Lastly, introducing learnable context in an image enhancement network will facilitate the development of new contextual mechanisms for effective image enhancement.

References

[1] Marcelo Bertalmío, Vicent Caselles, Edoardo Provenzi, and Alessandro Rizzi. Perceptual color correction through variational techniques. *TIP*, 2007. 2

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 2, 5, 6, 7

[3] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *ICCV*, 2017. 1

[4] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *TIP*, 2016. 2

[5] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 1, 2, 3, 5, 6, 7, 8

[6] Xinlei Chen and Abhinav Gupta. Spatial memory for context reasoning in object detection. In *ICCV*, 2017. 2, 3

[7] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ 2-nets: Double attention networks. In *NIPS*, 2018. 4

[8] 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 *CVPR*, 2018. 1, 3, 5, 6, 7

[9] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACM Multimedia*, 2018. 3

[10] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 2, 3

[11] 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 *ICME*, 2011. 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 *CVPR*, 2016. 1, 2, 6

[13] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *TOG*, 2017. 6, 7, 8

[14] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017. 3

[15] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *TIP*, 2016. 2, 5, 6, 8

[16] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NIPS*, 2018. 4

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[18] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *TOG*, 2018. 5, 6, 7

[19] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Wespe: weakly supervised photo enhancer for digital cameras. In *CVPRW*, 2018. 3

[20] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *TIP*, 1997. 2, 6

[21] Daniel J Jobson, Zia-ur Rahman, and Glenn A Woodell. Properties and performance of a center/surround retinex. *TIP*, 1997. 2

[22] Ahmet Serdar Karadeniz, Erkut Erdem, and Aykut Erdem. Burst photography for learning to enhance extremely dark images. *arXiv preprint arXiv:2006.09845*, 2020. 6

[23] Filippos Kokkinos and Stamatios Lefkimmiatis. Image demosaicking using a cascade of convolutional residual denoising networks. In *ECCV*, 2018. 2

[24] Edwin H Land. The retinex theory of color vision. *Scientific american*, 1977. 2

[25] Stamatios Lefkimmiatis. Universal denoising networks: A novel CNN architecture for image denoising. In *CVPR*, 2018. 2

[26] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *TIP*, 2018. 2

[27] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *CVPR*, 2018. 2, 3, 6

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

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

[30] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via residual learning. In *ICME*, 2019. 6

[31] J. Morovič. Color gamut mapping. 2008. 7

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3

[33] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *CVPR*, 2018. 5, 6

[34] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *CVPR*, 2017. 3

[35] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 2

[36] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew. Color image processing pipeline. *IEEE Signal Processing Magazine*, 2005. 8

[37] Wenqi Ren, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, Junping Du, and Ming-Hsuan Yang. Low-light image enhancement via a deep hybrid network. *TIP*, 2019. 1, 3

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 3

[39] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma. Msr-net: Low-light image enhancement using deep convolutional network. *arXiv*, 2017. 2

[40] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 1927. 7

[41] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019. 1, 2, 3, 5, 6, 7

[42] Shuhang Wang, Woon Cho, Jinbeum Jang, Mongi A Abidi, and Joonki Paik. Contrast-dependent saturation adjustment for outdoor image enhancement. *JOSA A*, 2017. 1

[43] Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *TIP*, 2013. 2, 6

[44] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *FG*, 2018. 6

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 4

[46] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2, 3, 5, 6, 8

[47] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

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

[49] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new image contrast enhancement algorithm using exposure fusion framework. In *CAIP*, 2017. 6

[50] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 3

[51] Lu Yuan and Jian Sun. Automatic exposure correction of consumer photographs. In *ECCV*, 2012. 1

[52] Syed Waqas Zamir, Aditya Arora, Salman Khan, Fahad Shahbaz Khan, and Ling Shao. Learning digital camera pipeline for extreme low-light imaging. *arXiv preprint arXiv:1904.05939*, 2019. 6

[53] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2

[54] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. *arXiv preprint arXiv:1905.04161*, 2019. 1, 3, 5, 6, 8

[55] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 4

[56] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, 1994. 1