Learning to Restore Low-Light Images via Decomposition-and-Enhancement

Ke Xu[1,2] Xin Yang[1,†] Baocai Yin[1,3] Rynson W.H. Lau[2,†]
[1]Dalian University of Technology [2] City University of Hong Kong [3]Pengcheng Lab

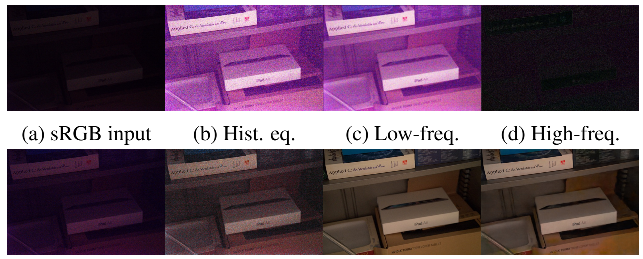
Abstract

Low-light images typically suffer from two problems. First, they have low visibility (i.e., small pixel values). Second, noise becomes significant and disrupts the image content, due to low signal-to-noise ratio. Most existing low-light image enhancement methods, however, learn from noise-negligible datasets. They rely on users having good photographic skills in taking images with low noise. Unfortunately, this is not the case for majority of the low-light images. While concurrently enhancing a low-light image and removing its noise is ill-posed, we observe that noise exhibits different levels of contrast in different frequency layers, and it is much easier to detect noise in the low-frequency layer than in the high one. Inspired by this observation, we propose a frequency-based decomposition-and-enhancement model for low-light image enhancement. Based on this model, we present a novel network that first learns to recover image objects in the low-frequency layer and then enhances high-frequency details based on the recovered image objects. In addition, we have prepared a new low-light image dataset with real noise to facilitate learning. Finally, we have conducted extensive experiments to show that the proposed method outperforms state-of-the-art approaches in enhancing practical noisy low-light images.

1. Introduction

Low-light imaging is very popular, for various purposes, e.g., night-time surveillance and personal scenery imaging at sunset. However, the visibility of low-light images in the standard RGB (sRGB, 24 bits/pixel) space does not match with human perception, due to quantization. This low visibility hinders vision tasks (*e.g.*, object detection [31] and tracking [8]), or image editing tasks (*e.g.*, image matting [45]). Hence, recovering low-light images is essential.

Typical image enhancement methods [46, 51, 24, 7, 40, 34, 48, 4] propose to recover low-light images to match with human perception. These methods rely on users to have good photographic skills in taking images with low noise, so that these methods can focus on learning to manipulate



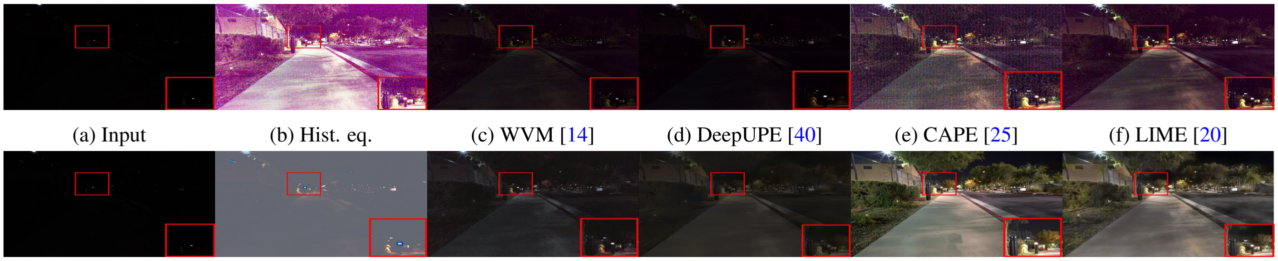
(a) sRGB input (b) Hist. eq. (c) Low-freq. (d) High-freq.
(e) DeepUPE [40] (f) DSLR [24] (g) Ground truth (h) Ours

Figure 1. Given a low-light sRGB image of 24-bit color depth (a), typical enhancement methods cannot produce a pleasant image with details recovered and noise suppressed (b, e, f). To illustrate our idea, we apply a Gaussian filter to decompose (b) into a low-frequency layer (c) and a high frequency layer (d), and observe that the low-frequency layer preserves sufficient information for recovering objects and colors, which can then be used to enhance high-frequency details. This inspires us to learn a decomposition-and-enhancement method for low-light images (h).

the tones, colors or contrasts of the images. As such, they cannot be used to enhance majority of the practical low-light images with noise, which are taken by casual users. Figure 1 shows one example, where image contents are not only buried by low pixel intensity values, but also disrupted by noise, due to the inherent low signal-to-noise ratio (SNR) at low light [6]. Existing enhancement methods may either enhance both the noise and scene details (Figure 1(b, f)), or fail to recover the low visibility of low-light images (Figure 1(e)). In addition, these enhanced images still have low SNRs, providing limited useful contextual information for detecting noise from scene details. Hence, they fail existing image denoising methods [11, 49, 50, 27, 37, 32, 19].

In this paper, we address the low-light sRGB image enhancement problem, which involves two issues: image enhancement as well as denoising. Our motivation is based on two observations. First, the image low-frequency layer preserves more information, *e.g.*, objects and colors, and is less affected by noise (Figure 1(c)) than the image high-frequency layer (Figure 1(d)). This suggests that it is easier to enhance the low-frequency image layer than to directly enhance the whole image. Second, the very low intrinsic dimensionality of image primitives makes it possible for neural networks to learn a full knowledge of image primitives [29, 41]. Hence, given the low-frequency informa-

[†] Xin Yang and Rynson Lau are the corresponding authors. Rynson Lau led this project.



(a) Input (b) Hist. eq. (c) WVM [14] (d) DeepUPE [40] (e) CAPE [25] (f) LIME [20]

(g) HDRCNN [12] (h) DRHT [46] (i) DSLR [24] (j) SID [6] (k) Ground truth (l) Ours

Figure 2. While existing methods ((c) to (j)) generally fail to enhance the input noisy low-light image (a), our method produces a sharper and clearer result with objects and details recovered (l).

tion of primitives, it is possible for a network to reconstruct the whole primitives by inferring the corresponding high-frequency information. With such a prior, we can then learn to enhance high-frequency details from the recovered low-frequency layer.

These two insights inspire us to learn a frequency-based low-light image decomposition-and-enhancement model. To this end, we propose a novel neural network that leverages an Attention to Context Encoding (ACE) module to adaptively select low-frequency information for recovering the low-frequency layer and noise removal in the first stage, and select high-frequency information for detail enhancement in the second stage. We also propose a Cross Domain Transformation (CDT) module to leverage multi-scale frequency-based features for noise suppression and detail enhancement in the two stages. As shown in Figure 2, our method can enhance the noisy low-light sRGB image with contents/details recovered and noise suppressed.

In summary, the main contributions of this work are:

1. We propose a novel frequency-based decomposition-and-enhancement model for enhancing low-light images. It first recovers image contents in the low-frequency layer while suppressing noise, and then recovers high-frequency image details.
2. We propose a network, with an Attention to Context Encoding (ACE) module to decompose the input image for adaptively enhancing the high-/low-frequency layers and a Cross Domain Transformation (CDT) module for noise suppression and detail enhancement.
3. We prepare a low-light image dataset with real noise and corresponding ground truth images, to facilitate the learning process.

Extensive experiments verify the superior performance of the proposed method over the state-of-the-art approaches.

2. Related work

Low-light image enhancement. A line of methods enhance low-light images using different image-to-image regression functions. Represented by histogram equalization [36] and gamma correction, global and local contrast enhancement operators are proposed based on detecting semantic regions (*e.g.*, face and sky) [25], matching region templates [23] or contrast statistics in image boundaries and textured regions [38]. Advanced deep learning based methods learn the mapping functions from high-quality user retouched images or images taken using high-end cameras, using bilateral learning [15], intermediate HDR supervision [46], adversarial learning [24, 7], or reinforcement learning [34, 48]. Another line of works are retinex-based image enhancement methods [20, 14, 51, 5, 40, 47], which decompose the input low-light image into illumination and reflectance, and then enhance the illumination of the image.

However, existing enhancement methods may fail to recover low-light images, due to their low SNRs, as shown in Figure 2. The key reason is that these methods [24, 34, 7, 48, 46] typically assume the images to be taken by photographic experts with insignificant noise levels. Hence, they are unable to enhance noisy low-light images.

Recently, there are also some enhancement methods [6, 22] proposed to directly retouch the camera raw data into high quality output images. Particularly, Chen *et al.* [6] proposed to learn raw-to-image models to generate noise-suppressed, enhanced images from noisy raw images. However, models trained on the raw domain cannot be applied to regular sRGB images, which is the most widely adopted color space [10], as the linear raw data is significantly different from the non-linear sRGB data [44]. Besides, raw data is usually unavailable due to the lack of expertise or unknown protocols. In this paper, we focus on enhancing noisy low-light sRGB images.

Image denoising. Single image denoising is an active research topic in computer vision, and it often functions as pre-/post-processing for other vision tasks. Many methods have been developed based on image priors such as self-similarity [3, 11], sparsity [13, 30], and low rank [18, 43]. Deep learning has also been widely applied to the denoising problem [33, 49, 50, 27, 37, 32]. These denoisers typically learned from synthetic datasets that assumed additive, white or Gaussian noise. They often fail to remove real noise, which exhibits different patterns. Recent works attempted to improve the performances of denoisers in denoising real

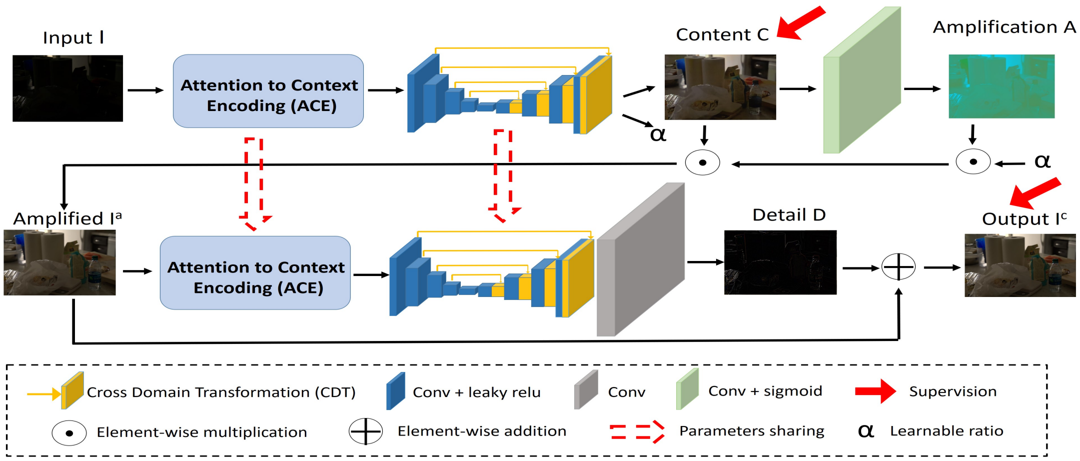


Figure 3. Overview of the proposed model. In the first stage, the network enhances the low-frequency contents of input image I with noise suppressed, and then amplifies it to produce I^a. In the second stage, the network infers the high-frequency details from I^a to produce the output enhanced image I^c.

images, by synthesizing noise in the raw data domain [2], constructing real image dataset [1], developing joint training strategy of both synthetic and real images [19], or unsupervised learning [28].

However, it is non-trivial to remove noise from low-light images simply by pre-/post-processing with existing denoising methods. On the one hand, low pixel values make it difficult to provide sufficient contextual information for detecting/removing noise before enhancing the low-light images. On the other hand, noise can be unpredictably amplified after applying existing enhancement methods, producing images that still have low SNRs and hence difficult for further denoising. To address this limitation, we propose in this paper to learn a deep enhancement model to enhance the low-light images while removing noise, in an end-to-end recurrent manner.

3. Proposed Model

Our method is inspired by two observations. First, it is easier to enhance the low-frequency layer of a noisy low-light image, compared to directly enhancing the whole image. This is because noise in the low-frequency layer is easier to detect and then suppress. Image illumination/colors can then be properly estimated by analyzing the global properties of the image low-frequency layer. Second, it is known that primitive parts of natural images, *e.g.*, edges and corners, have very low intrinsic dimensionality [29]. Such low dimensionality implies that a small number of image examples are sufficient to represent the image primitives well [41]. Hence, given the low-frequency information of the primitives, we may be able to infer the corresponding high-frequency information.

Based on these two observations, our proposed model, as shown in Figure 3, has two main stages. In the first stage, we propose to learn a low-frequency image enhancement function $C(\cdot)$, and then an amplification function $A(\cdot)$ for color recovery. By jointly modeling the mapping from $C(\cdot)$ to $A(\cdot)$, the network does not have to learn both global information (*e.g.*, illumination) and local information (*e.g.*, color) at the same time, resulting in a more effective enhancement. Formally, given a low-light sRGB image I, the first stage enhancement can be written as:

$$I^a = \alpha A(C(I)) \cdot C(I), \tag{1}$$

where I^a is the amplified low-frequency layer. Note that A is different from the illumination map in retinex-based methods, as we estimate a relative amplification map to a learnable global ratio α from the enhanced content C. In other words, $\alpha A(\cdot)$ can be interpreted as an error map that enhances C in the self-attention manner.

In the second stage, we propose to learn high-frequency detail enhancement function $D(\cdot)$, based on I^a from the first stage, instead of directly restoring the high-frequency details from the original input image I, which is noisy. $D(\cdot)$ is then modeled in a residual manner, and the final enhanced image can be obtained as:

$$I^c = I^a + D(I^a). \tag{2}$$

Figure 4 visualizes the output of each step of our model.

Our model uses two novel modules, the Attention to Context Encoding (ACE) module and the Cross Domain Tranformation (CDT) module. They are explained below.

3.1. ACE Module

The goal of the ACE module is to learn frequency-aware features for image decomposition. To do this, we extend the non-local operation [42], originally proposed for encoding long-range relations, to select frequency adaptive contextual information. Figure 5 shows the block diagram.

We use the first ACE module in Figure 3 for explanation. Given the input features $x_{in} \in R^{H \times W \times C}$, we first use two

(a) Input (b) Hist. eq. (c) Naive Reg. (d) C (e) A (f) I^a (g) D (h) I^c (i) Ground truth

Figure 4. Internal visualization (d-h) verifies the effectiveness of the proposed model, against naive image-to-image regression (c).

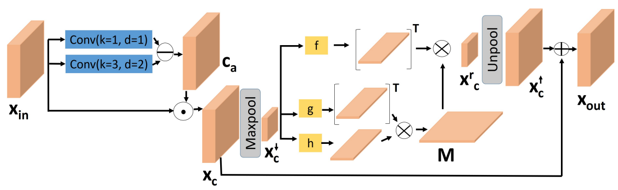


Figure 5. Overview of the proposed ACE module. It aims to decompose the image into frequency-based layers for adaptive enhancement in the two stages.

groups of dilated convolutions (with kernel size/dilation rate of 1/1 and 3/2), denoted as f_{d1} and f_{d2}, to extract features in different receptive fields. We then compute a contrast-aware attention map C_a between these two features as:

$$C_a = sigmoid(f_{d1}(x_{in}) - f_{d2}(x_{in})). \tag{3}$$

C_a indicates the pixel-wise relative contrast information, where pixels of high contrasts are regarded as belonging to the high-frequency layer. We then compute the inverse map $\overline{C}_a = 1 - C_a$ to select features from x_{in} to represent the low-frequency contents as: $x_c = \overline{C}_a \cdot x_{in}$. We further shrink the selected features x_c via max-pooling to obtain compact features $x_c^{\downarrow}$ and to reduce GPU memory and computations for establishing the non-local pixel-to-pixel dependence. Formally, given $x_c^{\downarrow} \in R^{H' \times W' \times C}$, the non-local context encoding process can be written as:

$$x_c^r = g(x_c^{\downarrow})^\top \times h(x_c^{\downarrow}) \times f(x_c^{\downarrow})^\top, \tag{4}$$

where g, h, f represent groups of operations (convolution, reshaping and matrix transpose) that first compute a pixel affinity table $M \in R^{H'W' \times H'W'}$ and then compute non-locally enhanced features x_c^r by considering the relations of each pixel to all other pixels. Finally, we obtain the frequency-aware non-locally enhanced features $x_{out} = Unpool(x_c^r) + x_c$ in a residual manner to facilitate the learning process. Note that the two ACE modules in Figure 3 share their weights. The second ACE module uses the contrast-aware attention map C_a, instead of the inverse map $\overline{C}_a$, to learn the image details from the features representing the high-frequency layer. Figure 6 shows two ACE attention maps ($\overline{C}_a$ from the first stage and C_a from the second stage) and their corresponding decomposed feature maps ($\overline{x}_c$ from the first stage and x_c from the second stage).

3.2. CDT Module

A good understanding of the global properties of low-light images can help recover the lighting and image contents. To do this, we propose the CDT module, as shown in Figure 7, to increase the receptive fields while bridging the gap between features in the low-light domain and in the enhanced domain. Sharing a similar spirit as [39] in increasing the receptive fields for more global information, the CDT module is specially designed to concurrently address the domain gap problem, *i.e.*, frequency-aware features extracted in the noisy low-light domain versus those in the enhanced domain.

(a) Input (b) Hist. eq. (c) Ours (d) GT

(e) $\overline{C}_a$ (f) C_a (g) $\overline{x}_c$ (h) x_c

Figure 6. Visual example of attention maps in the two-stage ACE module and the decomposed feature maps. $\overline{C}_a$ ($1st$ stage) tends to highlight background regions, while C_a ($2nd$ stage) attends more to foreground objects for reconstructing high-frequency details.

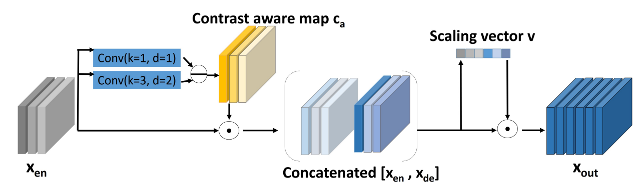


Figure 7. Overview of the proposed CDT module. It aims to increase the receptive fields while bridging the gap between the low-light domain and the enhanced domain.

Specifically, in the first stage, the noisy features from the encoder x_{en} are first spatially reweighed via the self-derived inverse contrast-aware map $\overline{C}_a$ to filter out high contrast information, before concatenating with features x_{de} from the corresponding decoder. We then compute global scaling vectors v from the concatenated features $[x_{en}, x_{de}]$, for adaptively re-scaling the features from different domains in a channel-wise manner. In the second stage, we use the contrast-aware attention map C_a, instead of the inverse map $\overline{C}_a$, to learn image details, similar to the ACE module.

3.3. Proposed Dataset

To facilitate the learning of the proposed model, we have prepared a new low-light dataset of real noisy low-light and ground truth sRGB image pairs.

Noise in low-light. We prepare our training data based on the SID dataset [6], which consists of raw data and ground truth image pairs. This raw data was collected when imaging in low-light with short exposure time (typically 0.1s or 0.04s). Their corresponding ground truth images were taken with long exposure time (typically 10s or 30s), where noise is negligible. However, the linear camera raw data is significantly different from the non-linear sRGB data, particularly in terms of noise [2] and image intensity [46]. As a result, models trained on raw data cannot be directly applied to sRGB images. To address this problem, we have considered several key steps (*i.e.*, exposure compensation, white balance and de-linearization) in the image formation pipeline, and manipulated their operations in order to model real-world noisy low-light sRGB images taken from different cameras.

Exposure compensation. Auto-exposure algorithms aim to automatically determine the exposure time and camera gain based on the light intensity perceived by the sensor. They are usually black-boxes and vary across cameras. To augment the diversity of this exposure time, we randomly sample the exposure compensation value from the range of $[0EV, 2EV]$ at intervals of $0.5EV$.

White balance. White balance algorithms aim to correct unrealistic casts via estimating the per-channel gain [16]. They are also unknown and vary across cameras. We augment it by randomly choosing the color temperature from the range of $[2100K, 4000K]$, which represents the color temperatures of typical household lighting and Sunrise/Sunset lighting, according to the Kelvin temperature color chart [9].

De-linearization. As the non-linearity introduced by the camera response function varies across cameras and is difficult to reverse-engineer [17], we apply the gamma function as the de-linearization function, as suggested in [12].

Using the above settings, we have produced a total of 4,198 image pairs for training and 1,196 image pairs for testing. Experimental results in Figures 9 and 10 show that the proposed network trained on our data can generalize well on images from other image formation pipelines.

3.4. Training

Loss function. We use L2 loss to measure the reconstruction accuracy in the two-stage training process. Specifically, in the first stage, to encourage our network to focus on predicting the low frequency components of the input image, we prepare the corresponding ground truth, denoted as I_f^{gt}, by using the guided filter [21] to filter out the high-frequency details while maintaining the main structures and contents of the ground truth image. Formally, the reconstruction loss can be written as:

$$L_{acc} = \lambda_1 \left\| C - I_f^{gt} \right\|_2 + \lambda_2 \left\| I^c - I^{gt} \right\|_2 , \qquad (5)$$

where C, I^c, I_f^{gt}, I^{gt} are the reconstructed image content, the recovered image, ground truth of the low-frequency layer, and ground truth of the enhanced image, respectively. λ_1 and λ_2 are balancing parameters.

We also incorporate the perceptual loss by comparing the VGG feature distances of I^c and I^{gt}, using L1 loss, as:

$$L_{vgg} = \lambda_3 \left\| \Phi(I^c) - \Phi(I^{gt}) \right\|_1 , \qquad (6)$$

where Φ is the VGG net, and λ_3 is a balancing parameter.

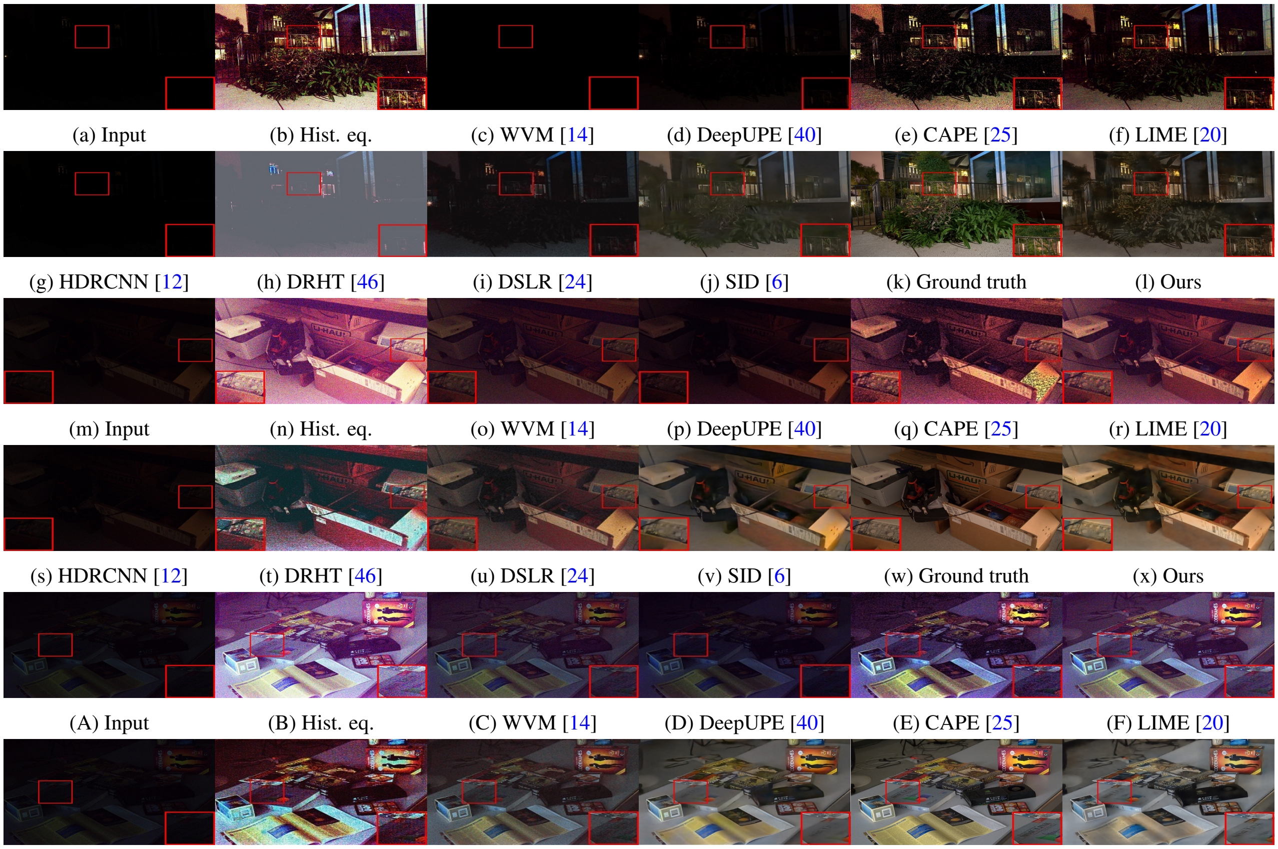
4. Experiments

We have implemented the proposed model in the Pytorch framework [35], and tested it on a PC with an i7 4GHz CPU and a GTX 1080Ti GPU. As we train our model from scratch, the network parameters are initialized randomly, except the learnable amplification ratio α, which is initialized to 1. Standard augmentation strategies, *i.e.*, scaling, cropping, and horizontal flipping, are adopted. During training, we randomly crop patches of resolutions 512×384 from the scaled images of resolution 2048×1536. For loss minimization, we adopt the ADAM optimizer [26] for 400 epochs, with an initial learning rate of $3e^{-4}$ and divided by 10 at the 250th epoch. λ_1, λ_2 and λ_3 are set to 1, 1, and 0.1, respectively. It takes $0.33s$ for the proposed network to process one image of resolutions 1024×768.

To evaluate the performance of the proposed method on enhancing low-light images, we quantitatively and visually compare our method to 9 state-of-the-art enhancement methods with available codes, including JieP [5], LIME [20], WVM [14], DSLR [24], CAPE [25], DRHT [46], DeepUPE [40], HDRCNN [12] and SID [6]. We use PSNR and SSIM for quantitative measurement.

4.1. Comparing to State-of-the-Arts

Visual comparisons. We first visually compare results of the proposed method to the state-of-the-art image enhancement methods. Figure 8 shows the results of different methods on three input low-light images (a, m, A), which were taken by a Sony camera. We can see that WVM [14] and DeepUPE [40] fail to enhance these images (c, d, o, p, C, D). Since they are based on decomposing the input image into reflectance and illumination, when an input image is of low-light, they are unable to decompose it accurately. LIME [20] can enhance the images (f, r, F), as it directly estimates the illumination map without decomposing the input image. However, it enhances both details and noise together. Similarly, the gamma correction based method CAPE [25] also jointly enhances the details and noise together (e, q, E). DRHT [46] fails to enhance the noisy low-light images (h, t, H), as noise can deteriorate both the HDR reconstruction and tone mapping processes. DSLR [24] is trained to regress a low-quality image into a high-quality one. While it can somewhat enhance the images, it fails to remove noise (i, u, I). Since the original SID [6] model



(a) Input (b) Hist. eq. (c) WVM [14] (d) DeepUPE [40] (e) CAPE [25] (f) LIME [20]
(g) HDRCNN [12] (h) DRHT [46] (i) DSLR [24] (j) SID [6] (k) Ground truth (l) Ours
(m) Input (n) Hist. eq. (o) WVM [14] (p) DeepUPE [40] (q) CAPE [25] (r) LIME [20]
(s) HDRCNN [12] (t) DRHT [46] (u) DSLR [24] (v) SID [6] (w) Ground truth (x) Ours
(A) Input (B) Hist. eq. (C) WVM [14] (D) DeepUPE [40] (E) CAPE [25] (F) LIME [20]
(G) HDRCNN [12] (H) DRHT [46] (I) DSLR [24] (J) SID [6] (K) Ground truth (L) Ours

Figure 8. Visual results of state-of-the-art methods and ours on input low-light images (a, m, A). Red boxes indicate the noisy regions where most existing methods fail. The input images were taken by a Sony camera.

(trained on raw domain) cannot be directly applied to sRGB images, we re-train it on the sRGB images. We can see that the SID model tends to remove noise and details, resulting in blurred images (j, v, J). In contrast, our results (l, x, L) show that the proposed method can successfully enhance the image content and details while suppressing noise.

Figure 9 shows results of another three input low-light images (taken by an iPhone camera). While state-of-the-art methods generally fail to remove noise and enhancing contents/details at the same time, our method produces visually more convincing results, even for the more challenging textured images (l, x). Figures 8 and 9 demonstrate the good generalization ability of the proposed model/dataset on images taken by different types of cameras.

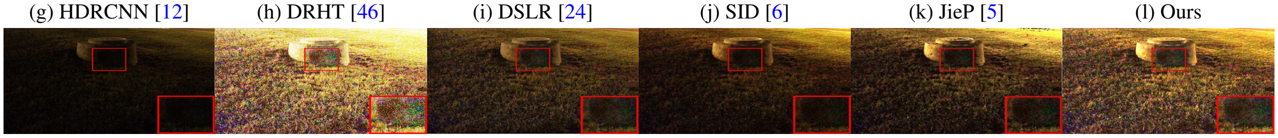
Quantitative comparisons. We have also quantitatively compared our method to the state-of-the-art enhancement methods. As shown in Table 1, the proposed method outperforms these existing enhancement methods by a large margin. Note that we have also pre-processed the input images before feeding them to two methods [14, 5], by amplifying these image pixel intensities with pre-defined ratios as in [6] or by applying histogram equalization. However, the results are the same as those without pre-processing. This indicates that enhancing noisy low-light images via decomposing images into reflectance and illumination is not suitable. In contrast, our frequency-based decomposition-and-enhancement can successfully decouple the image enhancement and denoising problem.

We also compare our method with SID [6], which was originally proposed to enhance low-light images in the raw domain, in both sRGB and raw domains. Specifically, in the sRGB domain, we apply two strategies: directly using the original SID model trained on raw images (denoted as SID), and using a retrained SID model on sRGB images in our training set (denoted as SID*). In the raw domain, we retrain our model using the raw data. We can see that our method outperforms SID [6] in both sRGB and raw domains. We further compare our method to the newest method [40] in both sRGB (retrained on our dataset) and raw domains. These results show that our model is more effective in enhancing low-light images with noise, than directly learning the image-to-image [6] or image-to-illumination [40] regression models.

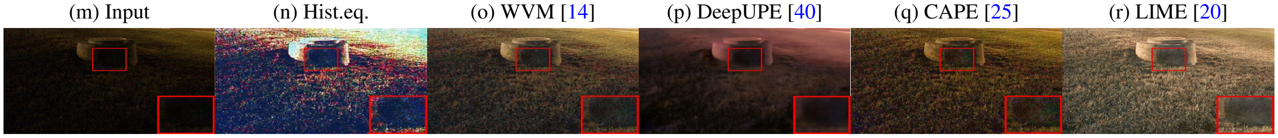
Finally, we compare our method to different combinations of existing enhancement and denoising methods. Specifically, we choose one classic denoising method B-

(a) Input (b) Hist.eq. (c) WVM [14] (d) DeepUPE [40] (e) CAPE [25] (f) LIME [20]

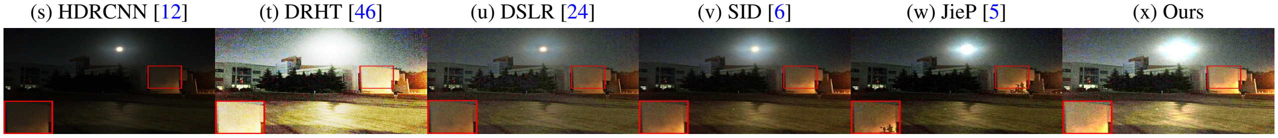
(g) HDRCNN [12] (h) DRHT [46] (i) DSLR [24] (j) SID [6] (k) JieP [5] (l) Ours



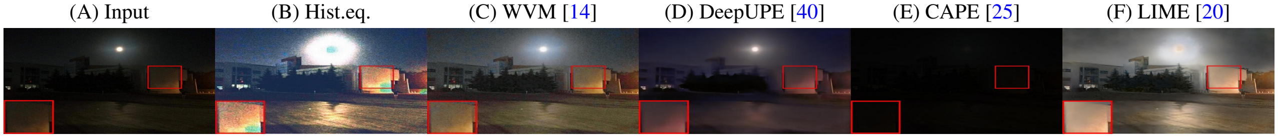
(m) Input (n) Hist.eq. (o) WVM [14] (p) DeepUPE [40] (q) CAPE [25] (r) LIME [20]



(s) HDRCNN [12] (t) DRHT [46] (u) DSLR [24] (v) SID [6] (w) JieP [5] (x) Ours



(A) Input (B) Hist.eq. (C) WVM [14] (D) DeepUPE [40] (E) CAPE [25] (F) LIME [20]

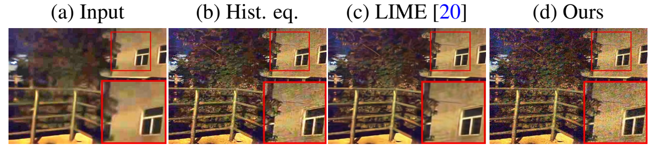


(G) HDRCNN [12] (H) DRHT [46] (I) DSLR [24] (J) SID [6] (K) JieP [5] (L) Ours

Figure 9. Visual results of state-of-the-art methods and ours on input low-light images (a, m, A). Red boxes indicate the noisy regions where most existing methods fail. The input images were taken by an iPhone camera. Results of our method in here as well as in Figure 8 demonstrate the generalization ability of the method on different camera types.

M3D [11] and one recent deep learning based denoising method xDnCNN [27] to pre-/post-process the low-light images (in the test set) before/after they are processed by enhancement method LIME [20]. We choose LIME [20] as it has the third best performance among the existing methods in Table 1. Although SID* [6] and DeepUPE* [40] have better performance, they are already trained on our dataset to remove noise. Hence, we do not use them here. Table 2 shows the results. We can see that directly applying existing denoising methods as a pre-/post-processing step to enhancement methods does not work well. As noise is already deeply buried into the image contents and details in low-light images, separately enhancing and denoising these images do not perform well. Instead, we suppress the noise in the low-frequency layer and then enhance the contents and details adaptively, producing better performances. Figure 10 shows some visual examples of combining existing enhancement and denoising methods. We can see that denoising followed by enhancement produces blurry results (e, f), due to the significant removal of image details in the denoising step. Although enhancement followed by denoising can produce relatively sharper results (g, h) in comparison to (e, f), respectively, the results are more noisy as both noise and details are enhanced in the enhancement step. It is also interesting to note that none of these methods can recover the colors (caused by noise) well, *e.g.*, the purplish color of the tree. In contrast, our method can produce a sharp image (d), with noise suppressed and color recovered.

(a) Input (b) Hist. eq. (c) LIME [20] (d) Ours



(e) BM3D+ [11] (f) xDnCNN+ [27] (g) +BM3D [11] (h) +xDnCNN [27]

Figure 10. Comparison to different combinations of LIME [20] and two denoising methods (BM3D [11] and xDnCNN [27]). "X+" indicates using LIME for post-processing, while "+X" indicates using LIME for pre-processing. Red boxes indicate the noisy regions where most existing methods fail.

Input	Method	PSNR↑	SSIM ↑
sRGB	Hist. eq.	12.08	0.2236
	CAPE [25]	15.05	0.2306
	JieP [5]	11.93	0.0381
	WVM [14]	11.95	0.0382
	DeepUPE [40]	14.44	0.2208
	DeepUPE* [40]	21.55	0.6531
	DRHT [46]	11.85	0.0969
	HDRCNN [12]	12.64	0.1102
	DSLR [24]	17.25	0.4229
	LIME [20]	17.76	0.3506
	SID [6]	15.35	0.2418
	SID* [6]	21.16	0.6398
	Ours	**22.13**	**0.7172**
RAW	SID [6]	28.88	0.7870
	DeepUPE [40]	29.13	0.7915
	Ours	**29.56**	**0.7991**

Table 1. Comparison to the state-of-the-art enhancement methods. Best performance is marked in **bold**. Note that an * indicates that the model is retrained on our sRGB traning set.

Input	Method	PSNR↑	SSIM↑
sRGB	LIME [20]	17.76	0.3506
	LIME [20] + BM3D [11]	17.90	0.3610
	LIME [20] + xDnCNN [27]	17.75	0.3511
	BM3D [11] + LIME [20]	17.41	0.3273
	xDnCNN [27] + LIME [20]	17.75	0.3511
	Ours	**22.13**	**0.7172**

Table 2. Comparison to different combinations of enhancement and denoising methods. Best performance is marked in **bold**.

4.2. Internal Analysis

We begin by studying the effectiveness of the proposed ACE module. The first two rows of Table 3 show that removing the ACE module or replacing it by a non-local block [42] causes a performance drop, as noise can no longer be filtered out via image decomposition. This verifies the effectiveness of the proposed ACE module in learning to select beneficial features and suppress harmful features before encoding the non-local contexts. Similarly, removing the CDT module also causes a performance drop, which demonstrates the importance of having a large receptive fields while bridging the gap between the low-light and enhanced domains. We further note a performance drop caused by replacing contrast-aware map C_a of the CDT modules with C_a of the ACE module, which verifies the necessity of modeling multi-level contrast-aware information for noisy low-light images. We can also see that incorporating perceptual loss leads to better results as it provides regularization in the feature space.

Finally, we study the pipeline choices. We train our model to learn to enhance images using just one stage (denoted as Single Shot). We also train our model by directly using ground truth images to supervise the output of the first stage (denoted as $I_f^{gt} \rightarrow I^{gt}$), instead of using the ground truth of the low-frequency layer. Results are shown in the

Input	Method	PSNR↑	SSIM ↑
sRGB	w/o ACE	21.34	0.6439
	ACE → NL [42]	21.49	0.6477
	w/o CDT	21.47	0.6410
	$C_a^{CDT} \rightarrow C_a^{ACE}$	21.84	0.7006
	w/o perceptual loss	22.03	0.7033
	Single Shot	21.63	0.6713
	$I_f^{gt} \rightarrow I^{gt}$	21.76	0.6874
	Ours	**22.13**	**0.7172**

Table 3. Internal analysis of the proposed method.

(a) Input (b) Hist. eq. (c) Ours

Figure 11. A failure case. When all objects in the image are far away, our method as well as existing methods may not be able to select useful contexts from the surrounding areas.

$6th$ and $7th$ rows. It shows the advantage of learning a two-stage model over Single Shot. We can also see that using ground truth of the low-frequency layer to supervise the first stage produces better results than using the ground truth images, which verifies the importance of learning the decomposition-and-enhancement model.

5. Conclusion and Future Work

In this paper, we have studied the noisy low-light image enhancement problem. We have observed that noise affects images differently in different frequency layers. Based on this observation, we propose a novel frequency-based image decomposition-and-enhancement model to adaptively enhance the image contents and details in different frequency layers, while at the same time suppressing noise. We have also presented a network with the proposed Attention to Context Encoding (ACE) module for adaptively enhancing the high and low frequency layers, and Cross Domain Transformation (CDT) module for noise suppression and detail enhancement. To train our model, we have prepared a new low-light image dataset. Finally, we have conducted extensive experiments to verify the effectiveness of our method against state-of-the-art methods.

Our method does have limitations. It may fail in scenes with small objects, in which our network may not be able to extract meaningful contextual information from the surrounding areas in order to recover the contents, as shown in Figure 11. As a future work, we are interested in extending our enhancement model to consider semantic layouts of the scenes and using generative adversarial learning for synthesizing image details.

Acknowledgement. This work was partly supported by NNSFC Grants 91748104, 61972067, 61632006, U1811463, U1908214, 61751203; and the National Key Research and Development Program of China, Grant 2018AAA0102003.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 3

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 3, 5

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 2

[4] 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. 1

[5] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *ICCV*, 2017. 2, 5, 6, 7, 8

[6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 1, 2, 4, 5, 6, 7, 8

[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 *CVPR*, 2018. 1, 2

[8] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using CNN-based single object tracker with spatial-temporal attention mechanism. In *ICCV*, 2017. 1

[9] Wikipedia contributors. Color temperature. Available from: `https://en.wikipedia.org/wiki/Color_temperature`. 5

[10] Wikipedia contributors. sRGB. Available from: `https://en.wikipedia.org/wiki/SRGB`. 2

[11] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3D filtering. In *Proc. SPIE*, volume 6064, 2006. 1, 2, 7, 8

[12] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafa Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM TOG*, 2017. 2, 5, 6, 7, 8

[13] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE TIP*, 2006. 2

[14] Xueyang Fu, Delu Zeng, Yue Huang, Xiaoping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *CVPR*, 2016. 2, 5, 6, 7, 8

[15] Michaël Gharbi, Jiawen Chen, Jonathan Barron, Samuel Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. In *SIGGRAPH*, 2017. 2

[16] A. Gijsenij, T. Gevers, and J. van de Weijer. Computational color constancy: Survey and experiments. *IEEE TIP*, 2011. 5

[17] M. Grossberg and S. Nayar. What is the space of camera response functions? In *CVPR*, 2003. 5

[18] Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *CVPR*, 2014. 2

[19] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, 2019. 1, 3

[20] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 2017. 2, 5, 6, 7, 8

[21] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE TPAMI*, 2013. 5

[22] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. In *SIGGRAPH*, 2018. 2

[23] Sung Ju Hwang, Ashish Kapoor, and Sing Bing Kang. Context-based automatic local image enhancement. In *ECCV*, 2012. 2

[24] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. DSLR-quality photos on mobile devices with deep convolutional networks. In *ICCV*, 2017. 1, 2, 5, 6, 7, 8

[25] Liad Kaufman, Dani Lischinski, and Michael Werman. Content-aware automatic photo enhancement. *Computer Graphics Forum*, 2012. 2, 5, 6, 7, 8

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

[27] Idan Kligvasser, Tamar Rott Shaham, and Tomer Michaeli. xunit: Learning a spatial activation function for efficient image restoration. In *CVPR*, 2018. 1, 2, 7, 8

[28] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void - learning denoising from single noisy images. In *CVPR*, 2019. 3

[29] Ann Lee, Kim Pedersen, and David Mumford. The complex statistics of high-contrast patches in natural images. *SCTV*, 2001. 1, 3

[30] Jianwei Li, Xiaowu Chen, Dongqing Zou, Bo Gao, and Wei Teng. Conformal and low-rank sparse representation for image restoration. In *ICCV*, 2015. 2

[31] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[32] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas Huang. Non-local recurrent network for image restoration. In *NeurIPS*. 2018. 1, 2

[33] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and SeonJoo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *CVPR*, 2016. 2

[34] Jongchan Park, Joon-Young Lee, Donggeun Yoo, and In So Kweon. Distort-and-recover: Color enhancement using deep reinforcement learning. In *CVPR*, 2018. 1, 2

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NeurIPS Workshop*, 2017. 5

[36] Stephen Pizer, E. Philip Amburn, John Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart Ter Haar Romeny, and John Zimmerman. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 1987. 2

[37] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. In *NeurIPS*. 2018. 1, 2

[38] Ramirez Rivera, Byungyong Ryu, and O Chae. Content-aware dark image enhancement through channel division. *IEEE TIP*, 2012. 2

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4

[40] Wang Ruixing, Zhang Qing, Fu Chiwing, Shen Xiaoyong, Zheng Weishi, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 2019. 1, 2, 5, 6, 7, 8

[41] Jian Sun, Nan-Ning Zheng, Hai Tao, and Heung-Yeung Shum. Image hallucination with primal sketch priors. In *CVPR*, 2003. 1, 3

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3, 8

[43] Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *ICCV*, 2017. 2

[44] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *CVPR*, 2019. 2

[45] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. 2018. 1

[46] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson Lau. Image correction via deep reciprocating HDR transformation. In *CVPR*, 2018. 1, 2, 5, 6, 7, 8

[47] Zhenqiang Ying, Ge Li, Yurui Ren, Ronggang Wang, and Wenmin Wang. A new low-light image enhancement algorithm using camera response model. In *ICCV Workshops*, 2017. 2

[48] Runsheng Yu, Wenyu Liu, Yasen Zhang, Zhi Qu, Deli Zhao, and Bo Zhang. Deepexposure: Learning to expose photos with asynchronously reinforced adversarial learning. In *NeurIPS*, 2018. 1, 2

[49] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE TIP*, 2017. 1, 2

[50] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, 2017. 1, 2

[51] Qing Zhang, Ganzhao Yuan, Chunxia Xiao, Lei Zhu, and Wei-Shi Zheng. High-quality exposure correction of underexposed photos. In *ACM MM*, 2018. 1, 2