

Unsupervised Real-world Low-light Image Enhancement with Decoupled Networks

Wei Xiong¹, Ding Liu², Xiaohui Shen², Chen Fang², and Jiebo Luo¹

¹ University of Rochester

² ByteDance Inc.

Abstract. Conventional learning-based approaches to low-light image enhancement typically require a large amount of paired training data, which are difficult to acquire in real-world scenarios. Recently, unsupervised models for this task have been explored to eliminate the use of paired data. However, these methods primarily tackle the problem of illumination enhancement, and usually fail to suppress the noises that ubiquitously exist in images taken under real-world low-light conditions. In this paper, we address the real-world low-light image enhancement problem by decoupling this task into two sub-tasks: illumination enhancement and noise suppression. We propose to learn a two-stage GAN-based framework to enhance the real-world low-light images in a fully unsupervised fashion. In addition to conventional benchmark datasets, a new unpaired low-light image enhancement dataset is built and used to thoroughly evaluate the performance of our model. Extensive experiments show that our method outperforms the state-of-the-art unsupervised image enhancement methods in terms of both illumination enhancement and noise reduction.

Keywords: Low-light image enhancement, image generation, unsupervised learning, generative adversarial networks.

1 Introduction

Real-world low-light image enhancement is a challenging task since images captured under low-light conditions usually exhibit low illumination and contain heavy noise. Enhancing these images requires adjusting contrast and illumination, as well as suppressing the noise while preserving the details simultaneously. Traditional methods for this task primarily focus on adjusting brightness or contrast via a fixed tone-mapping [28], resulting in limited performance on challenging cases. Recently, learning-based methods have been utilized to learn content-aware illumination or contrast enhancement from data with deep neural networks [9,32]. In spite of achieving satisfactory performance, they heavily rely on pairs of low-light and corresponding normal-light images, which are expensive or even impossible to obtain in real-world scenarios. One alternative way to cheaply generate such training pairs is to synthesize a low-light image from its counterpart captured under a normal light condition [27]. However, due to

the significant signal distribution gap between synthesized dark images and ones taken under real-world low-light conditions, models trained on synthesized image pairs usually fail to generalize well in real-world scenarios [11,29].

To eliminate the reliance on paired data, several unsupervised deep learning-based methods have been developed for image enhancement. As a general-purpose method, unsupervised image-to-image translation models such as CycleGAN [42] and Unit [25] can be applied to image enhancement. These methods adopt generative adversarial networks [10] to encourage the distribution of the generated images to be close to that of the target images without paired supervision. Recently, the GAN-based models have been specially designed to address the task of illumination enhancement [27,6,15]. These unsupervised learning approaches are able to generate images with better illumination and contrast in some cases. However, they have common limitations in the real-world low-light image enhancement task in two aspects: 1) the contrast and illumination of enhanced images may be unsatisfactory and usually suffer from color distortion and inconsistency; 2) these methods primarily focus on illumination enhancement, without paying specific attention to noise suppression. As a result, heavy noise still remains, and may even be magnified in the enhanced images. Although a few prior works have been developed to remove noise from images without any ground-truth supervision, they either use traditional methods such as BM3D [7] that requires a given noise level as an input, or use learning-based denoising methods that are originally designed for synthesized noise such as additive white Gaussian noise [21,18,1]. In addition, the illumination and contrast are usually adjusted differently in regions with different lighting conditions, making the noise level spatially variant as well in the enhanced images. As a result, directly applying these denoising methods to real-world low-light images leads to poor results with either regions with remaining noise or over-smoothed ones.

To address these issues, we propose to decouple the whole unsupervised enhancement task into two sub-tasks: 1) illumination enhancement, and 2) noise suppression. We propose a two-stage framework to handle each sub-task respectively, where the enhanced results in illumination enhancement are used to guide the learning of spatially adaptive denoising. Specifically, in Stage I, a Retinex-based deep network is trained under a GAN framework in an unsupervised manner to enhance the illumination of low-light images while preserving the contextual details. A pyramid module [41] is customized and embedded in the generator network to enlarge its receptive field, and is shown to effectively mitigate the color distortion in results. In Stage II, unlike the conventional learning-based denoising models, we take the original low-light image, the enhanced image from Stage I, and an illumination mask indicating the enhanced illumination of the low-light image as inputs, and propose an unsupervised learning-based denoising model in another GAN framework to remove noise and generate the final enhanced image. With the image pairs as input, our denoising model in Stage II is explicitly guided by both the original lighting condition and enhanced illumination, and is capable of adaptively removing noise. Moreover, we design an adaptive content loss and construct pseudo triples, i.e., a low-light image, an

image after illumination enhancement, and a corresponding noise-free normal-light image with the same content, to facilitate learning the noise patterns and training this GAN framework without ground-truth supervision.

We evaluate the performance of our proposed approach over the LOw-Light (LOL) dataset [33] and an unpaired enhancement dataset from [15]. To further demonstrate the effectiveness of our method, we contribute an Unpaired Real-world Low-light image enhancement dataset (URL) for evaluation. Our dataset is composed of 1) low-light images captured under real-world low-light conditions with varying levels of noise, and 2) normal-light images collected from existing data galleries, which consist of diverse scenes ranging from outdoor scenes to indoor pictures. We compare our method with the state-of-the-art unsupervised learning-based enhancement methods on these datasets. Extensive experiments show that our method outperforms other methods in terms of both illumination enhancement and noise suppression.

In summary, our primary contributions are:

- We propose a decoupled framework for low-light image enhancement in a fully unsupervised manner;
- We facilitate unsupervised learning of our denoising model by constructing pseudo triples and propose an adaptive content loss to denoise regions guided by both the original lighting condition and enhanced illumination;
- We contribute an unpaired low-light image enhancement dataset containing varying noise and good diversity as an important complement to existing low-light enhancement datasets.

2 Related Work

2.1 Image Contrast Enhancement

Traditional image enhancement methods are primarily built upon histogram equalization (HE) or Retinex theory [19]. HE-based methods aim to adjust the histogram of pixel intensities to obtain an image with better contrast [30]. Retinex-based methods assume that an image is the composition of illumination map and reflectance, and thus low-light images can be restored by estimating the illumination map and reflectance map [16].

The performance of traditional methods may not be satisfactory enough especially in challenging scenarios. Recently, learning-based methods have been proposed to learn the contrast enhancement from data [8,12,22]. Later deep neural networks have been used and achieved promising results [27,9,33,4]. Due to the difficulty of acquiring paired data in real-world scenarios, several weakly supervised and unsupervised enhancement approaches have been proposed. Ignatov et al. propose a transitive GAN-based enhancement model that can be learned without paired data [14]. Chen et al. propose an unpaired model based on 2-way GANs to enhance images [6]. Jiang et al. further propose an unsupervised deep network which is specifically designed for low-light image enhancement [15].

2.2 Real-world Image Denoising

There have been a number of works for image denoising, including conventional methods such as BM3D [7] and Non-local means [2], and deep learning-based models such as DnCNN [38], Residual Dense Networks [40] and Non-local Recurrent Networks [24]. However, most of the models are limited to synthetic noise removal. Models trained with synthetic noise are difficult to generalize to real-world noise removal, since the distribution of real-world noise is different from the synthetic noise. To address this issue, real-world blind denoising models have been proposed. Xu et al. [36] design a multi-channel weighted nuclear norm minimization model to use channel redundancy. Guo et al. propose CBDNet [11] to directly learn a blind denoiser from real-world paired data. Kim et al. leverage a GAN based deep network for real-world noise modeling [17]. Other approaches [20,39,35] also show promising results.

Similar to image enhancement, most learning-based denoising models need to be trained with paired data, which is expensive to obtain for real-world noise removal task. Recently, several unsupervised denoising methods have been devised, including self-supervised learning approaches, such as Noise2Noise [21] and Noise2Void [18], as well as unpaired training approaches [5,37].

2.3 Unsupervised Image-to-Image Translation

The general-purpose unsupervised image-to-image translation methods can be applied to low-light image enhancement by learning a mapping from the low-light input to the enhanced images [25,42]. Usually a GAN-based model is adopted to encourage the generated images to be as realistic as the normal-light clean images. However, for real-world low-light image enhancement, conventional unsupervised models usually suffer from color distortion or perform poor noise removal.

3 Our Approach

As shown in Fig. 1, our approach for real-world low-light image enhancement consists of two stages. In the first stage, we perform illumination enhancement on the real-world low-light images while preserving contextual details. We adopt a Retinex-based network to predict the illumination map from the low-light input image, then generate the enhanced result. To preserve details from the input image, we constrain the enhanced image to be perceptually similar to the input image. Note that the output image may still contain noise. In the second stage, we propose an unsupervised learning-based denoising network to suppress the noise in the output image from the first stage as well as enhancing the contextual details. The denoising network takes the original low-light image, the enhanced image of stage I and an illumination mask as inputs, and outputs the final denoised image. The illumination mask indicates how much the illumination is improved in Stage I, which will be formulated later.

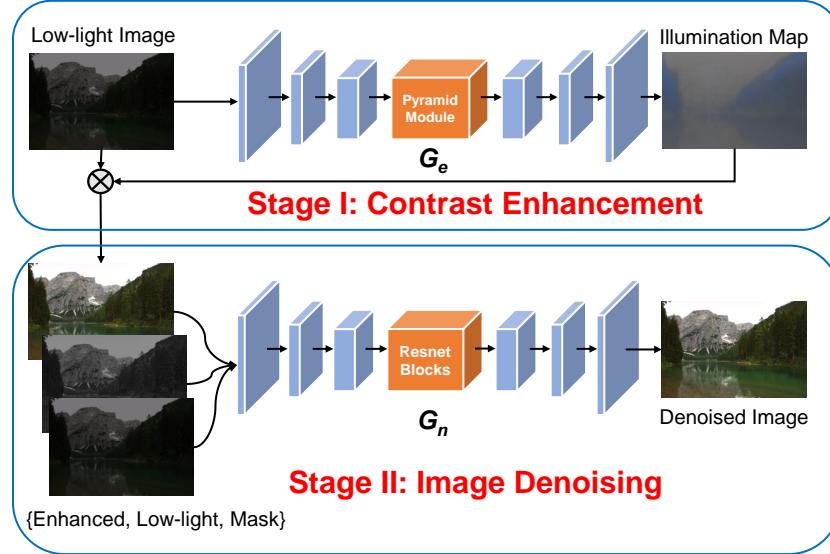


Fig. 1. An overview of our proposed decoupled network, which is composed of two stages. In Stage I, given a low-light image, we learn a deep network to predict an illumination map, then use Eq. 1 to obtain the contrast enhanced image, as illustrated with notation \otimes . Then in Stage II, we input the original low-light image, the enhanced image from Stage I, as well as the illumination mask, to generate an image with reduced noise and better details.

3.1 Stage I: Illumination Enhancement

Model architecture. Given a noisy low-light image I , our goal in this stage is to learn a model G_e to output an enhanced image \hat{I} . A straight forward solution is to use an image-to-image translation model to learn a mapping from input to the output. Conventional image translation models usually adopt a U-net [31] like architecture to directly predict the enhanced image from the low-light input image. However, under the unsupervised learning scheme, directly applying such architecture is easy to produce results with unstable illumination, such as color distortion or inconsistency [15].

To learn a more generalized model, we adopt a Retinex-based model to enhance the low-light images. Based on the Retinex theory, a low-light image I can be modeled as $I = S \circ R$, where S is the illumination, \circ denotes element-wise multiplication, and R is the reflectance. Similar to [33,32], we regard the reflectance as a well-exposed image \hat{I} , then we have $I = S \circ \hat{I}$. In reverse, the enhanced image \hat{I} can be recovered from the low-light image I given the predicted illumination map S , with the following equation:

$$\hat{I} = I / S. \quad (1)$$

where $/$ denotes element-wise division, the illumination map S is predicted by a deep neural network G_e and thus $S = G_e(I)$. As shown in Fig. 1, the input

low-light image is passed to the encoder, then an enhancing module, and then decoded by the decoder into an illumination map with RGB three channels. The enhanced image can then be obtained using Equation 1. The Retinex-based model has the advantage over the conventional image-to-image translation models in that the illumination maps for natural images usually have relatively simple forms, as indicated in [32], which facilitates the learning process, leading to a model with better generalization ability.

Besides the Retinex-based model, we customize a pyramid module [41] and embed it between the encoder and decoder of the network, in order to enlarge the receptive field of our network. In our module, we pool the feature maps into features at multiple resolutions, namely, 1×1 , 2×2 , 4×4 , 16×16 . At each resolution, feature maps are followed by a convolution layer and a ReLU[34] activation layer. Then the transformed feature maps are upsampled and concatenated, then fed into the following convolution layer. The fused feature maps are then decoded by the decoder to generate the illumination map S . With the pyramid module, the receptive field of the model is enlarged, and the network can perceive the illumination information at different spatial levels, which is beneficial to the enhancement. The detailed architecture of the encoder and decoder and the discriminator can be found in the supplementary material.

Loss Functions. In our work, G_e is trained with unpaired images. To achieve this goal, we adopt adversarial learning to encourage the distribution of the enhanced image \hat{I} to be close to that of the normal-light images. Specifically, we use two discriminators to distinguish the generated image from the real normal-light image. The global discriminator D_g takes the whole image as the input, and outputs the realness of the image. The local discriminator D_l takes random patches extracted from the image and outputs the realness of each patch. The global discriminator encourages the global appearance of the enhanced image to be similar to a normal-light image, while the local discriminator ensures that the local context (shadow, local contrast, highlight, etc.) can be as realistic as the real normal-light images.

We adopt the LSGAN version of relativistic average GAN loss for training the global discriminator. When updating the discriminator, we have:

$$L_D^g = \mathbb{E}_{x_r \in \mathbb{P}}[(D_g(x_r) - \mathbb{E}_{x_f \in \mathbb{Q}} D_g(x_f) - 1)^2] + \mathbb{E}_{x_f \in \mathbb{Q}}[(D_g(x_f) - \mathbb{E}_{x_r \in \mathbb{P}} D_g(x_r))^2] \quad (2)$$

When updating the generator, we have:

$$L_G^g = \mathbb{E}_{x_r \in \mathbb{P}}[(D_g(x_r) - \mathbb{E}_{x_f \in \mathbb{Q}} D_g(x_f))^2] + \mathbb{E}_{x_f \in \mathbb{Q}}[(D_g(x_f) - \mathbb{E}_{x_r \in \mathbb{P}} D_g(x_r) - 1)^2] \quad (3)$$

where \mathbb{P} and \mathbb{Q} are the real image (the normal-light images) distribution and the generated image distribution, respectively. x_r and x_f are samples from distribution \mathbb{P} and \mathbb{Q} , respectively.

We adopt LSGAN loss for training the local discriminator. When updating the discriminator, we have:

$$L_D^l = \mathbb{E}_{x_r \in \mathbb{P}}[(D_l(x_r) - 1)^2] + \mathbb{E}_{x_f \in \mathbb{Q}}[(D_l(x_f))^2] \quad (4)$$

When training the generator, we have:

$$L_D^l = \mathbb{E}_{x_f \in \mathbb{Q}}[(D_l(x_f) - 1)^2] \quad (5)$$

Similar to [15], we use a perceptual loss (computed on VGG features) between the output image and the input image to preserve content details from the input image. To prevent the output image from being as dark as the input image, we alleviate the influence of image brightness and force the network to focus on only the content preservation by using instance normalization on the VGG feature maps before performing the perceptual loss. Similar to the adversarial loss, we calculate perceptual loss on both the whole image and the image patches. We formulate the global perceptual loss L_P^g and the local version L_P^l as:

$$L_P^g = \|\Phi(I) - \Phi(\hat{I})\|_2^2/N, \quad (6)$$

$$L_P^l = \|\Phi(I_p) - \Phi(\hat{I}_p)\|_2^2/N, \quad (7)$$

where Φ denotes to VGG feature extractor, N is the number of elements in the image, I_p and \hat{I}_p are random patches extracted from I and \hat{I} , respectively.

By minimizing both the adversarial losses and the perceptual losses, we are able to learn a good illumination predictor and produce results without color distortion and preserve contextual details.

3.2 Stage II: Noise Suppression

In Stage I, in order to preserve the contextual details of the low-light image, strong perceptual loss has been imposed to train the model, and the noise may remain in the enhanced result. In dark regions, the noise may even be amplified due to the increase of brightness. To suppress noise and enhance the contextual details, in Stage II, we propose an unsupervised learning-based denoising model via generative adversarial networks that can adaptively remove noise in the enhanced images, as depicted in Fig. 1.

Our noise suppression model G_n adopts the original input image I , the enhanced image \hat{I} in Stage I, and an illumination mask M as inputs, and outputs the final image \tilde{I} that is clean and with enhanced illumination, i.e., $\tilde{I} = G_n(I, \hat{I}, M)$. M serves as an indicator showing the enhanced illumination from low-light image I to contrast enhanced image \hat{I} . We have:

$$M = \max(\text{illu}(\hat{I}) - \text{illu}(I), 0) \quad (8)$$

where $\text{illu}(\cdot)$ means extracting the illumination of an image. In our work, we directly use the gray-scale image as the illumination map. With the enhanced illumination M , the low-light image I and the contrast enhanced image \hat{I} , our denoising model is explicitly guided by the illumination conditions.

Model Architecture. Our denoising model G_n in this stage adopts an encoder-decoder architecture, with several convolutional blocks followed by several resnet blocks then decoded back to an image. We adopt a multi-scale discriminator [25] to predict the realness of images at multiple resolutions. The detailed architecture of the generator and discriminator can be found in the supplementary material.

Loss Functions. Since there is no ground-truth image for the input low-light image, to learn a noise-free image from its noisy counterpart, we adopt a LSGAN-based adversarial loss to encourage the generated image to be as clean as the real-world clean normal-light images. Note that the discriminator needs not only to judge whether the illumination and contrast of the generated image are realistic enough, but also need to judge whether the generated image is clean without any noise. Simply training a single discriminator with clean images or synthesized images to do both tasks is difficult. As our goal in this stage is noise suppression and detail enhancement, we aim to keep the color and brightness of the enhanced image from Stage I. Therefore, when feeding the discriminator, we first perform an instance normalization on both the synthesized image and the normal-light clean image to reduce the influence of image illumination, color and contrast.

During training, we randomly match the output image \tilde{I} to a normal-light clean image I_c . Our adversarial loss for training the discriminator D_n is:

$$L_D^n = \mathbb{E}_{I_c \in \mathbb{P}}[(D_n(Ins(I_c)) - 1)^2] + \mathbb{E}_{\tilde{I} \in \mathbb{Q}}[(D_n(Ins(\tilde{I}))^2)] \quad (9)$$

The corresponding loss for updating the generator is:

$$L_G^n = \mathbb{E}_{\tilde{I} \in \mathbb{Q}}[(D_n(Ins(\tilde{I})) - 1)^2] \quad (10)$$

where \mathbb{P} and \mathbb{Q} are the normal-light clean image distribution and the generated image distribution in Stage II, respectively. $Ins(\cdot)$ denotes to instance normalization.

Mere using the adversarial loss can cause color shifting problem, i.e., the color of the generated images can be easily distorted, since we only constrain the images after instance normalization to be similar to the normal-light images. As we have already obtained an image with a satisfactory contrast and color, in this stage, we only need to preserve the contrast and color from \hat{I} . Therefore, we use a color loss to constrain the generated image \tilde{I} to have the same color as the \hat{I} . Specifically, we first down-sample the images with average pooling to \tilde{I}_p^\downarrow and \hat{I}_p^\downarrow , then do the color matching in order to increase the robustness to heavy noise. We have:

$$L_{color} = \sum_p \angle(\tilde{I}_p^\downarrow, \hat{I}_p^\downarrow) / N_n \quad (11)$$

where p is the location of a pixel in the down-sampled image, $\angle(x, y)$ calculates the inner product between two 3-D vectors which are composed of RGB channels of a pixel location, N_n is the number of pixels in the down-sampled image.

Constructing Pseudo Triples for Self-supervised Learning. We estimate the noise in image \hat{I} as $I_n = \hat{I} - \tilde{I}$. Then given a randomly matched normal-light clean image \tilde{J} , we can simulate a pseudo noisy image \hat{J} by adding the estimated noise to the clean image, i.e.,

$$\hat{J} = \tilde{J} + I_n. \quad (12)$$

Since the input of our network G_n is a combination of an enhanced image and the low-light image, when constructing the pseudo noisy image, we also need a low-light version of the pseudo noisy image. To this end, we use gamma correction to decrease the brightness of \hat{J} , to obtain a pseudo low-light image J .

$$J = (\hat{J})^\lambda \quad (13)$$

where λ is estimated as $\lambda = \log \bar{I} / \log \tilde{I}$. \bar{I} and \tilde{I} are the average pixel value over all pixel locations of image I and \tilde{I} , respectively.

Similarly, we construct the illumination mask for the pseudo data as $M_J = \max(\text{illu}(\hat{J}) - \text{illu}(J), 0)$. We then input the simulated low-light image J , normal-light image (with noise) \hat{J} and illumination mask M_J to our network G_n , in order to predict the denoised image $J_c = G_n(J, \hat{J}, M_J)$.

Adaptive Content Loss for Pseudo Triples. To train the network, we adopt a content loss to constrain the generated image J_c to be perceptually close to the real clean image \tilde{J} . This is achieved by using both perceptual loss and L1 reconstruction loss on the pixel space between J_c and \tilde{J} . As different regions may have different lighting conditions, regions with significant brightness increase after the first stage may contain heavy noise, and regions without large brightness increase may contain less noise. When imposing the reconstruction, we encourage the network to focus more on dark regions where noise is usually heavier. We then formulate the adaptive content loss as:

$$L_C^{pseudo} = \sum_i \|M_J^i \circ (\Phi_i(J_c) - \Phi_i(\tilde{J}))\|_2^2 / N_i + \gamma_p \|M_J \circ (J_c - \tilde{J})\|_1 / N, \quad (14)$$

where M_J^i is the downsized version of M_J to match the spatial size of the VGG features. M_J^i serves as the weight mask for the feature matching loss at the i -th VGG layer. M_J serves as the weight mask for each pixel in the image. N is the number of elements in image J_c , N_i is the number of elements in the feature maps of the i -th layer in the VGG. $\Phi_i(I)$ is the feature in the i -th layer of VGG given the input image I . γ_p is the weight to balance the losses from the RGB image domain and the VGG feature domain. In our work, we choose the layers of “relu1_2”, “relu2_2”, “relu3_2”, “relu4_4”, “relu5_4” to perform both low-level and high-level feature matching, and γ_p is set as 10. We do not use instance normalization on the VGG feature maps, since we need preserve the color and contrast.

Content loss on real images. In order to make sure that the generated image \tilde{I} preserves contextual details of the input image \hat{I} , we also impose a perceptual loss as well as a reconstruction loss between the real images \tilde{I} and \hat{I} .

$$L_C^{real} = \sum_i \|\Phi_i(\hat{I}) - \Phi_i(\tilde{I})\|_2^2 / N_i + \gamma_c \|J_c - \tilde{J}\|_1 / N \quad (15)$$

where the layers and operations used are the same as L_C^{pseudo} . γ_c is a weight balance term similar to γ_p and set as 10 in our work.

The total loss for training G_n is a combination of all the losses.

$$L_{total} = L_G^n + \lambda_c L_{color} + \lambda_C^p L_C^{pseudo} + \lambda_C^r L_C^{real} \quad (16)$$

and we empirically find that setting $\lambda_c, \lambda_C^p, \lambda_C^r$ as 10, 1, 1, respectively, yields the best denoised result.

4 Experiments

In this section, we first compare the performance of each model with respect to only contrast/illumination enhancement on an unpaired enhancement dataset from EnlightenGAN [15]. Then we report results after both contrast enhancement and noise suppression on the challenging LOL dataset [33] and our collected unsupervised real-world low-light dataset (URL dataset), which both contain low-light images with noticeable noise. We include more experiment results in the supplementary materials.

4.1 Datasets

Unpaired Enhancement Dataset: Jiang et al. [15] collect an unpaired dataset for training contrast enhancement models. The training set is composed of 914 low-light images which are dark yet containing no significant noise, and 1016 normal-light images from public datasets. We use this dataset to compare the performance of contrast enhancement of each model. The evaluation set is composed of 148 low-light/normal-light image pairs from public datasets. All the images from both the training and evaluation sets have been resized to 400×600 .

LOw-Light (LOL) Dataset [33]: LOL dataset is composed of 500 low-light and normal-light image pairs and divided into 485 training pairs and 15 testing pairs. The low-light images contain noise produced during the photo capture process. Most of the images are indoor scenes. To adapt the dataset to our unsupervised setting, we adopt the 485 training images as our low-light train set, and adopt the normal-light images in the Unpaired Enhancement Dataset [15] as the normal-light train set. The testing images remain the same as the LOL dataset. All the images have a resolution of 400×600 .

URL Dataset: There are a quite limited number of real-world low-light datasets publicly available. Among the public low-light datasets, some of them are composed of synthetic images while many other datasets such as ExDark [26] or Adobe FiveK [3] contain dark images without significant noise. Therefore, these datasets do not meet the objective of our study. Therefore, we collect an Unsupervised Real-world Low-light dataset (URL dataset) composed of 414 real-world low-light images taken by iPhone-6s and 3,837 normal-light images selected from Adobe FiveK. To collect the low-light images, we first take photos with an

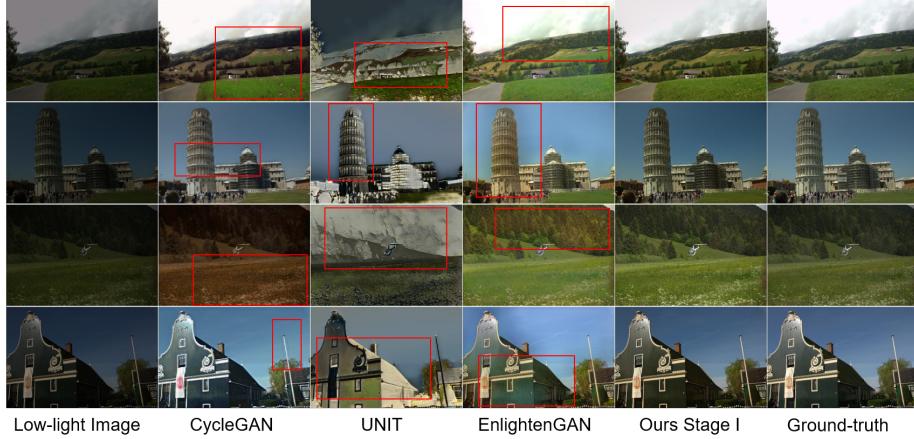


Fig. 2. Contrast enhancement results on Unpaired Enhancement Dataset from [15]. Please pay attention to the regions in the red boxes, where severe color distortion and contrast inconsistency take place.

Table 1. Quantitative results for contrast enhancement on Unpaired Enhancement Dataset. The best results are shown in bold.

Model	CycleGAN	UNIT	EnlightenGAN	Ours Stage I
PSNR	18.22	8.42	17.31	19.78
SSIM	0.7284	0.2549	0.8047	0.8197

iPhone-6s from various scenes in different cities around the world. We remove images that are too dark (cannot be recovered since details are lost), blurry, or with a high brightness. We also remove images that are very similar to other images in order to boost the diversity of the dataset. In the end, we are able to select 414 low-light images from over 4,000 photos. Our URL dataset is quite diverse, containing various scenes from both indoor and outdoor, and under different light conditions. Consequently, the level of noise contained in each image or even different regions of the same image varies considerably across the dataset. We divide the low-light images into 328 training images and 86 testing images. This dataset thus complements the existing datasets in those two regards. Note that there is no corresponding ground-truth image for each test image. Each low-light image is resized to 1008×756 .

4.2 Implementation Details

We use the Adam optimizer with a learning rate of 0.0001 and a batch size of 32 for training both stages. We train Stage I for 200 epochs using randomly cropped patches with size 320×320 , as our model needs to perceive the global information. Stage II is trained for 1,000 epochs with randomly cropped patches with size 128×128 , as our model in this stage primarily needs to capture the noise pattern in the local regions. The model architectures and other configuration details will be put in the supplementary materials.

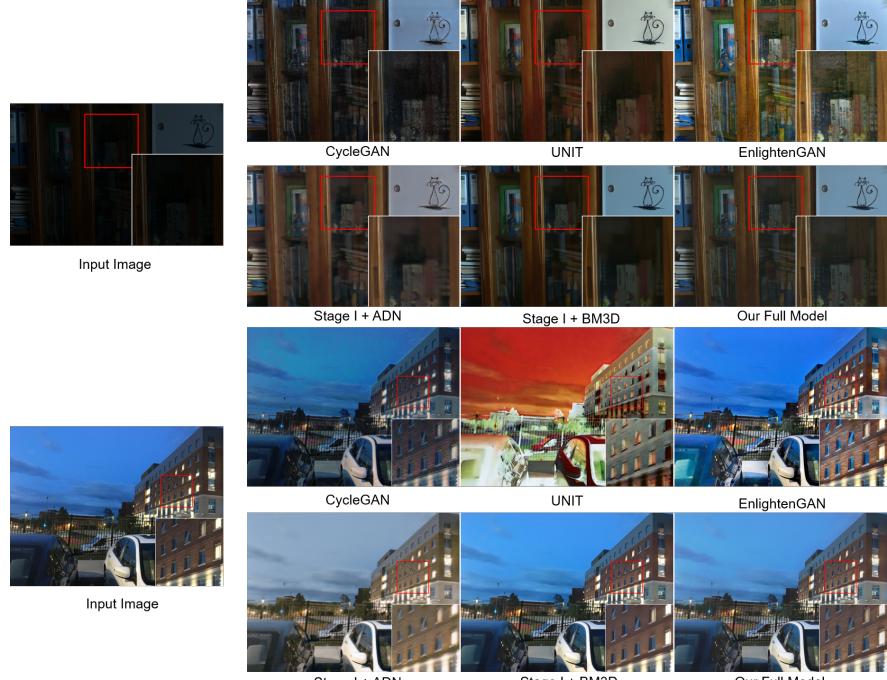


Fig. 3. Real-world low-light enhancement results (with denoising) on the LOL dataset (top) and URL dataset (bottom).

4.3 Experiments for Contrast Enhancement

We compare our Stage I model which does only contrast enhancement on the Unpaired Enhancement Dataset [15] with state-of-the-art models, including CycleGAN, UNIT and EnlightenGAN. All the models are trained on the training set, and evaluated on the evaluation set. As shown in Fig. 2, our model can generate normal-light images with reasonable contrast and color in both global and local regions. EnlightenGAN can produce visually pleasing images. However, they may still suffer from color distortion in several local regions, as indicated by the red boxes. The other unsupervised image translation methods can synthesis roughly good images. However, the color, contrast are not perfect.

We report the PSNR and SSIM of the generated images as a complementary to the visual results on Unpaired Enhancement Dataset. Results in Table. 1 show that our model performs better than the existing models, which are consistent with the visual results.

4.4 Experiments for Real-world Low-light Image Enhancement

Experiment Settings. We evaluate our full decoupled networks on the real-world low-light image enhancement datasets: LOL and our URL datasets, and compare it with the state-of-the-art unsupervised image translation or contrast enhancement models, including CycleGAN [42], UNIT [25] and EnlightenGAN [15]. We also compare our model with the combination of contrast enhancement

Table 2. Results on the LOL dataset.

Model	PSNR	SSIM
CycleGAN	14.75	0.6852
UNIT	15.49	0.7280
EnlightenGAN	18.36	0.7839
Stage I + BM3D	19.36	0.8154
Stage I + ADN	17.72	0.7776
Our Full Model	20.04	0.8216

Table 3. Configuration of different versions of our model.

Model	L_G^n	L_{color}	L_C^{pseudo}	L_C^{real}
Version 1	✓			
Version 2	✓	✓		
Version 3	✓	✓	✓	
Full Model	✓	✓	✓	✓
Plain	✓	✓	vanilla	✓

model and denoising model. Specifically, we compare our full model with our Stage I + BM3D and our Stage I + ADN [13,23]. BM3D [7] is a robust image denoising method. The limitation is that it requires a known noise level as input. To use BM3D in our task, we estimate a rough noise level for each test image, then apply BM3D on the test images. ADN is a recent work which proves to be effective for unsupervised artifact removal.

Qualitative results. Fig. 3 show the qualitative results on both the LOL dataset and our URL dataset. CycleGAN generates heavy artifacts and slightly suffer from color distortion. UNIT suffers from heavy color distortion on our dataset and cannot preserve details on LOL dataset. EnlightenGAN is able to improve the illumination of the images, but there are still noise and artifacts on the image. The visual results indicate that it is challenging to handle image contrast/illumination enhancement and denoising simultaneously with a single model for real-world low-light image enhancement.

However, simply cascading an contrast enhancement model with a denoising model still cannot produce satisfactory results. From Fig. 3, we see that the results of ADN still exhibit color distortion on our dataset and over-smoothing on LOL dataset. A possible reason may be that existing models primarily focus on denoising and pay less attention to maintaining a reasonable color of the image. Using BM3D to post-process the results of our Stage I yields good contrast and illumination. However, From Fig. 3, we observe that the highlighted regions on both LOL and URL are over-smoothed. Many content details are missing. A possible reason is that BM3D was proposed for synthetic noise removal, it may not generalize well to real-world denoising. Compared to existing methods, our full model learns to perform noise removal adaptively and considers preserving the details and color when denoising. Therefore, It is able to suppress noise as well as preserving more details. Results on Fig. 3 demonstrate the effectiveness of our method.

On LOL dataset, we also report PSNR and SSIM results in Table. 2. From Table. 2 we can see that our model is consistently better than the existing models, further demonstrating the superiority of our decoupled networks. We will supplement more results of other methods in the supplementary materials.

4.5 Ablation Study

In this section we study how each component of our model contribute to the final performance. We primarily analyze the components in our Stage II, which

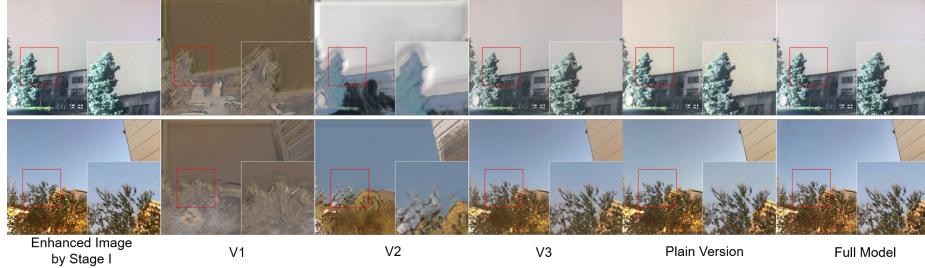


Fig. 4. Ablation Study on our URL dataset.

are the core contribution and play key roles in this work. Specifically, as shown in Table. 3, we compare the versions of our Stage II model with different losses imposed. Besides the versions regarding the loss functions, we also compare our model to a non-adaptive denoising model, which we call the Plain Model. Our full model is designed to remove noise adaptively, i.e., perform stronger noise suppression on regions where the brightness is significantly promoted after the contrast enhancement, and perform weaker noise suppression on regions where the brightness is not largely increased, since these regions may be originally brighter and therefore contain less noise. To validate the effectiveness of our approach, we compare our model in Stage II to a plain version, i.e., the generator only takes the enhanced image \hat{I} as input, and outputs the final denoised model. The four losses used in our full model remain the same, except we use the vanilla perceptual loss instead of adaptive content loss.

Experiments are conducted on our URL dataset and Fig. 4 shows the results. From the visual results we conclude that merely using adversarial loss (version 1) can help to smooth the image, but the color is shifting and cannot be well controlled. Using the color loss (version 2) does significantly help to constrain the output image to have the same color and contrast of the input image. However, without learning with pseudo data (i.e., loss L_C^{pseudo}), the content of the output is not explicitly controlled. As a result, this version does not learn any useful contents. When imposing the learning with pseudo triples (version 3), the network is able to produce good contents as well as performing noise suppression. However, several local regions still lack details. Please pay attention to the trees in these results in Fig. 4. The textures of the trees and leaves are smoothed in Version 3, and preserved in our full model, indicating that the content loss between the real input image and its output image further helps to preserve contextual details during denoising.

Compared our full model in Stage II to the Plain Model, we observe that the Plain Model cannot well suppress the noise. There are still notable noise on the sky and other places, as shown in Fig. 4. A possible reason is that the plain model may not be able to precisely perceive all the noise patterns under different illumination conditions. On the contrary, our model is explicitly guided by the illumination of the image, therefore it can capture the noise pattern under various illumination conditions more effectively and produce better results.

5 Conclusion

We propose decoupled networks to address the real-world low-light image enhancement problem. The model in Stage I is able to enhance the low-light image to generate an image with satisfactory contrast and color. The model in Stage II further denoises the enhanced image so that the final image is as clean as real-world normal-light clean images, while preserving good contrast, color and contextual details. We conduct experiments on three real-world datasets and show that our model outperforms the state-of-the-art models with respect to both contrast enhancement and image denoising.

References

1. Batson, J., Royer, L.: Noise2self: Blind denoising by self-supervision. arXiv preprint arXiv:1901.11365 (2019)
2. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: CVPR. vol. 2, pp. 60–65. IEEE (2005)
3. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input / output image pairs. In: The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition (2011)
4. Chen, C., Chen, Q., Xu, J., Koltun, V.: Learning to see in the dark. In: CVPR. pp. 3291–3300 (2018)
5. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: CVPR. pp. 3155–3164 (2018)
6. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6306–6314 (2018)
7. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on image processing **16**(8), 2080–2095 (2007)
8. Fu, X., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. In: CVPR. pp. 2782–2790 (2016)
9. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
11. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. In: CVPR. pp. 1712–1722 (2019)
12. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on Image Processing **26**(2), 982–993 (2016)
13. Haofu Liao, Wei-An Lin, J.Y.S.K.Z.J.L.: Artifact disentanglement network for unsupervised metal artifact reduction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2019)
14. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Wespe: weakly supervised photo enhancer for digital cameras. In: CVPR Workshops. pp. 691–700 (2018)

15. Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlightengan: Deep light enhancement without paired supervision. arXiv preprint arXiv:1906.06972 (2019)
16. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image processing **6**(7), 965–976 (1997)
17. Kim, D.W., Ryun Chung, J., Jung, S.W.: Grdn: Grouped residual dense network for real image denoising and gan-based real-world noise modeling. In: CVPR Workshops. pp. 0–0 (2019)
18. Krull, A., Buchholz, T.O., Jug, F.: Noise2void-learning denoising from single noisy images. In: CVPR (2019)
19. Land, E.H.: The retinex theory of color vision. Scientific american **237**(6), 108–129 (1977)
20. Lebrun, M., Colom, M., Morel, J.M.: Multiscale image blind denoising. IEEE Transactions on Image Processing **24**(10), 3149–3161 (2015)
21. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. In: ICML. pp. 2965–2974 (2018)
22. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. IEEE Transactions on Image Processing **27**(6), 2828–2841 (2018)
23. Liao, H., Lin, W., Zhou, S.K., Luo, J.: Adn: Artifact disentanglement network for unsupervised metal artifact reduction. IEEE Transactions on Medical Imaging (2019). <https://doi.org/10.1109/TMI.2019.2933425>
24. Liu, D., Wen, B., Fan, Y., Loy, C.C., Huang, T.S.: Non-local recurrent network for image restoration. In: Advances in Neural Information Processing Systems. pp. 1673–1682 (2018)
25. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems. pp. 700–708 (2017)
26. Loh, Y.P., Chan, C.S.: Getting to know low-light images with the exclusively dark dataset. Computer Vision and Image Understanding **178**, 30–42 (2019)
27. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)
28. Mantiuk, R., Daly, S., Kerofsky, L.: Display adaptive tone mapping. In: ACM SIGGRAPH 2008 papers, pp. 1–10 (2008)
29. Martin, C.H., Mahoney, M.W.: Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. arXiv preprint arXiv:1710.09553 (2017)
30. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
32. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: CVPR. pp. 6849–6857 (2019)
33. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. In: BMVC (2018)

34. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
35. Xu, J., Zhang, L., Zhang, D.: A trilateral weighted sparse coding scheme for real-world image denoising. In: ECCV. pp. 20–36 (2018)
36. Xu, J., Zhang, L., Zhang, D., Feng, X.: Multi-channel weighted nuclear norm minimization for real color image denoising. In: ICCV. pp. 1096–1104 (2017)
37. Yan, H., Tan, V., Yang, W., Feng, J.: Unsupervised image noise modeling with self-consistent gan. arXiv preprint arXiv:1906.05762 (2019)
38. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Transactions on Image Processing **26**(7), 3142–3155 (2017)
39. Zhang, K., Zuo, W., Zhang, L.: Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. IEEE Transactions on Image Processing **27**(9), 4608–4622 (2018)
40. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
41. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
42. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017)