Enhance Images as You Like with Unpaired Learning

Xiaopeng Sun[1*], **Muxingzi Li**[2*†], **Tianyu He**[2] and **Lubin Fan**[2]
[1]Xidian University
[2]Alibaba Group
xpsun@stu.xidian.edu.cn, {muxingzi.lmxz,timhe.hty}@alibaba-inc.com, lubinfan@gmail.com

Abstract

Low-light image enhancement exhibits an ill-posed nature, as a given image may have many enhanced versions, yet recent studies focus on building a deterministic mapping from input to an enhanced version. In contrast, we propose a lightweight one-path conditional generative adversarial network (cGAN) to learn a one-to-many relation from low-light to normal-light image space, given only sets of low- and normal-light training images without any correspondence. By formulating this ill-posed problem as a modulation code learning task, our network learns to generate a collection of enhanced images from a given input conditioned on various reference images. Therefore our inference model easily adapts to various user preferences, provided with a few favorable photos from each user. Our model achieves competitive visual and quantitative results on par with fully supervised methods on both noisy and clean datasets, while being 6 to 10 times lighter than state-of-the-art generative adversarial networks (GANs) approaches.

1 Introduction

Low-light image enhancement is fundamentally an image-to-image translation problem which aims to map low quality inputs to high quality versions. It is a task focusing on improving visual quality of an underexposed image which suffers from poor visibility, low contrast and noise. Recent works typically learn an one-to-one mapping functions from the perspective of paired data [Wei *et al.*, 2018], learning unpaired features [Jiang *et al.*, 2021] and brightness constraints [Guo *et al.*, 2020]. However, this mapping is not necessarily one-to-one, as one may want to generate from one input image multiple enhanced versions with different characteristics (lighting, tone, details etc), and meanwhile one high quality image can correspond to many low quality versions. This ill-posed nature indicates the unsuitability of paired supervision with one-to-one mapping assumption in image enhancement tasks, which exactly motivates our work. In this work, we define the task as improving the color, brightness and contrast of the input image conditioned on a given reference image. The use of conditioning enables one-to-many learning.

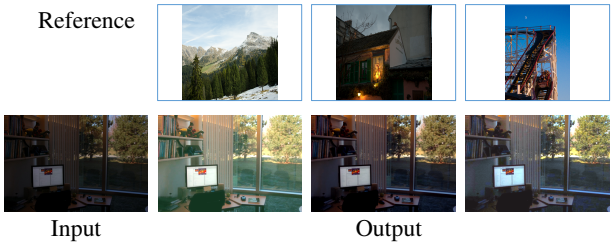

Figure 1: One-to-many Image Enhancement. We show an input low-light image at the bottom left and a set of normal-light reference images with different style on the first row. We then show our stylized results with enhanced light condition.

In this paper, we focus on learning an one-to-many mapping model without paired training samples. Concretely, as illustrated in Figure 1, we are able to translate a given low-light image to a normal-light one conditioned on the reference image (e.g., user preference on the image brightness, contrast, etc.). The conditional enhancement procedure is conducted by a U-Net Translator and a Modulation Code Generator (MCG). Specifically, the MCG generates a modulation code that fuses the characteristics of the learned features of both the low-light input image and the reference image. Meanwhile, the U-Net Translator performs conditional translation on the input low-light image with the assistance of our proposed Pixel-wise Self-Modulation (PSM) and Channel-wise Conditional-Modulation (CCM). The PSM module learns to adjust the mean and variance of the feature of input low-light image on each spatial location, while CCM is complementary to this operation. It performs channel-wise modulation conditioned on the modulation code generated by the MCG.

To enable unpaired learning, we optimize the model with four objective functions: 1) the idempotence loss that assumes a normal-light image should be mapped to itself when conditioned on itself. 2) the spatial consistency loss that facilitates the generated image to have more spatial coherence with the input. 3) the global color consistency loss that makes the overall color coherent with the input. 4) the GAN loss that tries to make the outputs more realistic.

[*]Equal contribution.
[†]Corresponding author.

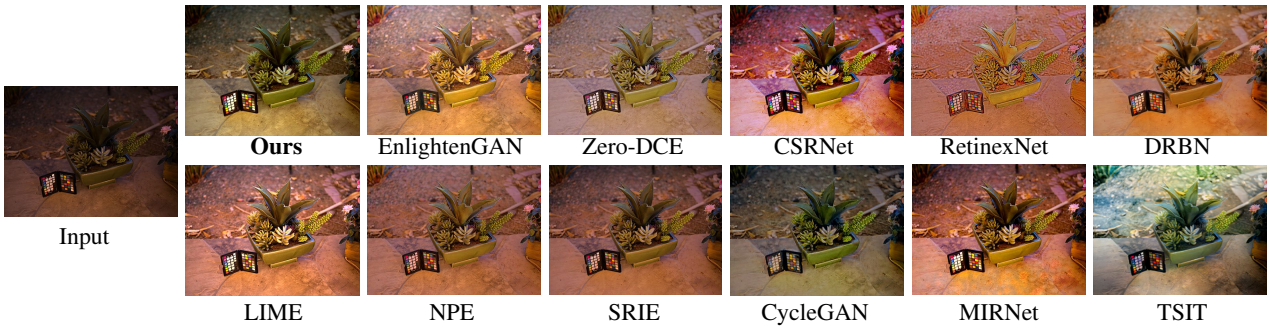

Figure 2: Our method outperforms the state-of-the-art baselines both qualitatively and quantitatively.

2 Related Works

We review related works in two main categories: image enhancement and conditional image generation. We mainly discuss the literature that addresses the problem of low-light image enhancement in unpaired and unsupervised setting, which is closely related to our setting.

2.1 Image Enhancement

Traditional Methods. There are two main categories of methods for low-light image enhancement, Histogram Equalization (HE) based methods [Pizer *et al.*, 1987] and Retinex [Land, 1977]. For example, LIME [Guo *et al.*, 2018] searches the maximum value in the RGB channels of the image to estimate the illumination of each pixel then rebuilt the illumination map with a structure prior. However, these methods have poor generalization ability, and often result in visible noise for real low-light images.

Learning-based Methods with Paired Supervision. Recently, the methods based on deep neural network achieve impressive results on low-light image enhancement. In enhancing natural images, there are many promising methods [Lore *et al.*, 2017; Wei *et al.*, 2018]. And for relight HDR images, HDR-Net [Gharbi *et al.*, 2017] utilized bilateral grid processing and local affine color transforms. Aiming at enhance raw sensor data, [Chen *et al.*, 2018a] proposed a "learning to see in the dark" methods that achieves impressive visual results.

Learning-based Methods with Unpaired Supervision. In image translation, there are many excellent works [Zhu *et al.*, 2017; Wang *et al.*, 2018] based on unpaired data. Focusing on image enhancement, several methods are proposed due to the difficulty of obtaining paired data in real scenes. [Yang *et al.*, 2020] train a deep recursive band network with paired-unpaired images. [Jiang *et al.*, 2021] propose EnlightenGAN that can be trained without low/normal-light image pairs. Zero-DCE [Guo *et al.*, 2020] estimates the pixel-wise and the high-order curves for dynamic range adjustment of a low-light image in an unsupervised way. These methods are one-to-one mapping of low light images to target domain. However, image enhancement is ill-posed and cannot be inverted with a deterministic mapping. While different from aforementioned works that map low-light images to a single enhanced distribution, we develop a lightweight conditional GAN, that learns a one-to-many relation from low-light to normal-light image space without paired datasets.

2.2 Conditional Image Generation

The generative adversarial networks [Goodfellow *et al.*, 2014] (GANs) employ a discriminator to distinguish the generated images from the real ones. Prior works have conditioned GANs (i.e., cGANs) on discrete labels [Mirza and Osindero, 2014], text [Reed *et al.*, 2016] or images [Isola *et al.*, 2017].

Among them, the most related direction to ours is translating an image from one domain to another, conditioned on a given reference image. Along this line, previous works in image style transfer introduce Conditional Instance Normalization (Conditional IN) and Adaptive Instance Normalization (AdaIN) to adjust the mean and the variance of the content input by style-specific parameters [Dumoulin *et al.*, 2017] or alternatively by directly replacing the mean and the variance with those of the style input [Huang and Belongie, 2017]. Basically, these normalization-based methods first normalize the features to a normal distribution, then denormalize them with a learned affine transformation whose parameters inferred from external data. Due to their flexibility, both were successfully adopted in various tasks with paired supervision [Brock *et al.*, 2018; Park *et al.*, 2019; Zhang *et al.*, 2020].

Most of conditional image generation works are trained with paired data, e.g., segmentation masks and images. While in this paper, we focus on unpaired conditional image generation, and achieve it with several proposed schemes that are elaborately tailored for image enhancement (see Fig. 2 for a qualitative comparison between our method and above mentioned baselines.).

3 Methodology

Problem Formulation. We aim at transferring a given low-light image to its normal-light counterpart according to user's preference without paired training samples. Formally, let $\mathbf{x}$ be an input image in the image space $\mathcal{X}$. We want to construct a conditional mapping $G(\mathbf{x}\,|\,\mathbf{y}_{\text{ref}}) : \mathbf{x} \in \mathcal{X} \rightarrow \mathbf{y} \in \mathcal{Y}$ which maps $\mathbf{x}$ to an image $\mathbf{y}$ in the target image space $\mathcal{Y}$ of normal-light images, conditioned on the reference image $\mathbf{y}_{\text{ref}}$. To this end, we propose to transfer the style information contained in $\mathbf{y}_{\text{ref}}$ through the form of a modulation code $\mathbf{c}_{\text{ref}}$ (see Fig. 4).

Approach Overview. To enhance the input low-light image to a normal-light one conditioned on a reference image, we employ a U-Net Translator that performs conditional

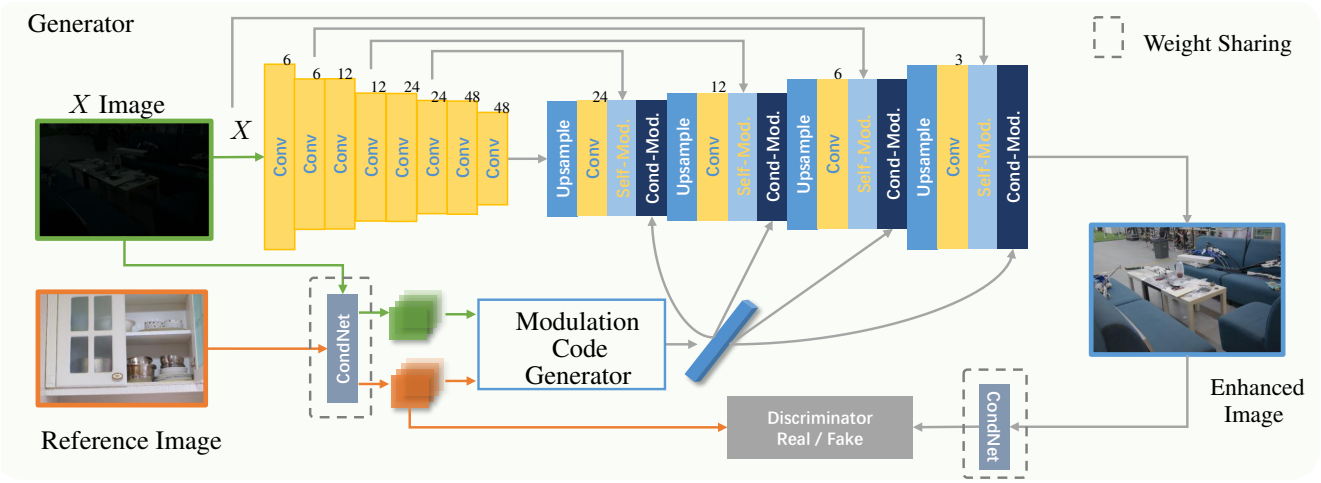

Figure 3: Overview of our Condition GAN for image enhancement.

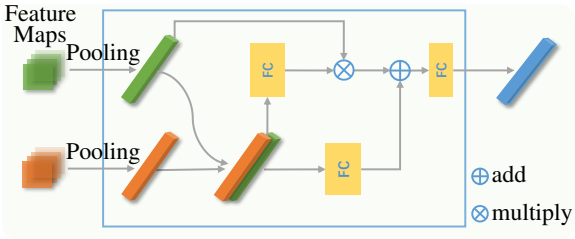

Figure 4: Modulation Code Generator

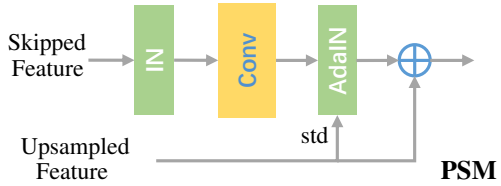

Figure 5: Pixel-wise Self-Modulation (PSM) block

translation on the input low-light image. Our U-Net Translator consists of two complementary modules: the Pixel-wise Self-Modulation (PSM) and the Channel-wise Conditional-Modulation (CCM), where both of them are designed to adjust the feature distribution of the low-light input image but from different aspects. Specifically, the PSM is designed to learn modulation parameters from previously upsampled features, while the CCM is designed to learn from the features of both the low-light and the reference image. In particular, the features fed into CCM are generated by our Condition Net (CondNet), which consists of three convolutional layers. To make the enhanced image natural, we also equip the model with a discriminator that distinguishes the feature outputted by CondNet from low- and normal-light domain. By enabling weight sharing, CondNet is encouraged to learn the difference , between the two image space, seen by both the generator and the discriminator. It prevents discriminator from cheating with discriminating based on other feature that is unrelated to the generator side.

3.1 Modulation Code Generator

A typical conditional image generation approach generates modulation parameters (e.g., learned scale and bias for AdaIN [Dumoulin *et al.*, 2017]) purely based on conditional input, which is the reference image $\mathbf{y}_{\text{ref}}$ in our setting [Park *et al.*, 2019; Zhang *et al.*, 2020]. However, for our unpaired image enhancement, there are several issues of generating modulation parameters from the reference image only: 1) Our goal is to enhance input image itself according to the reference image, which not only fully depends on the reference image but also needs to consider the property of the input image. 2) Since we do not have paired training samples, it is impossible to optimize the model with pair-wise constraint like Mean Square Error. Generating modulation code only conditioned on the reference may allow the network to take shortcuts and cheat on the loss function where low cost is achieved while no valid characteristic is extracted from the reference image. More specifically, the network would learn a *constant bias* while only take the reference image as input. Therefore, we combine the information from both the input image and the reference one to facilitate the learning process of Modulation Code Generator. Formally, we perform global average pooling on the outputs of CondNet, forming two feature vectors $\mathbf{x}^c$ and $\mathbf{y}^c_{\text{ref}}$ for the input image $\mathbf{x}$ and the reference image $\mathbf{y}_{\text{ref}}$ respectively. Our Modulation Code Generator can be formulated as:

$$\begin{aligned} \mathbf{c}_{\text{ref}} &= \texttt{fc}_{\texttt{out}}(\texttt{fc}_{\texttt{in}}(\mathbf{c}'_{\text{ref}}) \odot \mathbf{y}^c_{\text{ref}} \oplus \texttt{fc}_{\texttt{y}}(\mathbf{c}'_{\text{ref}})), \\ \text{where } \mathbf{c}'_{\text{ref}} &= \texttt{concat}(\mathbf{x}^c, \mathbf{y}^c_{\text{ref}}), \end{aligned} \tag{1}$$

`concat` and `fc` indicate concatenation operation and fully-connected layer respectively, $\odot$ and $\oplus$ denote element-wise multiplication and addition respectively.

3.2 U-Net Translator

Our U-Net Translator follows a typical encoder-decoder architecture with skip connections [Ronneberger *et al.*, 2015], which has strong capability on multi-scale texture preservation [Jiang *et al.*, 2021]. We tailor the standard U-Net archi-

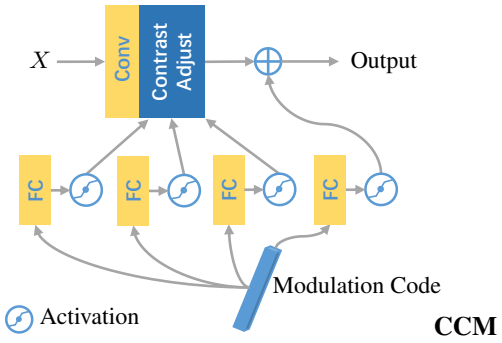

Figure 6: Channel-wise Conditional-Modulation (CCM) block

Figure 7: Ablation study on losses and the CCM block. On column 3-4, we verify the usefulness of our loss functions on two challenge cases of low-light images. In the last column, we show the results of replacing our CCM block with Global Feature Modulation (GFM) followed by ReLU, which justifies the effectiveness of our CCM block.

tecture for our task from three aspects: 1) we propose a Pixel-wise Self-Modulation (see Fig. 5) to match the statistics of the upsampled feature with the skipped feature. 2) we propose a Channel-wise Conditional-Modulation (see Fig. 6) to polish the feature with the generated modulation code (see Fig. 4), which is vital for conditional image enhancement. 3) we remove the original batch normalization layers which destroy the relative distribution across channels, as our Channel-wise Conditional-Modulation is learned more efficiently by seeing the original distributions. We elaborate the two proposed modules as follows.

Pixel-wise Self-Modulation (PSM). Commonly, the decoder of the U-Net combines the preceding feature generated from the encoder with the upsampled feature by concatenation or summation. However, in an enhancement setting, the feature generated from the encoder side generally represents the input in the low-light domain. While for the decoder side, we gradually generate normal-light representation for the input. Simply mixing them together without additional adjustment will lead to a domain gap between the encoder and decoder representation. Therefore, instead of a direct concatenation, we propose a Self-modulation Block to adjust the statistics of skipped feature by the upsampled feature from the previous layer. As illustrated in Figure 5, the skipped feature is first processed by an instance normalization layer and two 3×3 convolution layers with leaky ReLU, then fed into AdaIN [Huang and Belongie, 2017], whose mean and variance are calculated from the upsampled feature. The goal of our PSM block is to enhance the skipped feature, which consists of multi-scale texture information, in an adaptive way. Intuitively, it can be viewed as modulating the lower-level representation using the higher-level one. Thus, the PSM block is expected to automatically learn to enhance the image while preserving the detailed content. Note that, our PSM block is totally different from [Chen *et al.*, 2018b], where the feature is modulated by some input noise. In contrast, our modulation condition is provided by previous layer, targeting at matching the statistics between the skipped feature and the upsampled one.

Channel-wise Conditional-Modulation (CCM). Our CCM block plays two important roles: 1) it transfers the style of the reference image, which encoded in a modulation code, to the input low-light image. 2) it performs learnable non-linear transformation on the learned feature. Formally, let $\mathbf{c}_{\text{ref}}$ be the modulation code inferred from $\mathbf{y}_{\text{ref}}$, as shown in Figure 6, our CCM block first generates four coefficient vectors α^1_{ref}, α^2_{ref}, $\bar{\alpha}_{\text{ref}}$, and β_{ref} based on the modulation code $\mathbf{c}_{\text{ref}}$ through two fully-connected layers. Then the retouch operation $m(\cdot)$ can therefore be formulated as:

$$m(\mathbf{x}) = \begin{cases} \alpha^1_{\text{ref}} \odot (\mathbf{x} \ominus \bar{\alpha}_{\text{ref}}) \oplus \beta_{\text{ref}} & \text{if } \mathbf{x} > \bar{\alpha}_{\text{ref}} \\ \alpha^2_{\text{ref}} \odot (\mathbf{x} \ominus \bar{\alpha}_{\text{ref}}) \oplus \beta_{\text{ref}} & \text{if } \mathbf{x} \leq \bar{\alpha}_{\text{ref}}, \end{cases} \tag{2}$$

where $\odot$, $\ominus$ and $\oplus$ indicate element-wise multiplication, subtraction and addition respectively. A typical contrast and brightness adjustment operation in image processing can be formulated as $I^{\text{new}} = \alpha I + (1-\alpha)\beta + b$, where α is a scaling factor, β is the average intensity value, and b is the brightness adjustment coefficient. which can be simplified as $\alpha I + \gamma$. However, this linear operation has limited effect as compared to the more sophisticated curve adjustment. In previous work, non-linearity is achieved by introducing activation functions like ReLU into the main path [Jiang *et al.*, 2021; He *et al.*, 2020], or by using curve adjustment directly [Guo *et al.*, 2020]. We propose to simulate curve adjustment by compositing the function defined in (2) for multiple times.

3.3 Objective Functions

As we have encoded the brightness and the contrast information into the Condition Net, and further encoded the reference-controlled channel-wise modulation information into the modulation code, our U-Net translator only needs to carry forward content information, and is therefore suitable to take as input images from both low-light and normal-light image space. Our training involves feeding images from both space to the translator. We propose to use two non-reference pixel-wise losses, i.e. an idempotence loss and a spatial consistency loss, within the target image space, together with a global color consistency loss and a GAN loss to enable unpaired learning in U-Net Translator.

Idempotence Loss. The idempotence loss requires that a normal-light image should be mapped to itself when conditioned on itself. Let $\mathbf{y}$ be an image sampled from the normal light space, the loss is defined as

$$L_{\text{idem}} = \left\| G(\mathbf{y} \,|\, \mathbf{y}) - \mathbf{y} \right\|_1 \tag{3}$$

Spatial Consistency Loss. We adopt the spatial consistency loss [Guo *et al.*, 2020] between the generated reference, which encourages spatial coherence between two images in the form of consistent gradient variation in the local neighborhood. Let $\mathbf{y}_1, \mathbf{y}_2$ be two images sampled from the normal light space, the loss is given as

$$L_{\text{spa}} = \left\| \nabla G\big(\mathbf{y}_1 \,|\, \mathbf{y}_2\big) - \nabla \mathbf{y}_1 \right\|_1 \tag{4}$$

Global Color Consistency Loss. The relative strength of each color channel of the enhanced output should not deviate significantly from the input. To this end, We propose a global color consistency loss which prevents unrealistic color tone shift. Let $I^c(\cdot)$ denote the average intensity value of channel c of an layer-normalized image, we define the loss as

$$L_{\text{color}} = \sum_{c \in \{R,G,B\}} \Big(I^c\big(G(\mathbf{x} \,|\, \mathbf{y}_{\text{ref}})\big) - I^c\big(\mathbf{x}\big) \Big)^2 \tag{5}$$

GAN Loss. We slightly modify the standard GAN loss by letting the discriminator see two types of fake images, one generated from an low-light input $\mathbf{x}$ and the other generated from a normal-light input $\mathbf{y}$, both conditioned on a reference normal-light image $\mathbf{y}_{\text{ref}}$. Overall, the loss is written as

$$\begin{aligned} L_{\text{GAN}} = \quad & \mathbf{E}_{\mathbf{y} \sim \mathcal{Y}} \big[\log \big(D(\mathbf{y}) \big) \big] + \\ & \lambda \mathbf{E}_{\mathbf{x}} \big[\log \big(1 - D\big(G(\mathbf{x} \,|\, \mathbf{y}_{\text{ref}})\big) \big) \big] + \\ & (1-\lambda) \mathbf{E}_{\mathbf{y}, \mathbf{y}_{\text{ref}} \sim \mathcal{Y}} \big[\log \big(1 - D\big(G(\mathbf{y} \,|\, \mathbf{y}_{\text{ref}})\big) \big) \big] \end{aligned} \tag{6}$$

Overall, our final loss function is a combination of each individual constraint:

$$L_{\text{total}} = L_{\text{idem}} + L_{\text{spa}} + L_{\text{color}} + \alpha L_{\text{GAN}}. \tag{7}$$

4 Experiments

4.1 Dataset and Implementation Details

One of the main advantages of our unpaired setting for image enhancement is that we utilize a much larger collection of low-light and normal-light images without imposing given correspondences between the images, which is not the case for the methods designed on paired and fully supervised setting. Thereby, we assemble images from three different datasets [Wei *et al.*, 2018; Bychkovsky *et al.*, 2011; Loh and Chan, 2019] and ignore the paired information in each individual dataset if there is any, which leads to a larger and more diverse dataset that consists of 983 low-light and 5576 normal-light images. We follow the same practice of previous work[Yang *et al.*, 2020] to use part of the LOL dataset[Wei *et al.*, 2018] for training, and leaving the other part for testing. We then train our network on this unpaired dataset and compare to other methods with their pretrained models.

We implement our network with PyTorch on a Tesla GPU. Our network has $891,527$ parameters in total including the discriminator, leading to almost 10 times reduction in size as compared to EnlightenGAN [Jiang *et al.*, 2021] with $8,636,675$ parameters. The weights of each layer are initialized with random values sampled from a Gaussian with 0 mean and 1 standard deviation. We adopt Adam optimizer with default parameters and with learning rate set to 5×10^{-5}. We set the loss weight λ in Eq. (6) to 0.9, and α in Eq. (7) to 0.05 in all the tests. Our code can be found at https://github.com/sxpro/ImageEnhance_cGAN.

4.2 Ablation Study

We demonstrate the effectiveness of our choice of losses and the CCM block via ablation studies. We do not ablate the idempotent loss L_{idem} as it is the only loss that enforces content consistency in our setting, meaning that the generator would produce almost arbitrary results in absence of L_{idem}.

Contrast adjustment module. Our Channel-wise Conditional-Modulation (CCM) is designed to transfer the *style* of the reference image which can particularly capture the contrast and the brightness of the reference image. An alternative design choice would be adopting GFM followed by leaky ReLU as used in [He *et al.*, 2020]. However, this can result in non-realistic color tones as shown in the last column in Fig. 7. On the contrary, our proposed CCM component can properly capture the style feature of the reference image and smoothly transfer it to the output.

Spatial Consistency loss. The spatial consistency loss Eq. (4) can help the network to better learn and infer the normal-light image space explicitly where the spatial coherence between to normal-light images are promoted via this loss. We can observe the contribution of this loss in the third column of Fig. 7.

Color Consistency loss. The color consistency loss Eq. (5) ensures that the color distribution of the output image does not deviate too much from the input image though the light condition get enhanced significantly. Removing this loss can lead to results with undesirable color distribution (see the fourth column in Fig. 7 for an example).

4.3 Benchmark Evaluations

We compare our conditional GAN with several state-of-the-art methods, including SIRE [Fu *et al.*, 2016], LIME [Guo *et al.*, 2018], NPE [Wang *et al.*, 2013], RetinexNet [Wei *et al.*, 2018], DRBN [Yang *et al.*, 2020], CSRNet [He *et al.*, 2020], EnlightenGAN [Jiang *et al.*, 2021], Zero-DCE [Guo *et al.*, 2020], CycleGAN [Zhu *et al.*, 2017], TSIT [Jiang *et al.*, 2020], and MIRNet [Zamir *et al.*, 2020].

Specifically, Table 1 shows a quantitative comparison between our method and the other baselines on the PSNR, SSIM and the NIQE [Mittal *et al.*, 2012] metrics. Note that, even in a more challenging setup without paired information and conditional constraint, our method achieves the state-of-the-art performance on FiveK among unpaired methods. Fig. 8 shows a qualitative comparison. In LOL-#690, we can see our results have brightness and contrast with less noise, and we have enhanced the color and contrast of the trees and the building in LIME-#2.

5 Conclusion

In this paper, we propose a conditional GAN with tailored components including PSM and CCM for image en-

Method \ Metric		Unpaired	Conditional	LOL-690	FiveK
SRIE				15.35 \ 0.559 \ 7.4022	16.90 \ 0.750 \ 4.1352
LIME				17.97 \ 0.512 \ 8.2972	16.67 \ 0.772 \ 3.7043
NPE				17.62 \ 0.481 \ 8.5105	15.60 \ 0.736 \ 3.6475
RetinexNet				16.17 \ 0.420 \ 9.2652	11.89 \ 0.644 \ 4.4298
DRBN				18.71 \ 0.784 \ 4.5612	15.07 \ 0.562 \ 7.1623
CSRNet				15.69 \ 0.408 \ 8.1343	23.68 \ 0.896 \ 3.7492
EnlightenGAN		✓		18.89 \ 0.692 \ 5.0857	15.47 \ 0.734 \ 3.7616
Zero-DCE		✓		18.47 \ 0.598 \ 7.8224	13.01 \ 0.557 \ 7.3117
CycleGAN		✓		17.42 \ 0.576 \ 4.0663	17.04 \ 0.681 \ 4.8327
TSIT			✓	13.14 \ 0.533 \ 5.5965	14.35 \ 0.638 \ 5.3926
MIRNet		✓		12.90 \ 0.432 \ 4.2501	19.36 \ 0.806 \ 3.9225
Ours	Min.	✓	✓	12.24 \ 0.609 \ –	11.97 \ 0.655 \ –
	Avg.			17.00 \ 0.671 \ –	17.37 \ 0.750 \ –
	Max.			22.45 \ 0.732 \ 4.0733	20.87 \ 0.797 \ 4.0305

Table 1: PSNR(↑) \ SSIM(↑) \ NIQE(↓) metrics on the paired test set of datasets LOL [Wei *et al.*, 2018] starting from image #690, and FiveK [Bychkovsky *et al.*, 2011]. The arrow after each metric indicates whether a larger or a smaller value is better. As our method generates a distribution of output images given a set of reference images, we report the minimum, average and maximum values of PSNR and SSIM. As NIQE is a no ground-truth quality metric, we can thus select the reference which gives the best NIQE, from the reference set.

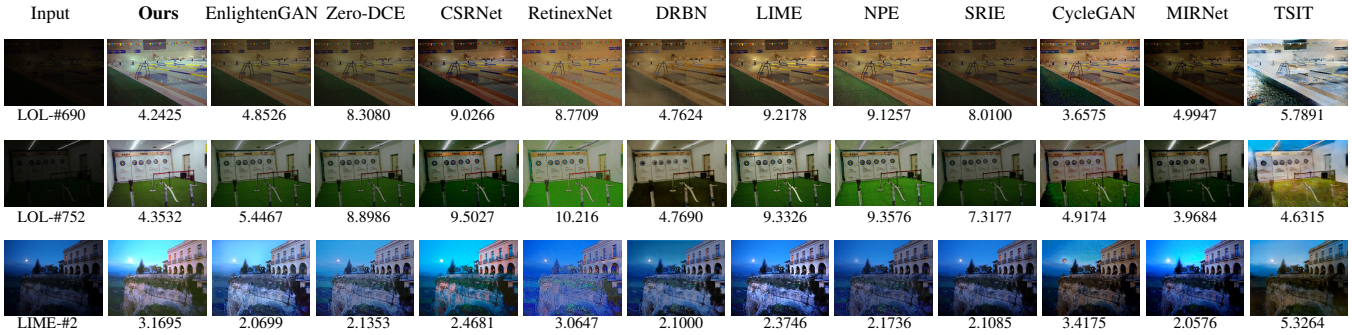

Figure 8: Qualitative comparison. We compare out method with the state-of-the-art baselines. The top two rows show results on the images from the test set, while the last row (and also Fig. 2) show results on the LIME dataset [Guo *et al.*, 2018]. We can see that our method has better generalization ability. Below each image, we also report the NIQE metric for each image. Best viewed by zooming in the electronic version.

hancement in an unpaired setting. We also propose task-specific losses including an idempotence loss, a spatial consistency loss, a global color consistency loss, which are combined with the standard GAN loss to encode the reference-controlled channel-wise modulation information and the brightness/contrast information of the input image and the reference image. Our design addresses the one-to-many mapping nature of the problem of image enhancement and achieves state-of-the-art performance on several standard datasets in a much lighter design with 10× less parameters as compared to the state-of-the-art [Jiang *et al.*, 2021]. We have justified the usefulness of our designed losses in image enhancement and we believe they can be applied to other image processing tasks such as style transfer or image synthesis due to the fact that our losses encode the general and global information of input images. Therefore, in the future work, we would like to investigate the generalization ability and effectiveness of our designed losses and investigate the performance of our PSM/CCM components in other network architecture.

Acknowledgments

We thank Jing Ren for useful suggestions on the manuscript and all the anonymous reviewers for their valuable comments.

References

[Brock *et al.*, 2018] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2018.

[Bychkovsky *et al.*, 2011] 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.

[Chen *et al.*, 2018a] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018.

[Chen *et al.*, 2018b] Ting Chen, Mario Lucic, Neil Houlsby, and Sylvain Gelly. On self modulation for generative adversarial networks. In *ICLR*, 2018.

[Dumoulin *et al.*, 2017] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017.

[Fu *et al.*, 2016] 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.

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

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 27, 2014.

[Guo *et al.*, 2018] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE TIP*, 2018.

[Guo *et al.*, 2020] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *CVPR*, 2020.

[He *et al.*, 2020] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *ECCV*, 2020.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[Jiang *et al.*, 2020] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020.

[Jiang *et al.*, 2021] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 2021.

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

[Loh and Chan, 2019] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *CVIU*, 2019.

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

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[Mittal *et al.*, 2012] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a completely blind image quality analyzer. *IEEE SPL*, 2012.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[Pizer *et al.*, 1987] Stephen M Pizer, E. Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *CVGIP*, 1987.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

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

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

[Wang *et al.*, 2018] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018.

[Wei *et al.*, 2018] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018.

[Yang *et al.*, 2020] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 2020.

[Zamir *et al.*, 2020] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, 2020.

[Zhang *et al.*, 2020] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, 2020.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.