Degrade Is Upgrade: Learning Degradation for Low-Light Image Enhancement

Kui Jiang[1], Zhongyuan Wang[1*], Zheng Wang[1], Chen Chen[2], Peng Yi[1], Tao Lu[3], Chia-Wen Lin[4]

[1] School of Computer Science, Wuhan University, Wuhan, China
[2] University of Central Florida
[3] Wuhan Institute of Technology
[4] National Tsing Hua University
{kuijiang, wangzwhu, yipeng}@whu.edu.cn, wzy_hope@163.com, chen.chen@crcv.ucf.edu, lutxyl@gmail.com, cwlin@ee.nthu.edu.tw

Abstract

Low-light image enhancement aims to improve an image's visibility while keeping its visual naturalness. Different from existing methods tending to accomplish the relighting task directly by ignoring the fidelity and naturalness recovery, we investigate the intrinsic degradation and relight the low-light image while refining the details and color in two steps. Inspired by the color image formulation (diffuse illumination color plus environment illumination color), we first estimate the degradation from low-light inputs to simulate the distortion of environment illumination color, and then refine the content to recover the loss of diffuse illumination color. To this end, we propose a novel Degradation-to-Refinement Generation Network (DRGN). Its distinctive features can be summarized as 1) A novel two-step generation network for degradation learning and content refinement. It is not only superior to one-step methods, but also capable of synthesizing sufficient paired samples to benefit the model training; 2) A multi-resolution fusion network to represent the target information (degradation or contents) in a multi-scale cooperative manner, which is more effective to address the complex unmixing problems. Extensive experiments on both the enhancement task and joint detection task have verified the effectiveness and efficiency of our proposed method, surpassing the SOTA by *0.70dB on average and 3.18% in mAP*, respectively. The code will be available soon.

1 Introduction

Low-light conditions cause a series of visibility degradation and even sometimes destroy the color or content of the image. Such signal distortion and detail loss often lead to the failure of many computer vision systems (Bae 2019; Huang et al. 2018; Yang et al. 2020a; Yu, Yang, and Chen 2021; Xu et al. 2021a; Zhong et al. 2021; Huang et al. 2021; Xu et al. 2021b). Therefore, there is a pressing need to develop algorithms to produce normally exposed outputs.

Early low-light enhancement works mainly focus on contrast enhancement to recover the visibility of dark regions (Land 1977). Since the hand-crafted priors and assumptions are introduced for specific scenes or degradation conditions, these techniques (Kaur, Kaur, and Kaur 2011) are less effective on scenarios when the predefined models do not hold. Recently, deep learning frameworks have emerged as a promising solution to low-light image enhancement (Zhu et al. 2020; Chen et al. 2018; Zhang et al. 2021; Jiang et al. 2021b; Zhu et al. 2020). The existing methods can be roughly divided into two categories regarding the imaging models under low-light conditions.

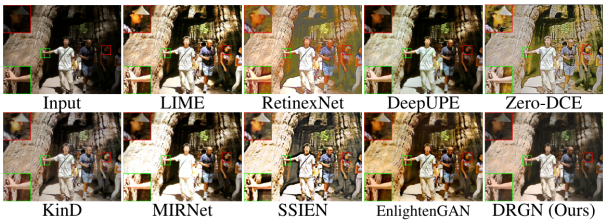

Figure 1: Results of different low-light image enhancement methods. RetinexNet (Wei et al. 2018), KinD (Zhang, Zhang, and Guo 2019), DeepUPE (Wang et al. 2019) and EnlightenGAN (Jiang et al. 2021b) light up the image contents, but cause severe color distortion. LIME (Guo, Li, and Ling 2017), Zero-DCE (Guo et al. 2020), MIRNet (Zamir et al. 2020) and SSIEN (Zhang et al. 2020) tend to produce over-exposed images. Our DRGN generates more realistic and credible textures with visually pleasing contrast.

Scheme One: An observed low-light image I_L is featured as a superposition of a normal-light image I with some perturbations I_N (noise, exposure, etc.) through a special combination manner $\theta(\cdot)$, defined as

$$I_L = \theta(I, I_N). \tag{1}$$

Given I_L, the goal of low-light image enhancement is to predict I. When $\theta(\cdot)$ is simplified as the pixel-wise summation (Gharbi et al. 2017), the effort is to learn the perturbation I_N and then subtract it from the observed image I_L. Some methods (Lv et al. 2018; Yang et al. 2020b) also attempt to directly estimate the optimal approximation of I. However, these methods rely heavily on abundant data, innovative architectures, and training strategies, ignoring the intrinsic degradation itself, resulting in unsatisfactory and unnatural contents in texture, color, contrast, etc.

Scheme Two: The low-light image I_L is mathematically modeled as a combination of the reflectance I_R and illumination I_T component through Eq. 2:

$$I_L = I_R \circ I_T, \tag{2}$$

*Corresponding Author

where $\circ$ denotes the element-wise product. With this kind of scheme, an image can be enhanced either by estimating and adjusting the illumination map (Li et al. 2018; Wang et al. 2019), or by learning joint features of these two components (Zhang, Zhang, and Guo 2019). It is known as the Retinex based method (Land 1977; Wei et al. 2018). These methods follow the layer decomposition paradigm and show impressive results in stretching contrast and removing noise in some cases, but still have three problems. 1) Traditional models rely on properly enforced priors and regularization. 2) Deep RetinexNets require a large number of paired training samples. 3) The enhanced reflectance I_R is often treated as the normal-light image (Yue et al. 2017), which often fails to produce high-fidelity and naturalness results due to the gap between the ideal situation and reality with simplified imaging models.

In sum, there are two main drawbacks of these two schemes. First, their performances *heavily rely on the diversity and quality of synthetic training samples*. Second, they *simplify the imaging model* when coping with the enhancement task, thus *sacrificing representation precision*, especially when being trained on a small dataset. More specifically, since the degradation additionally damages the texture, color, and contrast of normal-light images, we argue that subtracting the perturbation I_N from its low-light input I_L directly (**scheme one**) cannot fully recover multiple kinds of fine details and color distribution (refer to MIRNet (Zamir et al. 2020) in Figure 1). Meanwhile, adjusting the illumination map to enhance reflectance I_R (**scheme two**) cannot fully keep the visual naturalness of the predicted normal-light image (refer to RetinexNet (Wei et al. 2018) and DeepUPE (Wang et al. 2019) in Figure 1).

In light of the color image formulation, an image is composed of the diffuse illumination color and the environment illumination color, determined by the light source, albedo (material), environment light, noise, etc (Wu and Saito 2017). We thus promote **scheme one** by considering the blending relations between the normal-light image I and the degradation I_D, denoted as $f(I_D, I)$. The newly designed low-light image generation process can be modeled as

$$I_L = I_D + I + f(I, I_D) = I_D + \psi(I_D, I). \quad (3)$$

We argue $(I, f(I_D, I))$ represents the intrinsic manifold projection from I to I_L via degradation I_D. We formulate this complex mapping as a high-dimensional non-analytic transfer map $\psi(I, I_D)$ since there is no explicit analytic function to present the blending relations caused by the albedo (material) (I), environment light, noise, etc (I_D).

By doing so, we believe the degradation I_D can be learned from the low-light input I_L via a degradation generator (DeG), formulated as $\phi(I_L)$. Thus, given I_L, we use Eq. 4 to predict the normal-light image I as

$$I = \psi^{-1}(I_L - \phi(I_L)). \quad (4)$$

To alleviate the learning difficulty, our enhancement task is then decomposed into two stages: (1) simulating the degradation I_D via DeG, and (2) refining the color and contrast information by a refinement generator (ReG), which is parameterized by $\psi^{-1}(\cdot)$ (learning the inverse representation of $\psi(\cdot)$). More specifically, inspired by the impressive results of GAN in image synthesis and translation (Gatys, Ecker, and Bethge 2015; Bulat, Yang, and Tzimiropoulos 2018), we train a parametric generator to learn the degradation factors from the low-light input in the first stage, denoted as $\phi(I_L)$. Note that simulated degradation factors are also applied to another referenced high-quality dataset to generate synthetic low-light samples to mitigate the limitation of sample monotonicity and repetitiveness (More details on the data generation are described in Sec. 2.3). Meanwhile, we produce the base enhancement result I_B by removing the predicted degradation from the low-light input. Thus, the second stage takes I_B as input, and applies ReG to estimate the inverse transformation matrix to further refine the color and textural details.

DeG is implemented with a multi-resolution fusion strategy because the degradation always acts on various texture richness levels. We thus propose a multi-resolution fusion network (MFN) to learn the joint representation of multi-scale features. In the second stage, ReG shares the same architecture as DeG for simplicity, since the cooperative representation among the multi-scale features also contributes more to the recovery of image textures. Our main contributions are summarized as follows:

- We propose to decompose and characterize the low-light degradation. Thanks to the degradation generator, it can augment an arbitrary number of paired samples using a small synthetic dataset, and achieves impressive performance and robustness. Our proposed degradation formulation can be easily extended to other enhancement tasks, to help get rid of paired dataset collection. Based on this, we construct a novel two-stage Degradation-to-Refinement Generation Network (DRGN) by cascading DeG and ReG for low-light image enhancement tasks.
- Besides achieving impressive enhancement performance (surpassing KinD++ (Zhang et al. 2021) by 0.70dB in PSNR on average), our method is also competitive on the joint low-light image enhancement and detection tasks on ExDark (Loh and Chan 2019) and NightSurveillance (Wang et al. 2020) datasets.
- We also introduce two novel low/normal-light datasets (COCO24700 for training and COCO1000 for evaluation) based on COCO (Caesar, Uijlings, and Ferrari 2018) dataset, which can be used for the joint low-light image enhancement and detection tasks under extreme night-time degradation scenarios.

2 Method

Figure 2 shows the overall pipeline of our proposed low-light enhancement model – Degradation-to-Refinement Generation Network (DRGN). The objective of DRGN is mainly two-fold: one is to train a degradation generator (DeG) $\mathcal{G}_{de} \mapsto \phi(\cdot)$ to simulate the degradation from low-light inputs. The other is to estimate the inverse transformation matrix with a refinement generator (ReG) $\mathcal{G}_{re} \mapsto \psi^{-1}(\cdot)$ to refine color and textures with the augmented dataset by applying the predicted degradation priors.

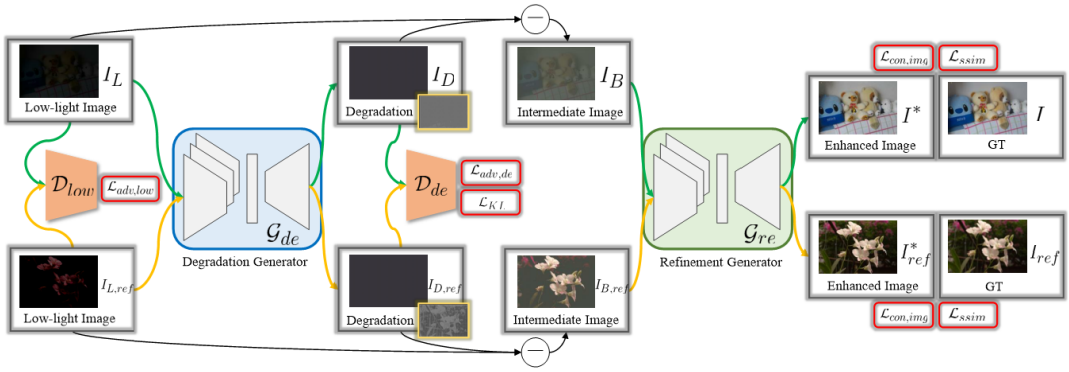

Figure 2: Architecture of Degradation-to-Refinement Generation Network (DRGN). It consists of two subnetworks to tackle degradation estimation and content refinement, respectively. A degradation generator (DeG) learns degradation I_D from the low-light input I_L in the first stage, and produces the base enhanced result I_B by removing I_D. Then, a refinement generator (ReG) takes I_B as input and produces the refined prediction (I^*) of the normal-light image (I). We also apply DeG to generate the synthetic paired samples $[I_{L,ref}, I_{ref}]$ to augment the sample space to help train these two generators.

2.1 Degradation Generator

Let I_L, I, and I_{ref} respectively denote the low-light input, normal-light image, and high-quality reference image (not in the paired training dataset). In the first stage, we learn the degradation through a generative model where all involved parameters can be inferred in a data-driven manner. Specifically, given a low-light input I_L, we train a degradation generator (DeG) to predict the degradation I_D from I_L, *e.g.*, $I_D = \mathcal{G}_{de}(I_L)$. Thus we can generate the synthetic low-light image $I_{L,ref}$ by combining I_D and I_{ref}. Moreover, we produce the degradation prediction $I_{D,ref}$ from $I_{L,ref}$ using the same generator DeG. More details about the data augmentation are described in Sec. 2.3.

Recall that the degradation style of the input I_L and the synthetic low-light image $I_{L,ref}$ should be shared, we regularize both the predicted degradation prior $[I_D, I_{D,ref}]$ and the low-light image pair $[I_L, I_{L,ref}]$ to be close. We introduce the generative adversarial loss (Ledig et al. 2017; Yi et al. 2020; Ma et al. 2019, 2020) and degradation consistency loss (the KL Divergence) to train both DeG and discriminator. Here, the generative adversarial loss between the low-light input and synthetic low-light image is defined as

$$\mathcal{L}_{adv,low} = -\log \mathcal{D}_{low}(I_L) - \log(1 - \mathcal{D}_{low}(I_{L,ref})). \quad (5)$$

In addition, the constraints on the predicted degradation priors $[I_D, I_{D,ref}]$ are expressed as

$$\begin{aligned} \mathcal{L}_{adv,de} &= -\log \mathcal{D}_{de}(I_D) - \log(1 - \mathcal{D}_{de}(I_{D,ref})), \\ \mathcal{L}_{kl} &= \sum p(I_D) log \frac{p(I_D)}{p(I_{D,ref})}, \end{aligned} \quad (6)$$

where $p(\cdot)$ denotes the target distribution. By setting α to 10^{-5} (Jiang et al. 2019) to balance the GAN losses and consistency loss, the total loss in the first stage is given by

$$\mathcal{L}_{DeG} = \alpha \times (\mathcal{L}_{adv,low} + \mathcal{L}_{adv,de}) + \mathcal{L}_{kl}. \quad (7)$$

2.2 Refinement Generator

We design the refinement generator (ReG) in the second stage using the same architecture of DeG for convenience. In the first stage, we also produce the base enhanced result I_B by removing I_D from the low-light input I_L. Thus ReG takes I_B as input and learns to refine the contrast and textural details. More importantly, except for the original sample pairs $[I_B, I]$, we also generate additional paired images $[I_{B,ref}, I_{ref}]$ for sample augmentation in the first stage. With the enrichment and augmentation of training samples, a more robust image enhancement model can be achieved. The procedures in the second stage can be formulated as

$$I^* = \mathcal{G}_{re}(I_B), I^*_{ref} = \mathcal{G}_{re}(I_{B,ref}), \quad (8)$$

where I^* and I^*_{ref} are the predicted normal-light images via ReG. Unlike the adversarial loss in the first stage, we introduce the content loss (the Charbonnier penalty function (Lai et al. 2017; Jiang et al. 2021a)) and the SSIM (Wang et al. 2004) loss between the predicted normal-light image and the high-quality ground truth to guide the optimization of ReG. The loss functions are expressed as

$$\begin{aligned} \mathcal{L}_{con,img} &= \sqrt{(I^* - I)^2 + \varepsilon^2} + \sqrt{(I^*_{ref} - I_{ref})^2 + \varepsilon^2}, \\ \mathcal{L}_{ssim} &= SSIM(I^*, I) + SSIM(I^*_{ref}, I_{ref}), \\ \mathcal{L}_{ReG} &= \mathcal{L}_{con,img} + \lambda \times \mathcal{L}_{ssim}, \end{aligned} \quad (9)$$

where λ is used to balance the loss terms, and experimentally set as -0.2. The penalty coefficient ε is set to 10^{-3}.

2.3 Data Augmentation

Previous hand-crafted methods, such as RetinexNet, provide an effective way for data synthesis, but still with limitations: 1) Misalignment errors of producing normal-light images; 2) Complex and diverse hyper-parameters and camera settings; 3) Limited and discrete sample space. Unlike them, our DRGN produces a *full-automatic pipeline to create infinite paired samples with controllable degradation strengths.*

Given a synthetic paired dataset $\mathcal{S} = \{I, I_L\}$ and a high-quality reference dataset $\mathcal{R} = \{I_{ref}\}$, we first train a generator to learn the degradation I_D from each I_L in the first stage. According to Eq. 3, the ideal low-light image can be produced via $I_L = I_D + \psi(I, I_D)$. In theory, there exits a **strictly high-quality** version of I_{ref} to fit Eq. 3, formulated as $I_{L,ref} = I_D + \psi(I^*_{ref}, I_D)$, where I^*_{ref} denotes

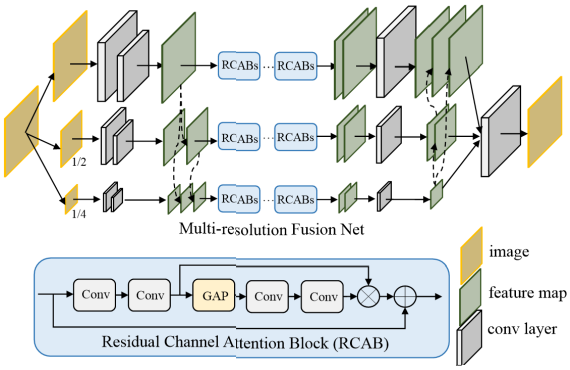

Figure 3: Pipeline of multi-resolution fusion network.

the normal-light version of I_{ref}, and $\psi(I^*_{ref}, I_D)$ is equal to I_{ref}, generated via $\psi(\cdot)$ and I_D. Since low-light image $I_{L,ref}$ and its high-quality version are not one-to-one correspondence to each other, we thus adjust the degradation mapping of $I_{L,ref}$. As shown in Figure 2, $I_{L,ref}$ is reformulated as $I_{L,ref} = I_{D,ref} + \psi(I_{ref}, I_{D,ref})$, which strictly fits the definition in Eq. 3. To maintain the diversity and reality of the degradation, we introduce GAN loss and consistency loss to regularize both predicted degradation $[I_D, I_{D,ref}]$ and low-light image pair $[I_L, I_{L,ref}]$ to have similar distributions.

Overall, the aforementioned procedure leads to a new paired dataset $\mathcal{T} = \{I_{ref}, I_{L,ref}\}$ during training. Therefore, our sample space consists of the original sample dataset $\mathcal{R}$ and the new synthetic dataset $\mathcal{T}$. Our degradation learning and data augmentation scheme is more like a transfer learning of degradation patterns from the normal-light sample domain to the low-light image domain.

2.4 Multi-resolution Fusion Network

Decomposing a complex task into several sub-problems and further learning their cooperative representation to yield a final solution is a practical scheme to promote the model performance (Papyan and Elad 2016; Fu et al. 2020). To this end, we propose to decompose the learning task (including degradation learning and content refinement) into multiple sub-spaces, and construct a multi-resolution fusion network (MFN) to represent the target information in a multi-scale collaborative manner, as demonstrated in Figure 3.

Taking the procedure of degradation learning in DeG for instance, we first generate the image pyramid from the low-light input via Gaussian sampling kernel $G(\cdot)$, and adopt initial convolutions (Zhang et al. 2018) to project the input into the feature space to extract initial features f^i_{ini}, expressed as

$$f^i_{ini} = \mathcal{H}^i_{ini}(G(I_L)), i \subseteq [1, n], \tag{10}$$

where n and i denote the sampling number and the corresponding i_{th} pyramid layer. In particular, the Gaussian sampling is activated except $i = 1$. Before passing f^i_{ini} into the backbone network, we sample the features of high-level branches with strided convolutions, and aggregate them with the features of lower-level branches to form the multi-scale representation (MSR), formulated as

$$f^i_{msr} = \mathcal{H}^i_{msr}(f^i_{ini}, f^j_{ini}), j \subseteq [1, i), \tag{11}$$

where $\mathcal{H}^i_{msr}(\cdot)$ denotes i_{th}-layer multi-scale fusion. After that, we apply the residual group $\mathcal{H}^i_{RG}(\cdot)$ composed of several cascaded residual channel attention blocks (RCABs), each containing multiple RCABs to build deep feature representations (Zhang et al. 2018), as shown in Figure 3. A 1×1 convolution is followed to perform the long-term fusion among the output and input of the residual group. The procedures above can be expressed as

$$f^i_{long} = \mathcal{H}^i_1([\mathcal{H}^i_{RG}(f^i_{msr}), f^i_{msr}]). \tag{12}$$

Through the n-layer pyramid network, we obtain n outputs. To better characterize the degradation, we utilize the deconvolution layer to up-sample the outputs of low-level pyramid layer, and then perform the multi-scale fusion with high-level features via a 1×1 convolution. Consequently, a reconstruction layer $\mathcal{H}_{rec}(\cdot)$ with the filter size of 3×3 is used to project the output of MFN back to the image space. The overall procedure can be summarized as

$$I_D = \mathcal{H}_{rec}(\mathcal{H}^j_{msr}(f^j_{long}, f^1_{long})), j \subseteq [2, n]. \tag{13}$$

3 Experiments

We carry out extensive experiments on synthetic and real-world low-light image datasets to evaluate our proposed DRGN. Nine representative low-light enhancement algorithms are compared, including LIME (Guo, Li, and Ling 2017), MIRNet (Zamir et al. 2020), KinD (Zhang, Zhang, and Guo 2019), KinD++ (Zhang et al. 2021), DeepUPE (Wang et al. 2019), RetinexNet (Wei et al. 2018), Zero-DCE (Guo et al. 2020), SSIEN (Zhang et al. 2020) and EnlightenGAN (Jiang et al. 2021b). Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) and Feature Similarity (FSIM) (Zhang et al. 2011) are employed for performance evaluation on synthetic datasets. We also conduct additional experiments on real-world low-light datasets and adopt three no-reference quality metrics (Naturalness Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2012), Spatial-Spectral Entropy-based Quality (SSEQ) index (Liu et al. 2014) and user study scores) for evaluation.

3.1 Implementation Details

Data Collection. Following the setting in RetinexNet (Wei et al. 2018), we use the LOL dataset for training, which contains 500 low/normal-light image pairs (480 for training and 20 for evaluation). The competing methods are also trained on LOL using their publicly released codes for a fair comparison. Similar to (Wei et al. 2018; Zhang, Zhang, and Guo 2019), we evaluate our model and other methods on three widely used synthetic low-light datasets: *LOL1000* (Wei et al. 2018), *Test148* (Jiang et al. 2021b) and *VOC144* (Lv et al. 2018). Besides, we also collect 63 real-world low-light samples from (Lee, Lee, and Kim 2012; Ma, Zeng, and Wang 2015), namely *Real63*, for evaluation. Finally, a synthetic low-light dataset (COCO1000) and a real-world low-light dataset (ExDark (Loh and Chan 2019)) are used to verify the performance on both image enhancement and detection tasks under low-light conditions.

Model	DL-DA	MSR	$\mathcal{L}_{KL}$	$\mathcal{L}_{ssim}$	PSNR	SSIM	FSIM
Input	–	–	–	–	12.86	0.610	0.787
w/o DL-DA	×	✓	×	✓	19.03	0.857	0.906
w/o MSR	✓	×	✓	✓	19.32	0.874	0.910
w/o KL	✓	✓	×	✓	19.61	0.881	0.914
w/o SSIM	✓	✓	✓	×	19.56	0.849	0.907
DRGN	✓	✓	✓	✓	**19.88**	**0.889**	**0.916**

Table 1: Ablation study of basic components on LOL1000 dataset. DL-DA, MSR, KL and SSIM stand for the degradation learning and data augmentation, multi-scale representation, KL divergence and SSIM constraint, respectively.

Figure 4: Visualization results on LOL1000 dataset.

Experimental Setup. In our baseline, the pyramid layer is empirically set to 3, corresponding to the number of RCABs and RCAB depths of [2, 3, 4] and [3, 3, 3], respectively, in the residual group. The training images are cropped to non-overlapping 96×96 patches to obtain sample pairs. Standard augmentation strategies, *e.g.*, scaling and horizontal flipping are applied. We use the Adam optimizer with a batch size of 16 for training DRGN on a single NVIDIA Titan Xp GPU. The learning rate is initialized to 5×10^{-4} and then attenuated by 0.9 every 6,000 steps. After 60 epochs on training datasets, we obtain the optimal solution with the above settings. Specifically, we train DeG for the first 20 epochs and then optimize ReG for 40 epochs.

3.2 Ablation Study

Basic Components. Since the proposed DRGN involves multiple enhancement strategies, including degradation learning and data augmentation (DL-DA) and multi-scale representation (MSR), we carry out ablation studies to validate the contributions of individual components to the final enhancement performance. We design two models by removing these two components in turn while keeping the similar computational complexity, termed *w/o* DL-DA and *w/o* MSR. Quantitative results on the LOL1000 dataset are tabulated in Table 1, showing that DRGN containing all modules significantly outperforms its incomplete versions. For example, with the degradation learning and data augmentation, DRGN gains about 0.85dB performance improvements over *w/o* DL-DA. We consider that applying a parametric generator to characterize the implicit distribution of degradation is more reasonable and practical, avoiding the dependence on hand-crafted priors. Furthermore, the training dataset can be automatically augmented and enriched with the newly generated synthetic paired samples, which promote the diversity and abundance of the training dataset. On the other hand, using the multi-scale representation strategy can cope with the learning task in multiple subspaces, thus effectively alleviating the learning difficulty with the joint representation among multi-scale features. The results in Figure 4 also demonstrate that the proposed modules can together produce results with more credible contents and faithful contrast.

Loss Function. We also evaluate the influence of KL divergence and SSIM loss on the enhancement performance by training another two models, *w/o* KL (optimized without KL divergence in DeG) and *w/o* SSIM (optimized without SSIM loss in ReG), for comparison. Comparison results on LOL1000 are shown in Table 1 and Figure 4. KL divergence is used to constrain DeG to keep the distribution consistency between the predicted degradation priors $[I_D, I_{D,resf}]$ and accelerate the training. SSIM loss can guide the model to produce results with better texture fidelity by enforcing structural similarity. Therefore, compared with *w/o* KL or *w/o* SSIM models, DRGN gains notable improvements in quality metrics and visual effects.

3.3 Comparison with State-of-the-Arts

Synthetic Data. Nine representative low-light image enhancement methods are employed for comparison. Quantitative results on LOL1000 (Wei et al. 2018), Test148 (Jiang et al. 2021b) and VOC144 (Lv et al. 2018) datasets are presented in Table 2. Our DRGN significantly outperforms the competing methods among all quality metrics. In particular, DRGN outperforms the current SOTA approaches – KinD++ (Zhang et al. 2021) and EnlightenGAN (Jiang et al. 2021b) – **in PSNR (SSIM) by 0.70dB (0.023) and 3.23dB (0.068) on average while enjoying 42.3% and 93.3% more efficiency on model parameters and computation cost.** These substantial improvements validate the efficacy of the proposed modules (degradation learning and data augmentation, and multi-scale fusion representation) for the accurate recovery of image content and contrast.

Efficiency also plays an important role in evaluating the model performance for real-time applications in particular. To this end, we construct a light-weight model, namely LW-DRGN. Specifically, we simplify the multi-resolution fusion network with the single-branch residual network to reduce the parameter size and the memory footprint. Meanwhile, an attention-guided dense fusion strategy is also used to replace the original skip connection between modules to promote the discriminative representation of multi-stage features. Quantitative results in Table 2 show that LW-DRGN yields an acceptable performance (0.07dB less than KinD on average), but **only requires 8.37%, 23.2% and 69.5% in parameters, computation complexity (Gflops) and inference time.** The high efficiency and effectiveness of LW-DRGN makes it practical for mobile devices and applications requiring real-time throughput.

Qualitative comparisons under different low-light conditions (diverse exposures and degradation) on three synthetic low-light datasets are depicted in Figure 5. The results show that DRGN can cope with all low-light cases, generating results with clear and credible details, as well as better fidelity and naturalness. Although other competitors can light-up contents in the dark regions or produce vivid visual effects (LIME and EnlightenGAN), the predicted results are still corrupted by halo artifacts and noises, and suffer from

Methods	LIME	Zero-DCE	KinD++	MIRNet	SSIEN	EnlightenGAN	DRGN (Ours)	LW-DRGN (Ours)
PSNR ↑	18.35/18.32/19.38	14.42/15.53/15.42	19.18/19.42/22.45	20.34/17.20/19.13	16.23/17.35/18.18	16.34/17.64/19.49	**20.41/19.88/22.86**	18.85/18.89/20.43
SSIM ↑	0.771/0.829/0.757	0.392/0.420/0.348	0.859/0.857/0.842	0.871/0.813/0.781	0.780/0.779/0.727	0.796/0.835/0.790	**0.880/0.889/0.858**	0.851/0.856/0.839
FSIM ↑	0.896/0.871/0.827	0.738/0.741/0.635	0.923/0.911/0.885	0.928/0.872/0.850	0.870/0.830/0.792	0.907/0.884/0.853	**0.940/0.916/0.894**	0.920/0.886/0.882
Average ↑	18.68/0.785/0.864	15.12/0.386/0.704	20.35/0.852/0.906	18.89/0.821/0.883	17.25/0.762/0.830	17.82/0.807/0.881	**21.05/0.875/0.916**	19.39/0.848/0.896
Par.(M) ↓	–	**0.079**	8.274	31.78	0.486	8.64	4.773	0.660
Time (S) ↓	–	**0.010**	0.392	0.630	0.028	0.057	0.152	0.041
GFlops (G) ↓	–	**11.74**	1670.91	554.32	77.86	37.02	110.59	18.15

Table 2: Quantitative performance comparison of the compared methods on Test148/LOL1000/VOC144. LW-DRGN denotes our light-weight model. We also show the number of model parameters (Million), average inference time (Second) and GFlops used for 384 × 384 images, except for the traditional prior-based algorithm LIME. The Bold and underline present the best and second performances, respectively.

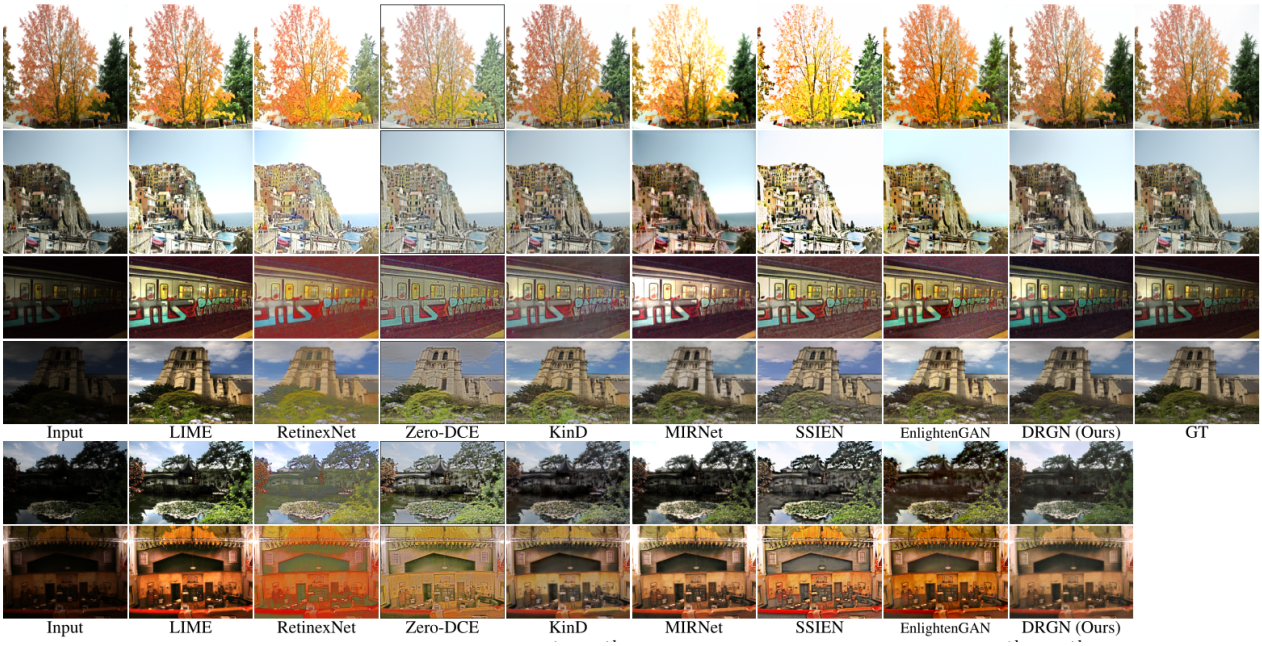

Figure 5: Restoration results on synthetic ($1^{st} - 4^{th}$ rows) and and real-world scenarios ($5^{th} - 6^{th}$ rows).

color distortion as well. For example, as observed from the "Train" and "House" images in Figure 5, only DRGN restores credible image details and approximates the contrast more similar to the ground-truth, while the competitors fail to solve the degradation and their results exhibit apparent color deviation. Moreover, we observe an interesting phenomenon from these results. The competitors tend to produce brighter outputs no matter what level of degradation (slight or severe degradation conditions), since they complete the relighting task using a unified enhancement module (see the first (slight degradation) and third (severe degradation) scenarios in Figure 5). However, DRGN can cope with different degradation conditions and produces results with better naturalness and texture fidelity. This is because DRGN takes advantage of the intrinsic degradation priors of the input and employs the data augmentation to improve the training. Moreover, we also provide the fitting curve based on the histogram of "Y" channel on four synthetic scenarios in Figure 6, showing that our results are close to the distribution of ground truth.

Real-world Data. Although recent years have witnessed significant progress on the low-light image enhancement task, recovering credible contents and visually pleasing contrast from real-world low-light scenarios is still challenging. Inspired by (Lee, Lee, and Kim 2012; Ma, Zeng, and Wang 2015), we conduct experiments on the Real63 dataset to further verify the effectiveness of DRGN. Besides two reference-free metrics (NIQE and SSEQ), we also perform a user study to evaluate the performance quantitatively. Specifically, a total of 30 volunteers independently score the visual quality of the enhanced images in the range of 1.0 to 5.0 (worst to best quality) in terms of the content naturalness (CN) and color deviation (CD). The quantitative results are included in Table 3. Again, our proposed DRGN achieves the best and second-best average scores of NIQE and SSEQ. Moreover, DRGN yields the highest average user study scores, including CN and CD, for a total of 63 scenarios. Figure 5 shows the visual comparison results. DRGN can infer more realistic and credible image details in dark regions and well adjust the image contrast, whereas other methods tend to produce under- or over-exposure results with noticeable color distortion (please refer to the "sky" and "plant" in the 5^{th} scenario.). These results confirm that applying degradation learning and data augmentation can promote robustness and generality in real-world scenarios.

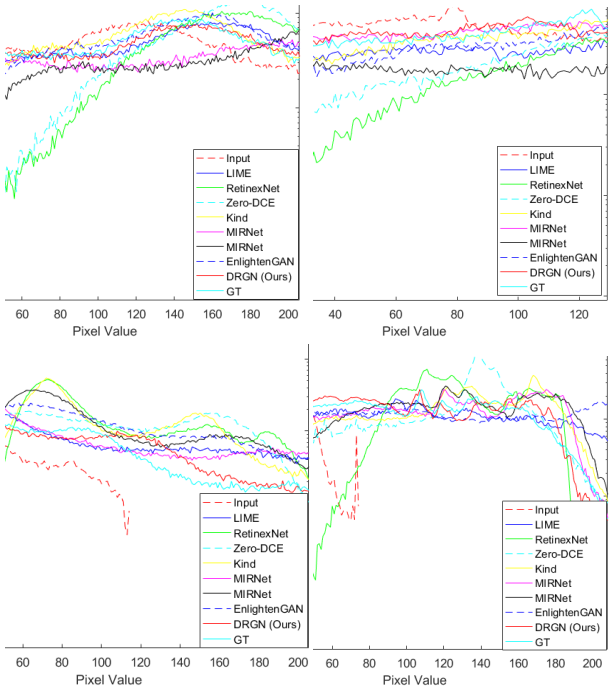

Figure 6: The fitting results based on the histogram curve of Y channel in YCbCr on four synthetic scenarios in Figure 5.

Methods	LIME	KinD	MIRNet	SSIEN	DRGN (Ours)
NIQE ↓	3.117	3.551	3.287	4.153	**3.051**
SSEQ ↓	24.45	29.48	26.69	32.20	**23.53**
CN ↑	3.587	3.458	3.569	3.026	**4.012**
CD ↑	3.531	3.762	3.625	3.148	**3.984**

Table 3: Comparison results on Real63 dataset. CN and CD are the user study scores. Bold and underline present the best and second performances, respectively.

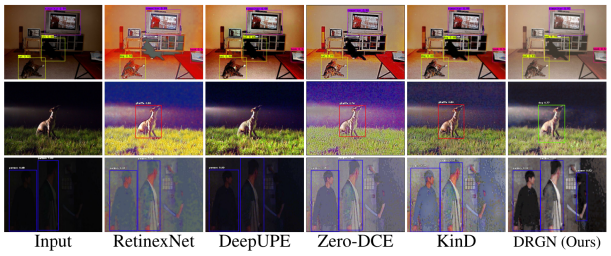

Figure 7: Examples of joint low-light image enhancement and object detection on COCO1000 and ExDark.

3.4 Evaluation via Downstream Vision Tasks

Lighting up the image under low-light conditions while recovering credible textural details is meaningful for many high-level vision applications such as object and pedestrian detection. This motivates us to investigate the impact of image enhancement performance on the accuracy of object detection. Therefore, we conduct experiments for the joint image enhancement and object detection tasks. We adopt the real-world low-light dataset (ExDark (Loh and Chan 2019)) and the synthetic low-light dataset (COCO1000) for evaluation, where the samples are of diverse exposures and degradation and most of them are from night scenes. Our proposed DRGN, KinD (Zhang, Zhang, and Guo 2019), Zero-DCE (Guo et al. 2020) and DeepUPE (Wang et al. 2019) are directly applied to the low-light samples to generate the predicted normal-light ones. We then use the pre-trained YOLOv4 (Bochkovskiy, Wang, and Liao 2020) detector for object detection. Qualitative results, including the image enhancement performance and the precision of object detection, are tabulated in Table 4 (a) and (b). Both the enhancement performance and detection precision on predicted normal-light images by DRGN show a significant improvement over the competing low-light enhancement methods, surpassing the top-performing method KinD by $1.91dB$ and 2.42% in terms of PSNR and mAP on COCO1000 dataset. DRGN also achieves the best results of 63.83% IoU and 67.72% precision on ExDark. Visual comparisons in Figure 7 clearly demonstrate that low-light degradation can greatly affect the readability and recognition, which hinders the object detection accuracy. The enhanced images by our proposed DRGN method obviously promote the detection precision.

	Methods	Input	Zero-DCE	KinD	DRGN
Enhance	PSNR ↑	11.58	15.07	18.56	**20.47**
	SSIM ↑	0.586	0.368	0.748	**0.790**
Detection	Pre. (%)	–	25.72	26.54	**26.81**
	IoU (%)	–	76.19	77.01	**77.35**
	mAP (%)	–	41.85	43.49	**45.91**

(a) COCO1000

	Methods	Input	Zero-DCE	KinD	DRGN
Enhance	NIQE ↓	4.887	3.825	3.770	**3.712**
	SSEQ ↓	31.58	34.27	36.04	**30.65**
Detection	Pre. (%)	63.95	63.91	64.86	**67.72**
	IoU (%)	53.28	58.48	59.29	**63.83**
	mAP (%)	69.34	68.21	71.45	**74.63**

(b) ExDark (Loh and Chan 2019)

Table 4: Comparison results of joint low-light image enhancement and object detection. Image Size: 640× 480; Algorithm: YOLOv4. We report the results on the original images and enhanced images with our DRGN.

4 Conclusion

We proposed a novel Degradation-to-Refinement Generation Network (DRGN) for low-light image enhancement. To achieve high-naturalness recovery, we decompose the enhancement task into two stages: degradation learning and content enhancement. Using degradation learning, a number of low/normal-light image pairs can be generated to augment the sample space to promote the training. Moreover, the multi-resolution fusion strategy further improves the feature representation. Consequently, the recovered normal-light images are more realistic with high naturalness due to the degradation learning, data augmentation and multi-scale feature fusion. In particular, we create two low/normal-light datasets (COCO24700 for training and COCO1000 for evaluation) for the joint low-light image enhancement and detection tasks. Extensive experimental results on synthetic and real-world low-light samples and downstream vision tasks (object and pedestrian detection) demonstrate the superiority of DRGN over the competing methods.

Acknowledgments

This work is supported by National Natural Science Foundation of China (U1903214, 62071339, 61872277, 62072347, 62171325), and Guangdong-Macao Joint Innovation Project (2021A0505080008).

References

Bae, S. 2019. Object Detection Based on Region Decomposition and Assembly. In *AAAI*, 8094–8101.

Bochkovskiy, A.; Wang, C.-Y.; and Liao, H.-Y. M. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv e-prints*, 2004.10934.

Bulat, A.; Yang, J.; and Tzimiropoulos, G. 2018. To Learn Image Super-Resolution, Use a GAN to Learn How to Do Image Degradation First. In *ECCV*, 187–202.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2018. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 1209–1218.

Chen, C.; Chen, Q.; Xu, J.; and Koltun, V. 2018. Learning to see in the dark. In *CVPR*, 3291–3300.

Fu, X.; Liang, B.; Huang, Y.; Ding, X.; and Paisley, J. 2020. Lightweight Pyramid Networks for Image Deraining. *IEEE Transactions on Neural Networks and Learning Systems*, 31(6): 1794–1807.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. In *Bernstein Conference 2015*, 219–219.

Gharbi, M.; Chen, J.; Barron, J. T.; Hasinoff, S. W.; and Durand, F. 2017. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4): 1–12.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *CVPR*, 1777–1786.

Guo, X.; Li, Y.; and Ling, H. 2017. LIME: Low-Light Image Enhancement via Illumination Map Estimation. *IEEE Trans. Image Process.*, 26(2): 982–993.

Huang, W.; Hu, R.; Wang, X.; Liang, C.; and Chen, J. 2021. Occluded suspect search via channel-guided mechanism. *Neural Comput. Appl.*, 33(3): 961–971.

Huang, W.; Liang, C.; Yu, Y.; Wang, Z.; Ruan, W.; and Hu, R. 2018. Video-Based Person Re-Identification via Self Paced Weighting. In *AAAI*, 2273–2280.

Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Wang, Z.; Wang, X.; Jiang, J.; and Lin, C.-W. 2021a. Rain-free and residue hand-in-hand: A progressive coupled network for real-time image deraining. *IEEE Trans. Image Process.*, 30: 7404–7418.

Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Gu, K.; and Jiang, J. 2019. ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination. *IEE Trans. Multim.*, 22(10): 2734–2747.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021b. Enlightengan: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.*, 30: 2340–2349.

Kaur, M.; Kaur, J.; and Kaur, J. 2011. Survey of contrast enhancement techniques based on histogram equalization. *International Journal of Advanced Computer Science and Applications*, 2(7).

Lai, W.-S.; Huang, J.-B.; Ahuja, N.; and Yang, M.-H. 2017. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 624–632.

Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 237(6): 108–129.

Ledig, C.; Theis, L.; Huszr, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; and Shi, W. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *CVPR*, 105–114.

Lee, C.; Lee, C.; and Kim, C.-S. 2012. Contrast enhancement based on layered difference representation. In *ICIP*, 965–968.

Li, C.; Guo, J.; Porikli, F.; and Pang, Y. 2018. LightenNet: a convolutional neural network for weakly illuminated image enhancement. *Pattern Recognition Letters*, 104: 15–22.

Liu, L.; Liu, B.; Huang, H.; and Bovik, A. C. 2014. No-reference image quality assessment based on spatial and spectral entropies. *SPIC*, 29(8): 856–863.

Loh, Y. P.; and Chan, C. S. 2019. Getting to know low-light images with the Exclusively Dark dataset. *Comput. Vis. Image Underst.*, 178: 30–42.

Lv, F.; Lu, F.; Wu, J.; and Lim, C. 2018. MBLLEN: Low-Light Image/Video Enhancement Using CNNs. In *BMVC*, 220.

Ma, J.; Yu, W.; Liang, P.; Li, C.; and Jiang, J. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48: 11–26.

Ma, J.; Zhang, H.; Shao, Z.; Liang, P.; and Xu, H. 2020. GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans. Instrum. Meas.*, 70: 1–14.

Ma, K.; Zeng, K.; and Wang, Z. 2015. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.*, 24(11): 3345–3356.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a completely blind image quality analyzer. *IEEE Signal Process. Lett.*, 20(3): 209–212.

Papyan, V.; and Elad, M. 2016. Multi-Scale Patch-Based Image Restoration. *IEEE Trans. Image Process.*, 25(1): 249–261.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 6849–6857.

Wang, X.; Chen, J.; Wang, Z.; Liu, W.; Satoh, S.; Liang, C.; and Lin, C. 2020. When Pedestrian Detection Meets Nighttime Surveillance: A New Benchmark. In *IJCAI*, 509–515.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *BMVC*, 155.

Wu, J.-H.; and Saito, S. 2017. Interactive relighting in single low-dynamic range images. *ACM TOG*, 36(2): 1–18.

Xu, X.; Liu, L.; Zhang, X.; Guan, W.; and Hu, R. 2021a. Rethinking data collection for person re-identification: active redundancy reduction. *Pattern Recognition*, 113: 107827.

Xu, X.; Wang, S.; Wang, Z.; Zhang, X.; and Hu, R. 2021b. Exploring Image Enhancement for Salient Object Detection in Low Light Images. *ACM Trans. Multim. Comput. Commun. Appl.*, 17(1s): 1–19.

Yang, T.; Zhu, S.; Chen, C.; Yan, S.; Zhang, M.; and Willis, A. 2020a. Mutualnet: Adaptive convnet via mutual learning from network width and resolution. In *ECCV*, 299–315. Springer.

Yang, W.; Wang, S.; Fang, Y.; Wang, Y.; and Liu, J. 2020b. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, 3063–3072.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Lu, T.; and Ma, J. 2020. A progressive fusion generative adversarial network for realistic and consistent video super-resolution. *IEEE Trans. Patt. Anal. Mach. Intell.*

Yu, W.; Yang, T.; and Chen, C. 2021. Towards Resolving the Challenge of Long-tail Distribution in UAV Images for Object Detection. In *WACV*, 3258–3267.

Yue, H.; Yang, J.; Sun, X.; Wu, F.; and Hou, C. 2017. Contrast enhancement based on intrinsic image decomposition. *IEEE Trans. Image Process.*, 26(8): 3981–3994.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M. H.; and Shao, L. 2020. Learning Enriched Features for Real Image Restoration and Enhancement. In *ECCV*, 492–511.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.*, 20(8): 2378–2386.

Zhang, Y.; Di, X.; Zhang, B.; and Wang, C. 2020. Self-supervised image enhancement network: Training with low light images only. *arXiv e-prints*, 2002.11300.

Zhang, Y.; Guo, X.; Ma, J.; Liu, W.; and Zhang, J. 2021. Beyond Brightening Low-light Images. *International Journal of Computer Vision*, 129(4): 1013–1037.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 286–301.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, 1632–1640.

Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; and Lin, C.-W. 2021. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE Trans. Circuits Syst. Video Technol.*

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2020. Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network. In *AAAI*, volume 34, 13106–13113.