DiffLoss: Unleashing Diffusion Model as Constraint for Training Image Restoration Network

Jiangtong Tan, Hu Yu, Jie Huang, Zizheng Yang, and Feng Zhao*

MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
{jttan, yuhu520, hj0117, yzz6000}@mail.ustc.edu.cn, fzhao956@ustc.edu.cn

Abstract. Image restoration aims to enhance low-quality images, producing high-quality images that exhibit natural visual characteristics and fine semantic attributes. Recently, the diffusion model has emerged as a powerful technique for image generation, and it has been explicitly employed as a backbone in image restoration tasks, yielding excellent results. However, it suffers from the drawbacks of slow inference speed and large model parameters due to its intrinsic characteristics. In this paper, we introduce a new perspective that implicitly leverages the diffusion model to assist the training of image restoration network, called DiffLoss, which drives the restoration results to be optimized for naturalness and semantic-aware visual effect. To achieve this, we utilize the mode coverage capability of the diffusion model to approximate the distribution of natural images and explore its ability to capture image semantic attributes. On the one hand, we extract intermediate noise to leverage its modeling capability of the distribution of natural images, which serves as a naturalness-oriented optimization space. On the other hand, we utilize the bottleneck features of diffusion model to harness its semantic attributes serving as a constraint on semantic level. By combining these two designs, the overall loss function is able to improve the perceptual quality of image restoration, resulting in visually pleasing and semantically enhanced outcomes. To validate the effectiveness of our method, we conduct experiments on various common image restoration tasks and benchmarks. Extensive experimental results demonstrate that our approach enhances the visual quality and semantic perception of the restoration network. Code is available at https://github.com/JosephTiTan/DiffLoss.

Keywords: Image restoration · Diffusion model · Perception quality · Low-for-high

1 Introduction

In complex imaging environments, the quality of imaging often suffers from unpredictable degradations, such as low-light conditions, heterogeneous media,

* Corresponding author.

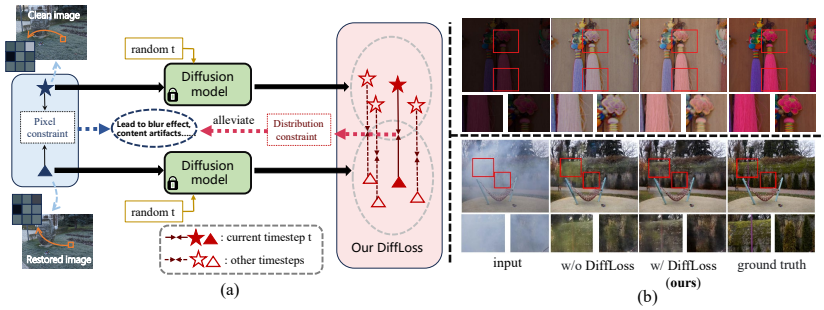
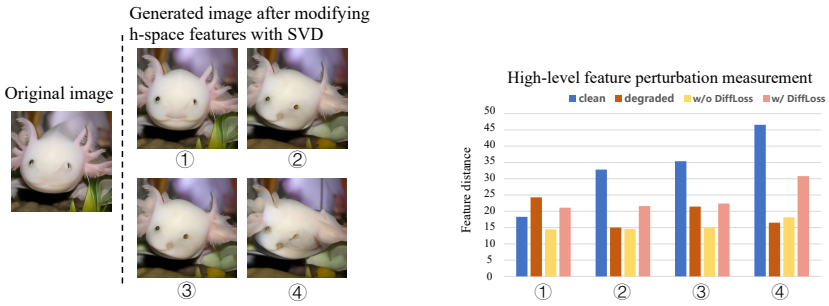


Fig. 1: (a): Illustration of the effect of our method to convert pixel constraint into distribution constraint.(b): Visual comparison between baselines with and without our DiffLoss. The top is produced with IAT [11] on LOL dataset [59] and the bottom is produced with MSBDN [14] on NH-HAZE dataset [2]. Previous loss is limited to the pixel level and suffers from unnaturalness problem, with color shift and content artifacts. Our DiffLoss leverages the powerful modeling capability of the diffusion model on the distribution of natural images, resulting in more natural outcomes in the image restoration process. It is worth mentioning that DiffLoss is an optimization strategy. The improvement should be compared with the baseline method, instead of other restoration methods or the ground-truth images.

leading to information loss both in content and color and unnatural visual effects. Besides, these degradations also causes adverse effects on high-level computer vision tasks such as object detection [8] and scene understanding [50, 51]. Image restoration aims to solve these issues by recovering high-quality images from degradations to achieve better natural visual quality and fine semantic attributes. This task has been a longstanding and challenging problem, drawing the attention of researchers. Previous traditional methods utilized the inherent and general statistical property of degraded images to design algorithms with prior knowledge for image restoration [4, 15, 16, 24]. However, these hand-crafted image priors are drawn from specific observations with limited robustness, which may not be reliable for modeling the intrinsic characteristics of degraded images. With the development of deep learning, image restoration has made marvelous progress, where most of them rely on designing advanced architectures to learn degradations to clean mappings [14, 21, 33, 37, 62, 67]. However, due to the severe degradation and limited model capacity, the naturalness of the restored results is restricted by color and texture distortions, as shown in Fig. 1. In addition to further improving image visual quality, preserving semantic attributes is also significant for image restoration. Existing methods try to solve the problem by either involving multi-stage training or utilizing complex network structures [30, 65, 66], which result in inconvenience during implementation. With the emergence of the diffusion model, it has shown remarkable performance as a backbone in image restoration tasks with naturalness and semantic-aware visual results [10, 29, 38, 49, 56, 61]. However, these methods attempt to integrate dif-



(a) Illustration of images after modifying diffusion model's h-space features with SVD. From ① to ④, the h-space perturbation increases.

(b) Histogram of high-level feature distance of results on low-light image enhancement with synthesized data from ImageNet dataset.

Fig. 2: As h-space changes, the image gradually loses its original semantics completely, from ① to ④ in (a). We also employ the output features of the ResNet50 network to measure the distance of high-level features in the images with different types of degradations to show the change of semantic attributes, as depicted in the histogram (b). With h-space perturbation increase, clean images and restored images with DiffLoss exhibit systematic variations in semantic attributes, while the degraded images and restored images without DiffLoss show minimal changes. It means that degradations undermine the semantic attributes of images, but our DiffLoss can restore this. Note that the diffusion model and ResNet50 are both trained on the ImageNet dataset.

fusion models into the network architecture that has the limit of slow inference speed and accounting huge memory cost when deploying.

In this study, to address these issues, instead of creating new network architectures, we empower existing network frameworks with powerful prior on natural image distribution modeling and high-level semantic space. Inspiringly, the diffusion model has exhibited strong natural image generation capability [13] and possible semantic potential [3], which motivates us to exploit the "implicit" usage of the diffusion model and leverage it as an optimization prior to improve naturalness and semantic attributes for image restoration.

In this paper, we propose a new perspective to address these issues by introducing a naturalness-oriented and semantic-aware optimization mechanism using a diffusion model, dubbed DiffLoss, which has two parts. (1) **Natural image distribution prior.** Drawing inspiration from the diffusion model's remarkable capacity to cover distributions in natural image generation, we leverage the Markov chain sampling characteristic of the diffusion model to project the restored results of existing networks into the noise sampling space and utilize it as a constraint to enhance the natural visual representation of images. Its improvement for natural visual performance can be observed in Fig. 1. (2) **High-level semantic space prior.** The bottleneck feature of diffusion models, also dubbed h-space feature, is verified to be a natural high-level semantic space in [23,27] and is shown in Fig. 2. For both clear images and images with different degradations, we first add noise to them and then input them into the diffusion

model. During the generation, we perform SVD decomposition on the feature in h-space and perturb its principal components. As the perturbation changes, the semantic appearance in the generated images also changes. We extract the high-level features of these images by ResNet50 that is pre-trained on ImageNet for classification and measure the L2 distance of the perturbed image features from original image features. We find the perturbation in h-space can disrupt high-level features of clean images and restored images applying our method. For degraded images and restored images without our method, the disturbance is slight, which implies degradations corrupt the semantic attributes of the images, and DiffLoss can act as a constraint for semantic recovery in image restoration tasks.

For implementation, we employ the unconditional diffusion model pre-trained on the ImageNet dataset as an optimization prior to exploit its immense clean data distribution. The diffusion model takes the restored image and ground truth as inputs and projects them into the distribution sampling space and semantic space by extracting their sampling noise distribution and bottleneck feature. Within these spaces, the empowered restoration network can get a more natural result and preserve more semantic attributes by pulling the intermediate sampling noise distribution and bottleneck feature of the restored image closer to that of the corresponding clear image. Different from previous methods that focus on dedicated model design or directly employing diffusion model as the restoration model, our DiffLoss works as a general and novel auxiliary training mechanism, which can endow existing restoration methods with both more natural and semantic-aware results utilizing this effective training strategy. Additionally, it's especially beneficial for empowering parameter-limited models as it involves naturalness prior. We also compare it with other loss functions that assist restoration networks, demonstrating excellent performance. Moreover, our method does not involve additional computations in the inference stage and is easy for implementation. The effectiveness of our method has been verified on substantial common image restoration tasks, including image dehazing, image deraining, and low-light image enhancement.

Overall, our contributions can be summarized as follows:

- We introduce the naturalness and semantic-aware modeling paradigm into the restoration network by embedding the diffusion model as an auxiliary training mechanism, which has not been explored before. This training strategy alleviates the low-quality issues caused by unnatural visual quality and a lack of semantic attributes in existing methods.
- Specifically, our approach leverages the fixed diffusion model to enable the extraction of intermediate sampling noise and semantic information, and it yields more natural and semantic-aware restoration when the DiffLoss is minimized.
- Extensive experiments demonstrate that our DiffLoss empowers existing restoration methods compared with other loss functions, helps train efficient models, and improves classification ability on data with varying degradations without involving additional computations in the inference stage at all.

2 Related Work

2.1 Visual and Semantic Improvement for Image Restoration

Image restoration refers to the process of improving the quality of a degraded or damaged image by removing various types of degradations. With the advent of deep learning, numerous methods leveraging deep neural networks for image restoration have emerged, such as low-light enhancement [58, 70, 71, 74], super-resolution [36,72,73], dehazing [4,7,15,24,33,34,68], deraining [19,35,39,43,46], deblurring [5, 9, 20, 32, 47]. However, the severe degradation limits the naturalness of restored results, causing color and texture distortions. Moreover, it is proven that solely relying on visual quality metrics during the restoration process without considering the semantic aspects of quality will negatively impact performance in downstream tasks. Some efforts [30,65,66] have attempted to improve the semantic attributes of restored network output images through time-consuming training strategies and complex networks, but they often result in inconvenience and redundancy in practical applications.

2.2 Diffusion Models for Image Restoration

Existing diffusion models used in image restoration tasks can be roughly categorized into two classes: conditional diffusion and unconditional diffusion.

Conditional Denoising Diffusion Probabilistic Models (DDPMs) [10, 29, 42, 48, 49, 60, 61] are usually combined with specific image restoration task, such as image super-resolution [49], image deblurring [61], and image deraining [60]. SR3 trains a conditional diffusion model for image super-resolution with the low-resolution images as condition. Whang et al. [61] proposed the "predict and refine" strategy and learned the residual with conditional diffusion model in image deblurring task. RainDiffusion [60] combines cycle-consistent architecture with diffusion model to achieve unsupervised image deraining.

Unconditional DDPMs [29,38,40,56] are usually integrated into general image restoration task. For example, RePaint [38] solves inpainting problem by employing unconditional diffusion process in the unmasked region and reverse back to solve boundary inconsistency. DDRM [29] uses SVD to decompose the degradation operators and embeds unconditional diffusion model into unsupervised posterior sampling method to solve various linear inverse problems. DDNM [56] applies range-null space decomposition to degraded images and refines only the null-space contents during the reverse process to yield diverse results.

However, these methods aim to incorporate diffusion models as the backbone, which suffer from the drawbacks of slow inference speed and significant memory consumption during deployment. We circumvent these limitations from a new perspective by exploring the diffusion model as an auxiliary training mechanism to empower the learning capability of existing image restoration networks without involving additional computations in the inference stage.

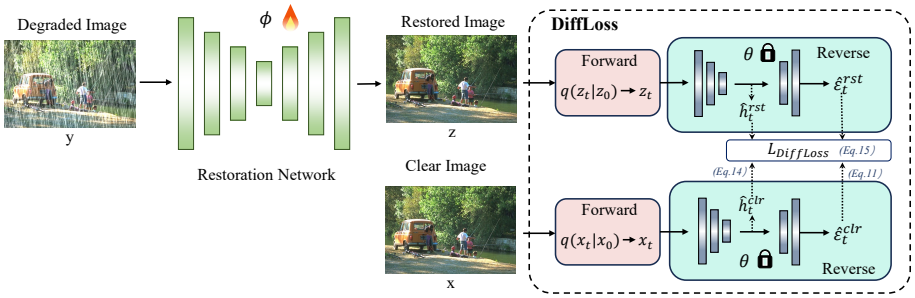


Fig. 3: Overview structure of our method. The parameter of DiffLoss is frozen during the training stage. For any existing restoration network, we train it with the aid of our DiffLoss to achieve higher natural visual and semantic performance. More implementation details of DiffLoss can be found in Fig. 4. During the inference stage, we only have the optimized restoration network without involving the DiffLoss.

3 Methodology

In this section, we first briefly introduce the denoising diffusion probabilistic models, followed by a detailed presentation of the DiffLoss we propose.

3.1 Denoising Diffusion Probabilistic Models

DDPM is a latent variable model specified by a T-step Markov chain, which approximates a data distribution $q(x)$ with a model $f_\theta(\cdot)$. It contains two processes: the forward diffusion process and the reverse denoise process.

The forward diffusion process. The forward diffusion process starts from a clean data sample x_0 and repeatedly injects Gaussian noise according to the transition kernel $q(x_t|x_{t-1})$ as follows:

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I), \tag{1}$$

where α_t can be learned by reparameterization [31] or held constant as hyperparameters, controlling the variance of noise added at each step. From the Gaussian diffusion process, we can derivate closed-form expressions for the marginal distribution $q(x_t|x_0)$ and the reverse diffusion step $q(x_{t-1}|x_t, x_0)$ as follows:

$$q(x_t|x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

$$q(x_{t-1}|x_t, x_0) = N(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t I), \tag{3}$$

where $\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t$, $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$, and $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$.

Note that the above-defined forward diffusion formulation has no learnable parameters, and the reverse diffusion step cannot be applied due to having no

access to x_0 in the inference stage. Therefore, we further introduce the learnable reverse denoise process for estimating x_0 from x_T.

The reverse denoise process. The DDPM is trained to reverse the process in Eq. (1) by learning the denoise network f_θ in the reverse process. Specifically, the denoise network estimates $f_\theta(x_t, t)$ to replace x_0 in Eq. (3). Note that $f_\theta(x_t, t)$ directly predicts the Gaussian noise ε, instead of x_0. While, the estimated ε deterministically corresponds to $\hat{x}_0$ via Eq. (2).

$$\begin{aligned} p_\theta(x_{t-1}|x_t) &= q(x_{t-1}|x_t, f_\theta(x_t, t)) \\ &= N(x_{t-1}; \mu_\theta(x_t, t), \textstyle\sum_\theta(x_t, t)), \end{aligned} \tag{4}$$

$$\mu_\theta(x_t, t) = \tilde{\mu_t}(x_t, x_0), \textstyle\sum_\theta(x_t, t) = \tilde{\beta}_t I. \tag{5}$$

Similarly, the mean and variance in the reverse Gaussian distribution 4 can be determined by replacing x_0 in $\tilde{\mu}_t(x_t, x_0)$ and $\tilde{\beta}_t$ with the learned $\hat{x}_0$

Training objective and sampling process. As mentioned above, $f_\theta(x_t, t)$ is trained to approach the Gaussian noise ε. Thus the final training objective is:

$$L = E_{t,x_0,\varepsilon} \left\| \varepsilon - f_\theta(x_t, t) \right\|_1 . \tag{6}$$

The sampling process in the inference stage is done by running the reverse process. Starting from a pure Gaussian noise x_T, we iteratively apply the reverse denoise transition $p_\theta(x_{t-1}|x_t)$ T times, and finally get the clear output x_0.

3.2 Overview Structure

As presented in Fig. 3, in our main setting, given the degraded image y, its corresponding clear image x, any existing restoration network $g_\phi(\cdot)$, and pre-trained diffusion model $f_\theta(\cdot)$ (with fixed parameter θ), we first get the restored output $z = g_\phi(y)$ from the restoration network. Then DiffLoss works as an auxiliary training mechanism, providing naturalness modeling ability and for restoration networks.

3.3 Detailed Design of DiffLoss

Originally, the forward diffusion process translates a clean data sample x_0 into Gaussian noise ε by gradually adding Gaussian noise to x_0 in a parameter-free manner. The reverse denoise process is trained to sample from Gaussian noise ε to generate clean images via gradually removing noise with the denoise network f_θ. However, both of these two processes have asymmetric input-output pairs and thousands of iterative steps. These properties are not suitable for direct loss design. For example, directly applying the reverse denoise process T times is time-consuming and impacts the backpropagation of gradients. Besides, since the DiffLoss takes the restored result z as input, it should start from the forward process and connect to the reverse process in a proper way. Because a single forward process is parameter-free and can be used as a learnable loss.

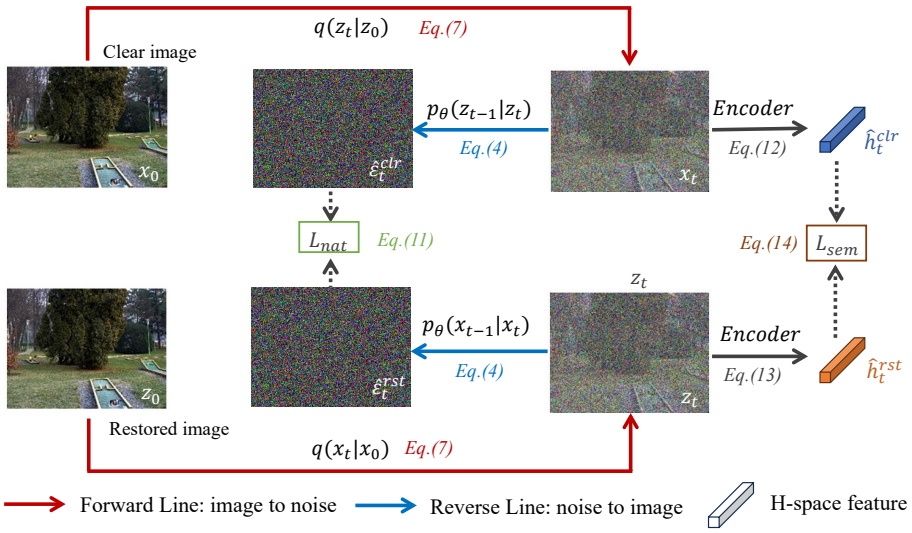


Fig. 4: Detailed design of DiffLoss. We devise the DiffLoss with t-step forward process and one-step reverse process. The t-step forward process can be directly achieved with Eq. (7). Then we project these noisy images into intermediate noise with Eq. (4) after being fed into the denoising UNet. We also get the h-space vector from the bottleneck of the UNet, which contains semantic information, as described in Eq. (12) and Eq. (13). The DiffLoss is designed to pull the output of the denoising UNet and the bottleneck feature closer.

To this end, we redesign the diffusion model delicately. We employ t-step forward diffusion process and one-step reverse denoise step to minimize time-consuming as well as get symmetric image-image input-output pairs. Specifically, as shown in Figs. 3 and 4, we integrate the forward diffusion process and reverse denoise process. In the forward diffusion process, we get the intermediate noisy image x_t via $q(x_t|x_0, \varepsilon)$. Note that t is obtained through uniform distribution sampling, which is expressed as Eq. (7). And Eq. (8) provides the way of reconstructing x_0 back:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1-\bar{\alpha}_t)}\varepsilon, \tag{7}$$

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - \sqrt{(\frac{1}{\bar{\alpha}_t} - 1)}\varepsilon. \tag{8}$$

Then comes the reverse denoise process. We need to implement Eq. (8) in a learnable way. Accordingly, we devise a way by respectively replacing ε with the diffusion model learned ones, as shown in Fig. 3.

Starting from clear image with added noise x_t, we get the pseudo Gaussian noise $\hat{\varepsilon}_t^{clr} = f_\theta(x_t, t)$. Then, the pseudo Gaussian noise map $\hat{\varepsilon}_t^{clr}$ is employed to

replace ε in Eq. (8). The inverse way $q(\hat{x_0}|x_t, \hat{\varepsilon}_t^{clr})$ is expressed as follows:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} x_t - \sqrt{(\frac{1}{\bar{\alpha}_t} - 1)} \hat{\varepsilon}_t^{clr}. \tag{9}$$

Similarly, the output results of the restoration network can undergo the same process to obtain the following equation.

$$\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} z_t - \sqrt{(\frac{1}{\bar{\alpha}_t} - 1)} \hat{\varepsilon}_t^{rst}. \tag{10}$$

Note that we can also obtain $\hat{x}_{t-1}$ and $\hat{z}_{t-1}$ using Eq. (4). We choose to directly constrain on $\hat{\varepsilon}_t^{clr}$ and $\hat{\varepsilon}_t^{rst}$ without reconstructing back to $\hat{x}_0$ and $\hat{z}_0$ or using $\hat{x}_{t-1}$ and $\hat{z}_{t-1}$ as constraint, because it has the best performance and is shown in Sec. 4.3. The loss function can be expressed as follows:

$$L_{nat} = \left\| \hat{\varepsilon}_t^{clr} - \hat{\varepsilon}_t^{rst} \right\|_2. \tag{11}$$

In this way, the restoration networks can harness the generative capabilities of the diffusion model, as well as the sampling space's reflection of the natural attributes of images, to obtain restored results with more naturalness.

If we decompose $f_\theta(x_t, t)$ into an encoder $\mathcal{E}_\theta(\cdot)$ and a decoder $\mathcal{D}_\theta(\cdot)$, then after obtaining the noisy version of clear image $\hat{x}_t$ and restored image $\hat{z}_t$, the form of semantic feature of them can be written as the following equations, respectively:

$$\hat{h}_t^{clr} = \mathcal{E}_\theta(x_t, t), \tag{12}$$

$$\hat{h}_t^{rst} = \mathcal{E}_\theta(z_t, t), \tag{13}$$

where $\hat{h}_t^{clr}$ and $\hat{h}_t^{rst}$ reflect bottleneck features in the middle layers of U-Net from the clear image and restored image, respectively. By reducing the gap between these two terms, we can preserve more semantic information in the restored image, which is expressed as follows:

$$L_{sem} = \left\| \hat{h}_t^{clr} - \hat{h}_t^{rst} \right\|_2. \tag{14}$$

Besides our newly employed DiffLoss, we preserve the traditional L2 loss between z and x for stable optimization. In conclusion, our DiffLoss and total losses used in the training stage are expressed as follows:

$$L_{DiffLoss} = L_{nat} + \lambda L_{sem}, \tag{15}$$

$$L_{total} = \|x - z\|_2 + \gamma L_{DiffLoss}. \tag{16}$$

where λ and γ are weight factors. We set the weights $\lambda = 0.01$ and $\gamma = 0.001$, which are discussed in Sec. 4.3. Besides, we also try adaptive weight, which is correlated with timestep t. Both strategies perform similarly. By employing these two strategies, the DiffLoss has the potential to enhance existing restoration methods, yielding restored images that are not only more natural but also reserve more semantic information.

Table 1: Quantitative comparison on three datasets. ↑ indicates that the larger the value is, the better the performance will be. " $*$ " refers to efficient model of the task.

Dehazing			Deraining			Low-light Enhancement		
Method	Dense-Haze		Method	Rain100H		Method	LOL	
	PSNR↑	SSIM↑		PSNR↑	SSIM↑		PSNR↑	SSIM↑
AOD-Net [33]	13.14	0.4144	DerainNet [18]	14.92	0.5920	EnlightenGAN [28]	17.48	0.6510
FFA-Net [44]	14.39	0.4524	RESCAN [35]	26.36	0.7860	RetiNexNet [58]	16.77	0.5620
AECR-Net [62]	15.80	0.4660	PreNet [46]	26.77	0.8580	DRBN [63]	19.55	0.7460
FSDGN [67]	14.34	0.4010	RCD-Net [55]	17.11	0.4634	DeepLPF [41]	18.44	0.7431
w/DiffLoss	14.66	0.4064	w/DiffLoss	17.67	0.4710	w/DiffLoss	19.51	0.7437
TaylorFormer* [45]	15.02	0.5178	EfDeRain* [22]	23.41	0.7524	IAT* [11]	19.89	0.7371
w/DiffLoss	15.19	0.5326	w/DiffLoss	24.54	0.7656	w/DiffLoss	20.11	0.7378

4 Experiments

In this section, we first introduce the datasets and implement details of our experiment. Then, we conduct experiments on existing restoration methods, including low light enhancement, image deraining and image dehazing, with and without our DiffLoss. To validate the effectiveness of our approach in low-level tasks, we conducted tests on the efficiency model. Additionally, we examined the performance of the model employing our method in the image classification task with degraded data. Experiments on several baselines and benchmarks demonstrate the effectiveness of our DiffLoss.

4.1 Experiment Setup

Datasets. We train and evaluate our models on both synthetic and real-world image restoration datasets, including low light enhancement, image deraining and image dehazing. For real-world challenging scenes, we adopt LOL dataset [59] for low light enhancement and Dense-Haze dataset [1] for image dehazing. For image deraining, we adopt Rain13K dataset [6] for training and Rain100H dataset [64] for testing. Finally, we use CUB dataset [54] for image classification task, and the degraded CUB dataset is obtained through synthetic methods from [26] to simulate degraded conditions, such as fog, rain, and low-light scenarios.
Comparison of baseline methods. We choose several classical and SOTA restoration methods as baselines, including IAT [11], DeepLPF [41] for low light enhancement, EfDeRain [22], RCD-Net [55] for image deraining, and TaylorFormer [45], FSDGN [67] for image dehazing. We separately train these baselines with and without our DiffLoss, with the same settings and implementation details. We compare these two settings on the above baselines qualitatively and quantitatively.
Implementation details. We choose existing restoration networks listed above as backbone. For the DiffLoss, we adopt the unconditional diffusion model pretrained on ImageNet dataset by [13]. Both the network architecture of restoration network and diffusion model need no modification. Besides, we fix the parameters of diffusion model. During the training stage, we use ADAM as the optimizer

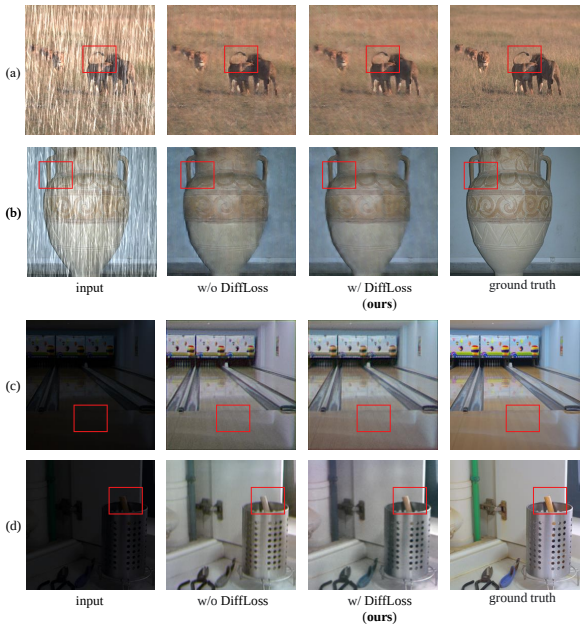


Fig. 5: Comparison of visual results on Rain100H and LOL datasets. (a): EfDeRain; (b): RCD-Net; (c): IAT; and (d): DeepLPF. Please zoom in for best view.

with the learning rate set to 1×10^{-4}. The batch and patch sizes are set to 4 and 256×256, respectively. The parameter size of diffusion model is 552.81 M. For image classification, we use VGG16 [52] and Resnet50 [25] pre-trained on clean CUB dataset as the recognition models to evaluate images restored by different methods. All the restoration models are trained with RTX 4090 GPU.

Evaluation metrics. We evaluate our method on two different metrics: PSNR, SSIM [57], LPIPS [69] and FID [53] which are well-known image quality assessment indicators.

4.2 Comparison with Baseline Methods on Several Benchmarks.

First of all, it is worthy to note that our method improves the naturalness of restored results instead of substantially removing more degradations.

Comparison on real-world degradation images. Table 1 compares the quantitative results of different methods on Dense-Haze, Rain100H, and LOL datasets for image dehazing, image deraining, and low light enhancement, respectively. On these datasets, Baselines with DiffLoss achieves better performance with most metrics. The results on these challenging real-world datasets effectively demonstrate the advantages and effectiveness of our approach.

We also specifically utilized the efficient model, including TaylorFormer, EfDeRain and IAT, and the experiments revealed that our method enables the

Fig. 6: Comparison of visual results on Dense-Haze dataset. We also show the visual comparison between different loss functions and our method.

Table 2: The results of Image Classification on CUB dataset among three different degradations. "Top-1 V" and "Top-1 R" refer to the Top-1 Accuracy (%) on pre-trained VGG16 [52] and Resnet50 [25], respectively.

Config	Dehazing		Deraining		Low-light Enhancement	
	Top-1 V	Top-1 R	Top-1 V	Top-1 R	Top-1 V	Top-1 R
w/o DiffLoss	37.5940	24.3966	68.6724	78.8966	15.2069	28.2069
w/ DiffLoss	46.7105	37.5906	70.0517	79.5690	35.7241	53.5690

efficient model to achieve better image restoration. This is attributed to the inherent generative capability of the diffusion model, which helps the efficient model learn the distribution of natural images.

We also show the visual comparison with other methods on real degradation images sampled from testing sets of Dense-Haze, Rain100H and LOL datasets. The visual results are presented in Figs. 5 and 6. The complex degradation distribution makes these datasets extremely challenging, and existing methods on these datasets usually suffer from unnaturalness problems. As shown in Figs. 5 and 6, these baselines produce relatively pleasing results. However, artifacts and blurs still emerge. In contrast, after being empowered by the DiffLoss, they produce results more visually pleasing than the baseline results. We also present a comparison with different loss functions in Fig. 6, which is described in the following sections. For better visualization, we denote the obvious region with the red rectangular box.

Improvement on image classification. We have demonstrated in Fig. 2 that h-space possesses semantic attributes. Therefore, we utilize h-space as a loss to preserve the semantic information for low-level tasks. We synthesized degraded images with low-light, haze, and rain using the origin CUB dataset. We train the model using common low-level datasets as mentioned before with and without DiffLoss and use the trained model to restore the degraded CUB dataset. We choose to use TaylorFormer, EfDeRain, and IAT for different degradations. Finally, we employ pre-trained VGG16 [52] and Resnet50 [25] networks to eval-

Table 3: The performance comparison of different loss functions on Dense-Haze [1] dataset with MSBDN [14] as the baseline. We train the baseline from scratch and choose the best performance in the first 40K iterations.

Label	LPIPS↓	PSNR↑	SSIM↑
L1 Loss	0.4948	15.228	0.4974
Perceptual Loss	0.4921	15.609	0.5011
Adversarial Loss	0.4848	15.593	0.4981
DiffLoss	0.4731	15.682	0.5088

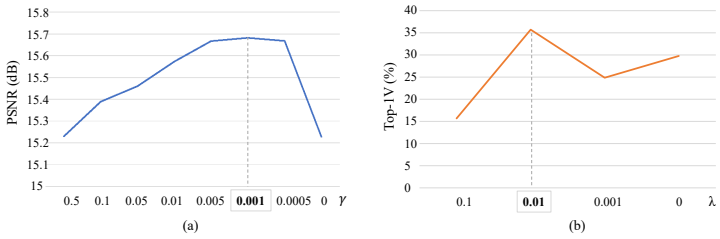


Fig. 7: Graph of PSNR and Top-1V(%) with different weights of DiffLoss during the training process on Dense-Haze [1] dataset with MSBDN [14] as the baseline. We train the baseline from scratch and choose the best performance in the first 40K iterations.

uate the accuracy of classification. From the Table 2, it is evident that with the assistance of our method, the model can preserve more semantic information after the restoration process. Please note that our focus is not on comparing with other methods but rather on improving existing approaches.

Comparison with other loss functions. Previous loss [12, 17] that assist restoration networks have the following drawbacks: (1) L1 or L2 loss works in pixel space, which may produce images deviated from natural distribution. (2) The VGG16 used in perceptual loss is pre-trained for high-level tasks, instead of low-level image restoration tasks. (3) Adversarial loss treats restoration network as generator and inserts an additional discriminator network and needs to train a discriminator for every restoration dataset, which is troublesome and time-consuming. The wide distribution and mode convergence property enables the diffusion model to be a powerful and general image prior, and fits for both general and specific low-level image restoration tasks. Besides, h-space in the diffusion model is also found to reflect the semantics of images. As shown in Fig. 6 and Table 3, our DiffLoss demonstrates excellent performance compared to other approaches.

By pulling the intermediate sampling stages and h-space closer to that of clear images and leveraging the distribution sampling property of diffusion model, the restored results can be optimized to be more natural and recognition-aware, which is difficult to achieve for the previous methods.

Table 4: The results of ablated models on Dense-Haze [1] dataset with GridNet [37] as the baseline.

Label	FID↓	PSNR↑	SSIM↑
a	429.73	14.010	0.3681
b	343.31	14.824	0.4429
c (**Ours**)	293.01	14.907	0.4666

4.3 Ablation Study

In this section, we perform several ablation studies to analyze the effectiveness of the proposed DiffLoss on Dense-Haze dataset. The studies include the following ablated models: (a) With constraints on $\hat{x}_0$ only. (b) With constraints on $\hat{x}_{t-1}$ only. (c) Ours (final setting). These models are trained using the same training setting as our method. The performance of these models is summarized in Table 4. Obviously, every design component in DiffLoss can elevate the performance.

We also experiment with different loss values to get the optimal one. As shown in Fig. 7, we experiment with different weights of DiffLoss and plot the PSNR-Loss weight curve. DiffLoss gets the optimal performance with loss weight $\gamma \in [0.0005, 0.005]$. In our method, we choose the loss weight of DiffLoss to be 0.001. This experiment is conducted with $\lambda = 0$. For the optimal choice of λ, we find that when $\lambda = 0.01$, it achieves the best performance. Note that in the experiments for selecting λ, we set $\gamma = 0.001$.

5 Conclusion

In this paper, we propose a new perspective from the correlation of degraded and natural image distribution that achieves effective image restoration. To achieve this, inspired by the powerful capability of the diffusion model for natural image sampling and generation, we embed the pre-trained diffusion model into the restoration network as an auxiliary training mechanism to empower the learning capability and semantic attributes of neural networks for effective naturalness image restoration. By equipping existing restoration networks with the DiffLoss in the training stage, we can substantially elevate their performance and yield more natural and semantic-aware restored images without involving additional computations in the inference stage.

Acknowledgments. This work was supported by the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

1. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense haze: A benchmark for image dehazing with dense-haze and haze-free images. In: Proceedings of the IEEE International Conference on Image Processing. pp. 1014–1018 (2019)
2. Ancuti, C.O., Ancuti, C., Timofte, R.: NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 444–445 (2020)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
4. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1674–1682 (2016)
5. Chakrabarti, A.: A neural approach to blind motion deblurring. In: Proceedings of the European Conference on Computer Vision. pp. 221–235. Springer (2016)
6. Chen, L., Lu, X., Zhang, J., Chu, X., Chen, C.: HINet: Half instance normalization network for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–192 (2021)
7. Chen, W.T., Ding, J.J., Kuo, S.Y.: Pms-net: Robust haze removal based on patch map for single images. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 11681–11689 (2019)
8. Chen, Y., Li, W., Sakaridis, C., Dai, D., Van Gool, L.: Domain adaptive faster r-cnn for object detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3339–3348 (2018)
9. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4641–4650 (2021)
10. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
11. Cui, Z., Li, K., Gu, L., Su, S., Gao, P., Jiang, Z., Qiao, Y., Harada, T.: You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction. arXiv preprint arXiv:2205.14871 (2022)
12. Deng, Q., Huang, Z., Tsai, C.C., Lin, C.W.: HardGAN: A haze-aware representation distillation gan for single image dehazing. In: Proceedings of the European Conference on Computer Vision. pp. 722–738. Springer (2020)
13. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
14. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2157–2167 (2020)
15. Fattal, R.: Single image dehazing. ACM Transactions on Graphics (TOG) **27**(3), 1–9 (2008)
16. Fattal, R.: Dehazing using color-lines. ACM Transactions on Graphics (TOG) **34**(1), 1–14 (2014)
17. Fu, M., Liu, H., Yu, Y., Chen, J., Wang, K.: DW-GAN: A discrete wavelet transform gan for nonhomogeneous dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–212 (2021)

18. Fu, X., Huang, J., Ding, X., Liao, Y., Paisley, J.: Clearing the skies: A deep network architecture for single-image rain removal. IEEE Transactions on Image Processing **26**(6), 2944–2956 (2017)
19. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 3855–3863 (2017)
20. Gao, H., Tao, X., Shen, X., Jia, J.: Dynamic scene deblurring with parameter selective sharing and nested skip connections. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 3848–3856 (2019)
21. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5812–5820 (2022)
22. Guo, Q., Sun, J., Juefei-Xu, F., Ma, L., Xie, X., Feng, W., Liu, Y., Zhao, J.: Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 1487–1495 (2021)
23. Haas, R., Huberman-Spiegelglas, I., Mulayoff, R., Michaeli, T.: Discovering interpretable directions in the semantic latent space of diffusion models. arXiv preprint arXiv:2303.11073 **3**(6) (2023)
24. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(12), 2341–2353 (2010)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
26. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
27. Jeong, J., Kwon, M., Uh, Y.: Training-free content injection using h-space in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5151–5161 (2024)
28. 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. IEEE Transactions on Image Processing **30**, 2340–2349 (2021)
29. Kawar, B., Elad, M., Ermon, S., Song, J.: Denoising diffusion restoration models. arXiv preprint arXiv:2201.11793 (2022)
30. Kim, I., Han, S., Baek, J.w., Park, S.J., Han, J.J., Shin, J.: Quality-agnostic image recognition via invertible decoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12257–12266 (2021)
31. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
32. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8183–8192 (2018)
33. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: AOD-Net: All-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4770–4778 (2017)
34. Li, B., Gou, Y., Liu, J.Z., Zhu, H., Zhou, J.T., Peng, X.: Zero-shot image dehazing. IEEE Transactions on Image Processing **29**, 8457–8466 (2020)

35. Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: Proceedings of the European Conference on Computer Vision. pp. 254–269 (2018)
36. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops. pp. 136–144 (2017)
37. Liu, X., Ma, Y., Shi, Z., Chen, J.: GridDehazeNet: Attention-based multi-scale network for image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7314–7323 (2019)
38. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
39. Luo, Y., Xu, Y., Ji, H.: Removing rain from a single image via discriminative sparse coding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3397–3405 (2015)
40. Mei, K., Nair, N.G., Patel, V.M.: Bi-noising diffusion: Towards conditional diffusion models with generative restoration priors. arXiv preprint arXiv:2212.07352 (2022)
41. Moran, S., Marza, P., McDonagh, S., Parisot, S., Slabaugh, G.: Deeplpf: Deep local parametric filters for image enhancement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12826–12835 (2020)
42. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. arXiv preprint arXiv:2207.14626 (2022)
43. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2482–2491 (2018)
44. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: FFA-Net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11908–11915 (2020)
45. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z.: Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12802–12813 (2023)
46. Ren, D., Zuo, W., Hu, Q., Zhu, P., Meng, D.: Progressive image deraining networks: A better and simpler baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3937–3946 (2019)
47. Ren, W., Zhang, J., Pan, J., Liu, S., Ren, J.S., Du, J., Cao, X., Yang, M.H.: Deblurring dynamic scenes via spatially varying recurrent neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 3974–3987 (2021)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
49. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4713–4726 (2022)
50. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: Proceedings of the European Conference on Computer Vision. pp. 687–704 (2018)
51. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision **126**(9), 973–992 (2018)

52. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
53. Soloveitchik, M., Diskin, T., Morin, E., Wiesel, A.: Conditional frechet inception distance. arXiv preprint arXiv:2103.11521 (2021)
54. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
55. Wang, H., Xie, Q., Zhao, Q., Meng, D.: A model-driven deep neural network for single image rain removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3103–3112 (2020)
56. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. arXiv preprint arXiv:2212.00490 (2022)
57. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**, 600–612 (2004)
58. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arxiv 2018. arXiv preprint arXiv:1808.04560
59. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018)
60. Wei, M., Shen, Y., Wang, Y., Xie, H., Wang, F.L.: Raindiffusion: When unsupervised learning meets diffusion models for real-world image deraining. arXiv preprint arXiv:2301.09430 (2023)
61. Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A.G., Milanfar, P.: Deblurring via stochastic refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16293–16303 (2022)
62. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10551–10560 (2021)
63. Xie, S., Ma, Y., Xu, W., Qiu, S., Sun, Y.: Semi-supervised learning for low-light image enhancement by pseudo low-light image. In: Proceedings of the 16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. pp. 1–6 (2023)
64. Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1357–1366 (2017)
65. Yang, Y., Wang, C., Liu, R., Zhang, L., Guo, X., Tao, D.: Self-augmented unpaired image dehazing via density and depth decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2037–2046 (2022)
66. Yang, Z., Huang, J., Chang, J., Zhou, M., Yu, H., Zhang, J., Zhao, F.: Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14059–14070 (2023)
67. Yu, H., Zheng, N., Zhou, M., Huang, J., Xiao, Z., Zhao, F.: Frequency and spatial dual guidance for image dehazing. In: Proceedings of the European Conference on Computer Vision. pp. 181–198. Springer (2022)
68. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3194–3203 (2018)
69. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)

70. Zhang, Y., Guo, X., Ma, J., Liu, W., Zhang, J.: Beyond brightening low-light images. International Journal of Computer Vision **129**, 1013–1037 (2021)
71. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 1632–1640 (2019)
72. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision. pp. 286–301 (2018)
73. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2472–2481 (2018)
74. Zhu, A., Zhang, L., Shen, Y., Ma, Y., Zhao, S., Zhou, Y.: Zero-shot restoration of underexposed images via robust retinex decomposition. In: 2020 IEEE International Conference on Multimedia and Expo. pp. 1–6 (2020)