This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Diffusion in the Dark: A Diffusion Model for Low-Light Text Recognition

Cindy M. Nguyen
Stanford University
cindyn@stanford.edu

Eric R. Chan
Stanford University
erchan@stanford.edu

Alexander W. Bergman
Stanford University
awb@stanford.edu

Gordon Wetzstein
Stanford University
gordonwz@stanford.edu

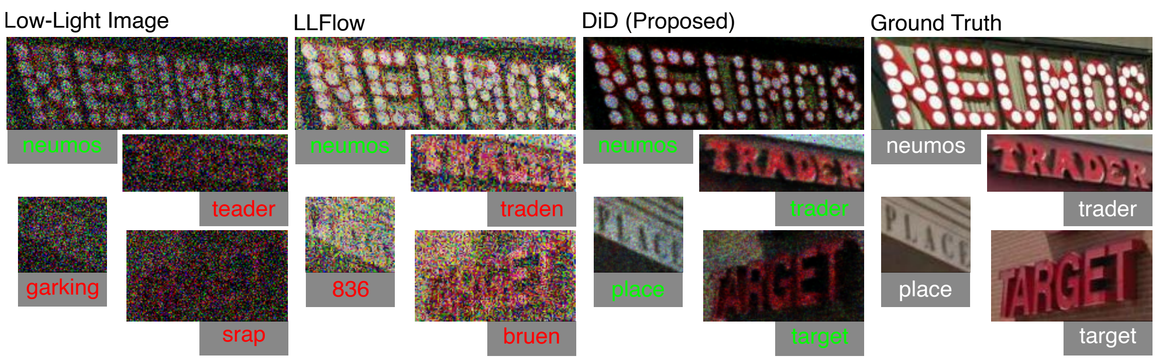

Figure 1. Low-light image reconstruction methods often are not able to recover the fine detail necessary for high-level tasks. We propose a diffusion model to reconstruct not only suitable well-lit images, but also fine details required for text recognition. Here, we recover fine details around lettering, allowing downstream models to accurately identify text. Red and green signify incorrect and correct predictions, respectively. White text is ground truth.

Abstract

Capturing images is a key part of automation for high-level tasks such as scene text recognition. Low-light conditions pose a challenge for high-level perception stacks, which are often optimized on well-lit, artifact-free images. Reconstruction methods for low-light images can produce well-lit counterparts, but typically at the cost of high-frequency details critical for downstream tasks. We propose Diffusion in the Dark (DiD), a diffusion model for low-light image reconstruction for text recognition. DiD provides qualitatively competitive reconstructions with that of state-of-the-art (SOTA), while preserving high-frequency details even in extremely noisy, dark conditions. We demonstrate that DiD, without any task-specific optimization, can outperform SOTA low-light methods in low-light text recognition on real images, bolstering the potential of diffusion models to solve ill-posed inverse problems. Our code and pretrained models can be found on `https://ccnguyen.github.io/diffusion-in-the-dark/`.

1. Introduction

Task automation has become ubiquitous. From the reading of license plates on highways to identifying groceries in a self-checkout line, automated tasks, powered by artificial intelligence, are everywhere and extremely reliant on visual cues, such as RGB images. However, real-world imaging is subject to noisy conditions, optical blurs, and other aberrations that make downstream applications challenging. Notably, image post-processing pipelines used to improve the quality of these images are often designed to fulfill perceptual and aesthetic requirements, as decided by a human expert. While these images may be useful for observation, said post-processing can fail to preserve high-frequency details, which may not be necessary for viewing pleasure but are critical for downstream applications, such as text recognition.

A particular challenge arises in low-light conditions. Low-light images can have extremely low-photon counts, making it difficult to resolve the low signal-to-noise ratios [9, 84]. Convolutional neural networks (CNNs) have emerged as useful tools for low-light reconstruction [10, 22, 50, 53, 83]. However, they are not very robust in low light as they fail to hallucinate details when there is very low signal to work from. Generative models, on the other hand, have proven successful at recovering signals from low light, thanks to their ability to model a distribution of well-lit images. These include generative adversarial networks (GANs) [33] and normalizing flows [78]. These methods are often designed to recover aesthetics, as seen in LLFlow [78] in Figure 1. We seek a method that not only recovers a well-lit image, but one that reconstructs high-frequency details useful for high-level tasks, such as text recognition. We focus on text recognition specifically because it requires fine details more so than other tasks, such as segmentation, as the goal is to predict entire words correctly.

Among the classes of generative models are diffusion models [28, 70, 71], which iteratively denoise from random noise to reconstruct desired data samples. Compared to other generative models, diffusion models are stable in training and provide diversity in reconstructions, which allows a higher probability of reconstructing an optimal signal. Thus, we propose Diffusion in the Dark (DiD), a diffusion-based method for low-light image reconstruction. We train a diffusion model to reconstruct well-lit images that not only are aesthetically pleasing but also preserve fine-grain detail necessary for text recognition better than state-of-the-art (SOTA) low-light methods do. Specifically, we make the following contributions:

- We introduce a novel low-light reconstruction method for text recognition using a conditional diffusion model. DiD can reconstruct images at different resolutions, while training only on patches, reducing training time and computational cost.
- We introduce key normalizations for training diffusion models on extremely dark or right-tailed data.
- Through evaluations of baselines and ablation studies, we demonstrate that DiD provides the best reconstruction of low-light images for text recognition, without any task specific design, when compared to that of SOTA reconstruction methods, while not reducing the aesthetic quality.

Without optimizing for a specific task, we demonstrate that DDPMs can reconstruct high-frequency detail better than exisiting generative models, and they show promise for other high-level tasks. DiD provides competitive quantitative results with the SOTA and consistently preserves high-frequency details in extremely dark, noisy conditions. We also show DiD performs well in reconstructing from unseen, real low-light scenes.

Diffusion models offer a new promising avenue for image reconstruction, one that is easy to train and can obtain better sample quality [16] over other generative models. It is vital to understand their potential and limitations in corner cases, such as low light.

As more images are consumed by high-level perception stacks, we must examine how to better design reconstruction methods for complex tasks, and our work provides an encouraging step in that direction.

2. Related Work

Low-light image enhancement. Classical low-light reconstruction methods include histogram equalization-based methods [1, 32] and Retinex-based methods [21, 23, 43, 59]. The former performs a global transformation of an image using color histograms, while Retinex-based methods decompose light into reflectance and illuminance properties and use these as bases for reconstruction. Burst averaging can also be used to mitigate noise in low-light scenarios [25, 45, 48, 54], but these methods typically require extensive alignment procedures during post-processing to prevent ghosting artifacts and multiple photos. We opt to do single-image reconstruction.

Newer approaches use deep learning to not only advance aforementioned classical methods [22, 47, 75, 83], but also bring new levels of robustness against extreme noise in low light. Zhang et al. [95] use a network, KinD, to decouple illumination and reflectance. To fix non-uniform lighting artifacts in KinD, Zhang et al. [94] developed KinD++, which uses multi-scale attention. Wang et al. [78] developed LLFlow, using normalizing flows to capture the manifold of well-lit images by mapping them to a Gaussian distribution. Given the difficulty of acquiring paired low-light/well-lit images, Jiang et al. [33] propose an unsupervised GAN, using a global-local–focused discriminator and self-regularizing attention maps. Zhou et al. [96] perform low-light reconstruction in conjunction with deblurring to address both problems. Concurrent with our work, Yuan et al. [90] use conditional Denoising Diffusion Probabilistic Models (DDPMs) with stochastic corruptions during training to enhance night sky appearance on a small-scale dataset. Their method focuses on hallucinating plausible star appearances, while we focus on recovering exposures and white balancing for visually appealing, diverse scenes.

Diffusion models. Diffusion models are a rising form of probabilistic generative models which can generate diverse, high-resolution images [16]. Diffusion models take many different forms including DDPMs [28], score-based generative modeling [71], and stochastic differential equations [73]. They all follow similar processes: a forward process which gradually adds noise to clean samples drawn from a prior distribution and a reverse process which reverses the corruption process to recover plausible samples from noise. Diffusion models have been successful at many challenging image-based tasks such as unconditional image generation [63], inpainting [6, 37, 51, 65], colorization [37, 65], image segmentation [2, 5], and medical imaging [72, 86]. We refer the reader to a survey [14] for more applications. Diffusion models offer an attractive alternative to other generative models, such as GANs and variational autoencoders (VAEs), thanks to their stability in training and ability to learn strong priors [16, 58]. We focus on DDPM, which uses a U-Net [64], simplifying the need for task-specific, tedious architecture design [65].

We highlight that generative models are known to generally perform *worse* on traditional metrics such as PSNR/SSIM. An L2 loss minimizes mean squared error (MSE), which conveniently maximizes PSNR. However, probabilistic generative models optimize for learning a representative distribution rather than learning a deterministic solution [31, 74], which would maximize MSE. *Thus, we are not optimizing for high PSNR/SSIM, nor do we expect that predictions from generative models provide the best PSNR/SSIM.* It is well known that MSE, and therefore PSNR, cannot capture perceptual similarities [18, 24, 68, 79–81]. Higher PSNR also does not necessarily correspond to greater photorealism [11, 42]. LPIPS [92] and FID/KID [8, 27] are more representative metrics. However, generative models are able to predict a wide range of exposure levels, and LPIPS is sensitive to different exposure levels, despite many exposure levels providing reasonable reconstructions. See the supplement for more details.

Domain transfer to text recognition. Past works have demonstrated that perceptual metrics such as PSNR/SSIM are not indicative of success in downstream high-level tasks, such as image classification and segmentation [17]. Task-specific imaging is an emerging paradigm, combating domain shift that high-level models experience on degraded images [39, 61, 77]. Diamond et al. [17] demonstrate that optimizing for perceptual metrics specifically can throw away details necessary for successful classification.

Modern scene text recognition (STR) methods [7, 20, 44, 44, 62, 88, 89, 91], as noted in a recent survey [49], only consider well-lit conditions. Their performances tend to suffer in uneven or poor lighting. Xue et al. [87] combine spatial- and frequency-based features to enhance details for recognition of low-light text. However, their method does not report very high precision or recall on standard well-lit text datasets. Hsu et al. [30] use a text-based loss, simulating low light from the ICDAR 2015 dataset [34]. Liu et al. [46] perform text recognition using feature pyramids. We demonstrate that, without any task-specific engineering, we reconstruct fine details to perform robustly in dark, noisy conditions using SOTA text recognition methods [4, 19, 20, 26, 57], such as PARSeq [7].

3. Method

We propose training a single DDPM to recover high-frequency details of full-resolution low-light image (Fig. 2). Training a full-resolution diffusion model is extremely computationally demanding, requiring days of training on multiple GPUs. Prior methods address this by using a cascading strategy, either training a single model in multiple phases [90] or training multiple models, each operating at a different resolution [29, 66]. We train a single model at multiple resolutions simultaneously, using a multi-scale patch-based approach. We describe the key design choices to train a single model on multiple scales that allows us to train on a single GPU (Sec. 3.1.1), the conditioning used in training (Sec. 3.1.2), and inference process to successively predict larger resolutions to get our final reconstruction (Sec. 3.1.3). We also describe our normalization scheme, which allows us to train on right-tailed data (Sec. 3.2).

3.1. Background

Diffusion models have different model families, including Variance Preserving (VP) [73], Variance Exploding (VE) [73], and Elucidating Diffusion Models (EDM) [36]. We use the EDM formulation which includes applying a higher-order Runge-Kutta method for sampling, preconditioning, and improved loss function.

Specifically, Karras et al. formulate their DDPM $\mathcal{D}(\boldsymbol{x};\sigma)$, where $\boldsymbol{x}$ is the noisy image and σ is the noise level, as a function that minimizes the expected MSE denoising error for samples drawn from the clean data distribution $\boldsymbol{p}_{\text{data}}$. The preconditioning, which uses a noise-level–independent skip-connection, allows $\mathcal{D}$ to estimate either the clean image $\boldsymbol{y}$ or noise $\boldsymbol{n}$. We choose to optimize a loss according to a prediction of $\boldsymbol{y}$, which can be expressed as

$$\mathbb{E}_{\sigma,\boldsymbol{y},\boldsymbol{n}}\left[\lambda(\sigma)||\mathcal{D}(\boldsymbol{y}+\boldsymbol{n};\sigma)-\boldsymbol{y}||_2^2\right], \tag{1}$$

where $\boldsymbol{y} \sim p_{\text{data}}$ and $\boldsymbol{n} \sim \mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$. This formulation allows us to use additional guiding losses on the predicted clean image $\boldsymbol{y}$.

3.1.1 Design Choices and Architecture

We discuss two key design choices for our method. The first is to operate on multiple scales. We found that training a 256×256 model to perform low-light reconstruction was too memory-intensive, so we opted to work on 32×32 patches. However, decomposing a low-light image to 32×32 patches, running DDPM on each patch, and then

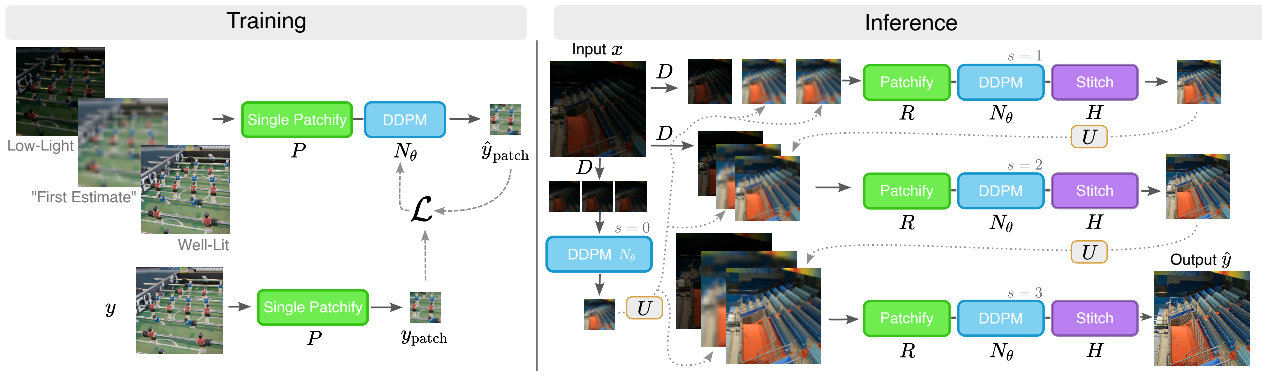

Figure 2. **Overview of DiD pipeline.** During training, we randomly crop 32×32 patches at multiple scales (noted with s) using Single Patchify and concatenate the low-light patches, low-resolution well-lit patches, and high-resolution well-lit patches together as conditioning images to denoise and reconstruct well-lit patches. During inference, we use our trained DDPM network N_θ on 4 scales successively, each time to get well-lit patches at progressively better resolution. The prediction at each scale is upsampled using U to use in the next scale.

stitching the patches together led to patch-to-patch inconsistencies in exposures and white balancing. See the supplement for examples. Here, there is no constraint to enforce all patches to have the same appearance. Thus, we have multiple scales, each using the recovered exposure from the first inference step and, later in the inference process, other recovered exposures from previous steps as conditioning.

A second choice is to have a single network training on randomly selected scales rather than have a network for each scale. While this may have resolved some scale ambiguities in prediction, we found that training 4 models, one for each scale, would take 4 times as much memory or time to achieve the same number of training iterations and reconstruction quality as 1 model, as shown in our ablations (Sec. 4.4).

3.1.2 Training Phase

Given low-lit/well-lit training pairs $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{R}^{256\times256\times3}$, let $\mathcal{S}$ be a random variable following the discrete uniform distribution over the set $\{0,1,2,3\}$. In each training iteration, we sample a random scale $s \sim \mathcal{S}$ and use the function $\gamma(s) = 2^{s+5}, \gamma : s \rightarrow Y$, in which Y is the set $\{32, 64, 128, 256\}$, to acquire a fixed resolution for scale s. These values are set to be multiples of the smallest patch size to simplify the upsampling further on. We use three conditioning inputs in each iteration:

- Low-light condition c_x: The low-light measurement. This image provides the basis for reconstruction. For each scale, we will downsample the measurement to the corresponding operating resolution $\gamma(s) \times \gamma(s)$.
- Well-lit condition c_{y_1}: The well-lit, but low-resolution prediction from the previous scale. This image provides an exposure level to condition from, which is closer to ground truth than c_x, but without well-lit high-frequency detail.
- First estimate condition c_{y_2}: The well-lit, but low-resolution prediction from $s = 0$. This image provides a globally uniform exposure level on which to condition, further constraining the recovered exposure level. During training, if $s \neq 0$, c_{y_2} is simulated by downsampling then upsampling the well-lit ground truth image.

Notably, if $s = 0$, we do not have a c_{y_1} or c_{y_2} from a previous scale to condition on, so we define our conditioning x_{patch_i} using the low-light input x_i and a bilinear downsampling operation $D(x, k)$ to reduce the low-light input to resolution $k \times k$. The conditioning input can be written as

$$x_{\text{patch}_i} = ([D(x_i, \gamma(0)), D(x_i, \gamma(0)), D(x_i, \gamma(0))]. \quad (2)$$

Note that $\gamma(0) \times \gamma(0)$ or 32×32 will be our fixed training resolution. We define $P(x)$ as a random 32×32 cropping function (called "Single Patchify"), $U(x, k)$ as an upsampling operation to bring x to resolution $k \times k$, and our noise as $\eta \sim \mathcal{N}(0, \sigma^2)$. If $s > 0$, we define the training pairs as

$$c_x = D(x_i, \gamma(s)), \quad (3)$$
$$c_{y_1} = U(D(y_i, \gamma(s-1)), \gamma(s)) + \eta, \quad (4)$$
$$c_{y_2} = U(D(y_0, \gamma(0)), \gamma(s)) + \eta, \quad (5)$$
$$(x_{\text{patch}_i}, y_{\text{patch}_i}) = (P([c_x, c_{y_1}, c_{y_2}]_i), P(y_i)). \quad (6)$$

We add η to better resemble the noisy predictions to later scales in the inference phase. We then apply the forward diffusion process of adding noise to x_{patch_i} [36]. We pass the corrupted images and a noise channel $n \in \mathcal{R}^{32\times32\times3}$ to our denoising network Ψ_θ to produce a reconstruction:

$$\hat{y}_{\text{patch}_i} = \Psi_\theta([x_{\text{patch}_i}, n]). \quad (7)$$

During training, we train on only patches of 32×32, but randomly select the scale of each image. This builds a robust model capable of reconstructing at multiple resolutions.

We evaluate the loss $\mathcal{L}_{\text{DiD}}(\hat{y}_{\text{patch}_i}, y_{\text{patch}_i})$ to learn the weights θ, using Eqn. (1). Denoising alone was not sufficient for the challenging task of low-light reconstruction. We add additional losses to each denoising step on the predicted clean image, applying MSE and LPIPS [92]. We chose MSE to help reconstruct better sharp details in the images and LPIPS [92] to enhance appearance. We find that these losses increase text recognition accuracy in addition to improving overall reconstruction.

3.1.3 Inference Phase

We follow a cascaded approach to regress our final image using a single model (Algorithm 1). We begin with the known low-light measurement x_i and compose the conditioning inputs. We apply the reverse diffusion process, and use the diffusion prediction at the current scale as input to the next scale. We continue this until we compose our final 256×256 resolution well-lit image. We observed that, even though the predictions are based on the same exposure instantiation under conditioning, exposure levels and white balancing still varied from patch-to-patch. To achieve full resolution consistency, we needed an additional step: Iterative Latent Variable Refinement (ILVR) [13]. ILVR guides the generative process by blending the high frequencies of the current denoised estimate with the noised, low-resolution version of the reference image. At each step of reverse denoising, we replaced the low-frequency details of the prediction at the current scale with low-frequency content extracted from our conditioning image: the low-resolution but well-lit reconstruction from the previous scale. This conditioning does not require any additional training as it is only used during inference.

Algorithm 1 Inference pipeline in DiD. $R(x)$ decomposes the images into M 32×32 patches ("Patchify"). $H(x)$ stitches the M reconstructed patches to a full resolution image.

$x^0_{\text{patch}} \leftarrow [D(x_i, \gamma(0)), D(x_i, \gamma(0)), D(x_i, \gamma(0))]$
$y^0_{\text{patch}} \leftarrow \Psi_\theta([x^0_{\text{patch}_i}, n])$ ▷ Enhance
for $s = 0, k{+}{+}$, while $s < 4$ **do**
  $c_x \leftarrow D(x_i, \gamma(s))$ ▷ Low-light condition
  $c_{y_1} \leftarrow U(y_i^{k-1}, \gamma(s)))$ ▷ Well-lit condition
  $c_{y_2} \leftarrow U(y_i^0, \gamma(s)))$ ▷ First estimate condition
  $x^s_{\text{patches}} \leftarrow R([c_x, c_{y_1}, c_{y_2}])$ ▷ Patchify
  $y^s_{\text{patches}} \leftarrow \Psi_\theta([x^s_{\text{patches}}, n])$ ▷ Denoise with ILVR
  $y^s_i \leftarrow H(y^s_{\text{patches}})$ ▷ Stitch patches
end for

3.2. Data Normalization

It is useful to standardize data by centering and dividing with the sample standard deviation, so that each datum is

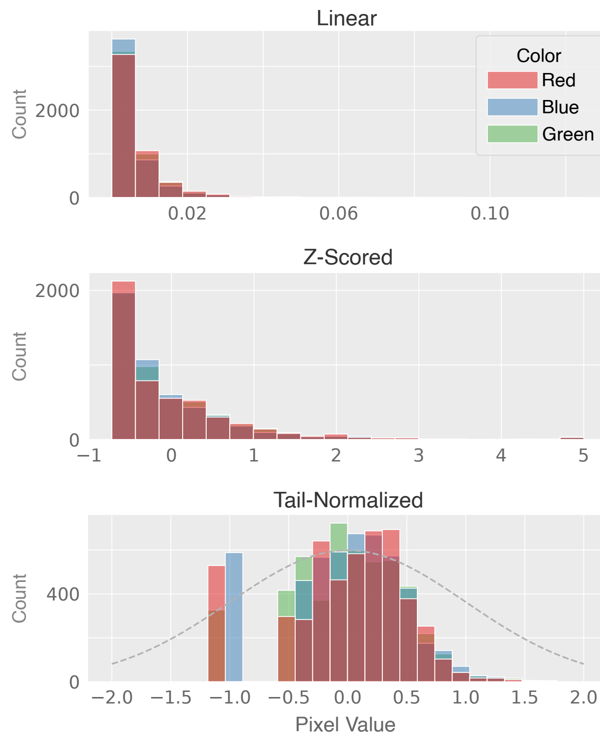

Figure 3. **Color pixel distributions for low-light data.** Right-tailed data do not follow assumptions made in training a diffusion model. We normalize the data to the appropriate range for training. **Top:** Data distribution of a random selection of 30 images from the LOL training set [83]. **Middle:** Data distribution of the same images after Z-scoring. The distribution is still right-tailed. **Bottom:** Data distribution of the same images from using tail-normalization (Sec. 3.2). The dotted line shows a true Gaussian distribution with $\mu = 0$ and $\sigma = 0.5$.

represented in common units. For example, DDPM [28] scales images to be within the range $[-1, 1]$. Given the right-tailed nature of low-light data (Fig. 3), we cannot follow typical Z-scoring [97]. Diffusion models require picking a noise schedule at training, particularly a σ_{min} and σ_{max}. The former is chosen such at that the lowest noise level is indistinguishable from images, and the latter is chosen such that the highest noise level is indistinguishable from white Gaussian noise. Since we are using σ values designed for images, we need our images to be roughly in the same range to satisfy these conditions. Thus, we normalize the data such that the distribution is between $[-1, 1]$ and follows a roughly Gaussian distribution with $\mu = 0$ and $\sigma = 0.5$. For right-tailed data, we found that taking the fourth root of the data, Z-scoring, and then dividing by two gave us a suitable distribution. This normalization is critical as observed in our ablation studies (Sec. 4.4).

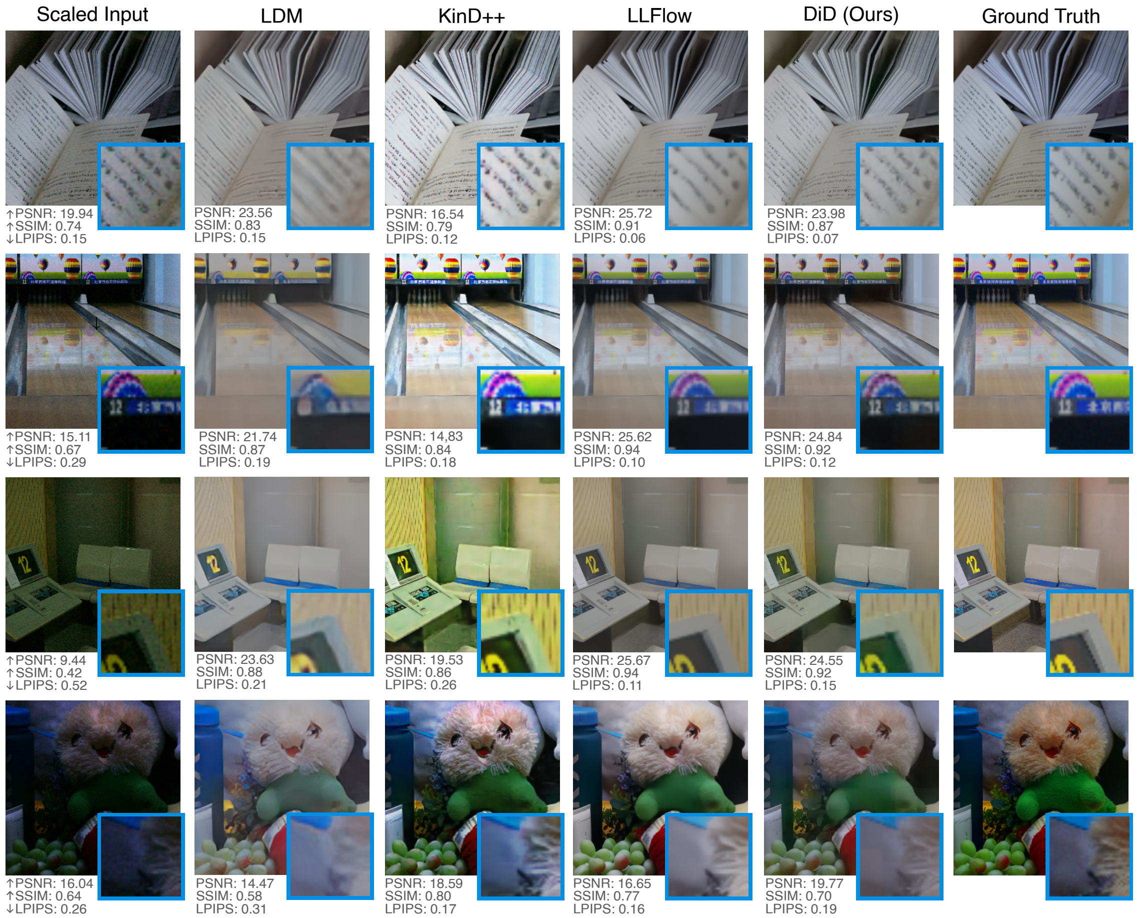

Figure 4. **Qualitative results on the LOL test dataset.** We show reconstructions from LDM [63] and the two best-performing low-light baselines, KinD++ [94] and LLFlow [78]. DiD reconstruction is on par with other low-light reconstruction methods while recovering more fine details such as handwriting and text on signs, such as the "12", with non-saturated appearances and reasonable exposure levels.

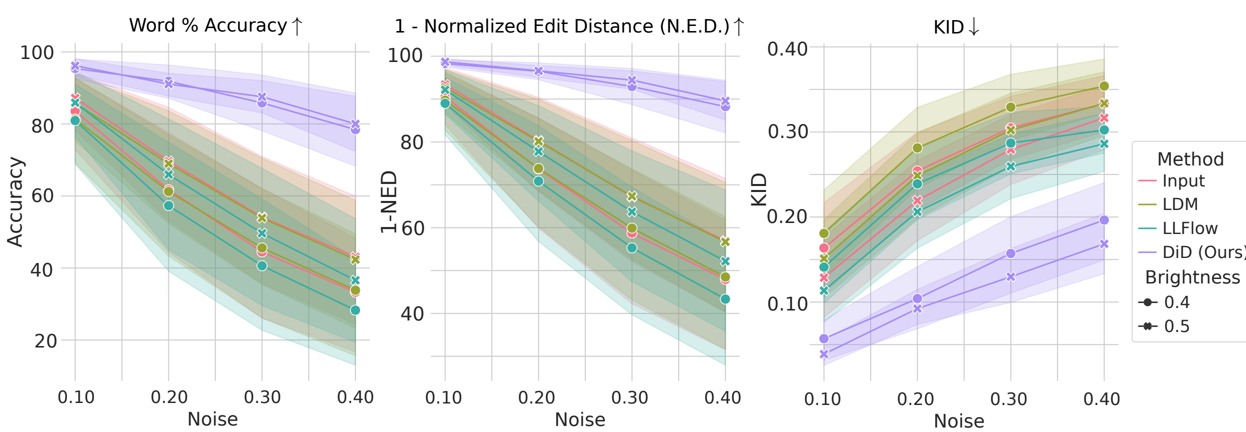

Figure 5. **Text recognition accuracy and reconstruction quality. Left and Center:** We plot word accuracy and 1-NED values at combinations of brightness and Poisson-Gaussian noise levels. Plotted points are the mean values of *all* tested STR datasets. As brightness decreases and noise increases, other low-light reconstruction methods fail to recover enough detail to perform accurate text recognition, while DiD performs consistently well. **Right:** We display KID values for LLFlow and DiD, affirming that DiD provides reconstructions closer to the distribution of ground truth images.

Table 1. **Quantitative comparison of low-light enhancement methods on the LOL test dataset [83]**. Diffusion models are separated from low-light–specific models. We report the average inference time of a 256×256 image on an NVIDIA RTX 3090. ♢ indicates methods that did not closely match their reported performance. We highlight the best and second best results using **bold** and <u>underline</u>, respectively.

Method	PSNR↑	SSIM↑	LPIPS↓	Time (s)	# Params
Zero-DCE [22]	14.67	0.68	0.35	0.06	79.4 K
LIME [23] ♢	14.22	0.65	0.37	1.10	N/A
EnlightenGAN [33]	16.84	0.76	0.33	0.08	8.6 M
RetinexNet [83]	16.77	0.59	0.47	0.22	444.6 K
RUAS [47] ♢	16.41	0.65	0.27	0.11	3.4 K
KinD [95]	20.39	0.88	<u>0.16</u>	0.33	8.0 M
KinD++ [94]	21.73	0.89	<u>0.16</u>	0.66	8.2 M
LLFlow [78]	**24.94**	**0.91**	**0.12**	0.26	38.9 M
DDRM [37]	16.41	0.65	0.21	9.03	552.8 M
LDM [63]	21.41	0.75	0.23	9.81	404.1 M
DiD (Ours)	<u>23.97</u>	<u>0.84</u>	**0.12**	6.64	55.7 M

4. Experiments

4.1. Implementation

We implement our framework with PyTorch on an NVIDIA Quadro RTX 8000. We use the ADAM optimizer [38] with a learning rate of 8×10^{-4} and betas $[0.9, 0.999]$. Our patch-based scheme reduces training time and computational demand, using only 1 GPU to train within 3 days. We train batch sizes of 160 for 3000 iterations. Given that we are working with small datasets, we found this number of iterations to be suitable for convergence. We augment the data by adding either random Gaussian blur or sharpening, as well as scaling brightness and saturation.

4.2. Dataset and Evaluation

We train on the LOw-Light dataset (LOL) [83], which contains 485 training and 15 test low-light/well-lit pairs. There are limited low-light text datasets [46, 87], but these unfortunately do not come with well-lit counterparts. We were interested in estimating not only accurate text, but also viable reconstructions from our method. Thus, we opt to use a benchmark low-light dataset for our training and evaluation.

We report PSNR, SSIM, and LPIPS [92]. Due to the ill-posedness of low-light reconstruction, there are many optimal solutions that do not share the same white balance and exposure level as its ground truth, meaning we do not surpass SOTA on PSNR/SSIM, but do match LPIPS performance. We are unable to provide KID [8] or FID [27] scores on LOL given the limited test set. However, for our text recognition task, we report KID for text reconstructions on much larger scene text datasets (Sec. 4.5). We demonstrate that our method exceeds other low-light reconstruction methods in low-light text recognition in Word Accuracy and Normalized Edit Distance, the Levenshtein distance between words [12, 69, 93]. We also demonstrate that our qualitative results are on par with SOTA performance for low-light reconstruction.

4.3. Baselines

To test our overall reconstruction quality, we compare against SOTA methods in low-light image enhancement, reporting results on the LOL test dataset. We include results from two other diffusion-based models: Denoising Diffusion Restoration Models (DDRM) [37] and LDMs [63]. DDRM applies a pretrained denoising diffusion model to an inverse problem. We pretrain the DDRM to denoise noisy ImageNet [15] images, and use the network to denoise a brightened version of the low-light images. LDMs use a pretrained VAE to encode images from pixel space to a latent space and trains a diffusion model in latent space.

For the non-deterministic models, we generate ten reconstructions for each image and pick the image that gives us highest PSNR. Although DiD does not have the best results numerically (Tab. 1), DiD performance is on par with that of SOTA, especially in LPIPS. DiD also performs the best of the diffusion-based models with significantly less trainable parameters and inference time. Despite not beating PSNR/SSIM, our generative model reconstructs high-frequency details better than SOTA low-light methods can (Fig. 4).

4.4. Ablations

To understand the contribution of each model component, we conduct several ablations (Tab. 2). We refer the number of models trained and the number of scales used in each model as the **model-to-scale ratio**. A 1:4 ratio means we train 1 model with 4 different resolutions, and a 2:1 ratio means we train 2 models, each corresponding to its own unique resolution. DiD, which uses a 1:4 ratio, outperforms all other ablations using a 1:4 models-to-scale ratios. We find that reducing the number of models or scales per model does not provide the level of conditioning information necessary for refined predictions. We also show that ILVR is critical for conditioning a prediction to be within a restricted exposure range. See the supplement for details on model-to-scale ratios and more ablations.

4.5. Low-light Text Recognition

We demonstrate our method's utility in low-light STR. We evaluate on real scene text datasets: IIIT5k-Words (IIIT5k) [55], ICDAR 2013 (IC13-1015) [35], Street View Text (SVT) [76], and SVT-Perspective (SVTP) [60]. We simulate capturing these images in low light by dimming its brightness and adding Poisson-Gaussian noise. More information on the noise model can be found in the supplement.

Table 2. **Ablation studies.** **Models/scales** refers to the number of models and scales for each model. **Noise** refers to the addition of noise on the conditioning image. **LPIPS** refers to an additional LPIPS loss. **Data** refers to data normalization. **Cond.** refers to adding c_{y_2} to the conditioning input. We highlight the best and second best results using **bold** and <u>underline</u>, respectively.

Models/scales	Noise	LPIPS	Data	Cond.	PSNR↑	SSIM↑	LPIPS↓
1 : 4	✗	✗	✗	✗	16.26	0.57	0.48
1 : 4	✗	✓	✓	✗	19.56	0.74	0.35
1 : 4	✓	✓	✗	✗	16.94	0.63	0.46
1 : 4	✓	✓	✓	✗	17.62	0.74	<u>0.31</u>
4 : 1	✓	✗	✓	✗	<u>19.63</u>	<u>0.80</u>	**0.14**
1 : 2	✓	✓	✓	✗	17.49	0.72	0.33
1 : 2	✓	✓	✓	✓	18.37	0.73	0.33
2 : 1	✓	✓	✓	✓	19.35	0.72	<u>0.31</u>
DiD (no ILVR)	✓	✓	✓	✓	17.78	0.72	0.36
DiD	✓	✓	✓	✓	**21.00**	**0.82**	**0.14**

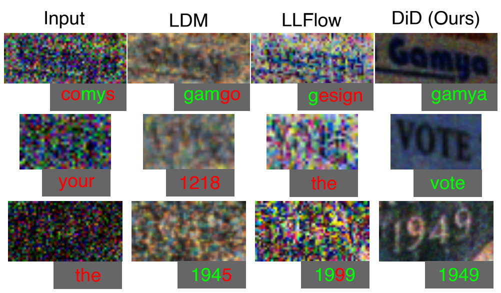

Figure 6. **Comparing text recognition predictions.** We show samples from real scene text datasets and the reconstructions from LDM [63], LLFlow [78], and DiD. The input has been scaled down by a factor of 0.4 with Poisson-Gaussian noise.

We found that more complex noise simulation is only relevant in the case of extremely low light [56, 84, 85]. The noise level is more challenging than that of LOL (compare inputs from Figure 4 and Figure 1). We use a SOTA STR method, PARSeq [7], to evaluate each low-light methods pretrained on LOL, on recovering text.

We report Word Accuracy (% Acc) and Normalized Edit Distance (1 - NED) (Fig. 5). DiD consistently performs well under extremely dark and noisy conditions, reporting $> 75\%$ accuracy even in the most dark and noisy setting, while other methods begin to fail as conditions worsen. We also present KID scores for LLFlow [78] and DiD in the supplement, which compares the distribution of the reconstructions to the distribution of the clean, well-lit ground truth STR images. We reconstruct images that not only provide successful text recognition in dark conditions, but also better follow the distribution of ground truth images than other methods do (Fig. 6).

4.6. Results on Other Datasets

We tested our method on Seeing in the Dark (SID) [10] Sony dataset, which consists of low-light RAW data with real noise. We trained our model on this RAW data, and found that its performance on the test set (PSNR: 23.98dB/SSIM: 0.78) was comparable to the original SID method (PSNR: 28.61dB/SSIM: 0.77). Our reconstruction allows for clear reconstruction of text in low light (Fig. 7). See our supplement for more dataset results.

Figure 7. **DiD reconstructs unseen, real low-light data.** Results from running PARSeq [7] on SID Sony low-light data [10]. Green signifies correct text recognition. Low-light inputs are scaled for visualization.

5. Discussion and Conclusions

The future of automation depends on robust, high-level algorithms performing on images from a wide range of conditions. We propose a low-light reconstruction method using DDPMs which, without any specific task-level design, outperforms SOTA low-light reconstruction methods on low-light text recognition.

Limitations and Future Work. Sampling from diffusion models requires multiple steps and may take prohibitively long in real-time scenarios. Fortunately, there is a fast growing family of methods to improve sampling time [40, 52, 67, 82] which can be applied.

We use an autoregressive patch-based method which requires multiple passes through the network, and if image resolutions are not a multiple of the patch size, we have to interpolate the image to a larger resolution. Future work would expand DiD for inference at all resolutions without interpolation, and LDMs [63] and similar models which diffuse in latent space [3, 41] are a promising direction.

Conclusion. As more tasks become automated, it is increasingly critical that reconstruction methods perform well in corner cases, such as dark and noisy conditions. Our method provides sharp results, which align well with perceptual quality and especially downstream tasks operating on low-light images.

Acknowledgements This project was in part supported by a PECASE from the ARO, by Samsung, and by the Stanford Institute for Human-Centered AI (HAI). We thank Yufei Wang for his help in running LLFlow. We thank Connor Lin and Nitish Padmanaban for their insights and comments on this work.

References

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(2):593–600, 2007. 2

[2] Tomer Amit, Eliya Nachmani, Tal Shaharbany, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 3

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 8

[4] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3113–3122, 2021. 3

[5] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[6] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 3

[7] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 178–196. Springer, 2022. 3, 8

[8] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 3, 7

[9] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 2

[10] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3291–3300, 2018. 2, 8

[11] Manri Cheon, Jun-Hyuk Kim, Jun-Ho Choi, and Jong-Seok Lee. Generative adversarial network-based image super-resolution using perceptual content losses. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[12] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 7

[13] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 5

[14] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022. 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 7

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[17] Steven Diamond, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. Dirty pixels: Towards end-to-end image processing and perception. *ACM Transactions on Graphics (TOG)*, 40(3):1–15, 2021. 3

[18] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129:1258–1281, 2021. 3

[19] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. *arXiv preprint arXiv:2205.00159*, 2022. 3

[20] Shancheng Fang, Zhendong Mao, Hongtao Xie, Yuxin Wang, Chenggang Yan, and Yongdong Zhang. Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[21] Xueyang Fu, Yinghao Liao, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A probabilistic method for image enhancement with simultaneous illumination and reflectance estimation. *IEEE Transactions on Image Processing*, 24(12):4965–4977, 2015. 2

[22] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020. 2, 7

[23] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2):982–993, 2016. 2, 7

[24] Prateek Gupta, Priyanka Srivastava, Satyam Bhardwaj, and Vikrant Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *2011 International Conference on Communication and Industrial Application*, pages 1–4. IEEE, 2011. 3

[25] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016. 2

[26] Yue He, Chen Chen, Jing Zhang, Juhua Liu, Fengxiang He, Chaoyue Wang, and Bo Du. Visual semantics allow for textual reasoning better in scene text recognition. In *Proceed-*

ings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 888–896, 2022. 3

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 3, 7

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 5

[29] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 3

[30] Po-Hao Hsu, Che-Tsung Lin, Chun Chet Ng, Jie Long Kew, Mei Yih Tan, Shang-Hong Lai, Chee Seng Chan, and Christopher Zach. Extremely low-light image enhancement with scene text restoration. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 317–323. IEEE, 2022. 3

[31] Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015. 3

[32] Haidi Ibrahim and Nicholas Sia Pik Kong. Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(4):1752–1758, 2007. 2

[33] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021. 2, 7

[34] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 3

[35] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1484–1493. IEEE, 2013. 7

[36] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 3, 4

[37] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 3, 7

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[39] Tzofi Klinghoffer, Siddharth Somasundaram, Kushagra Tiwary, and Ramesh Raskar. Physics vs. learned priors: Rethinking camera and algorithm design for task-specific imaging. In *2022 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2022. 3

[40] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 8

[41] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 8

[42] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 3

[43] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018. 2

[44] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8714–8721, 2019. 3

[45] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *ACM Transactions on Graphics (TOG)*, 38(6):164–1, 2019. 2

[46] Hang Liu, Mengke Yuan, Tong Wang, Peiran Ren, and Dong-Ming Yan. List: low illumination scene text detector with automatic feature enhancement. *The Visual Computer*, 38(9-10):3231–3242, 2022. 3, 7

[47] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021. 2, 7

[48] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics (TOG)*, 33(6):1–9, 2014. 2

[49] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129:161–184, 2021. 3

[50] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 2

[51] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[52] Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022. 8

[53] Paras Maharjan, Li Li, Zhu Li, Ning Xu, Chongyang Ma, and Yue Li. Improving extreme low-light image denoising via

residual learning. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 916–921. IEEE, 2019. 2

[54] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 2

[55] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *British Machine Vision Conference (BMVC)*. BMVA, 2012. 7

[56] Kristina Monakhova, Stephan R Richter, Laura Waller, and Vladlen Koltun. Dancing under the stars: video denoising in starlight. In *Proc. CVPR IEEE*, pages 16241–16251, 2022. 8

[57] Nguyen Nguyen, Thu Nguyen, Vinh Tran, Minh-Triet Tran, Thanh Duc Ngo, Thien Huu Nguyen, and Minh Hoai. Dictionary-guided scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7383–7392, 2021. 3

[58] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[59] Seonhee Park, Soohwan Yu, Byeongho Moon, Seungyong Ko, and Joonki Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Transactions on Consumer Electronics*, 63(2):178–184, 2017. 2

[60] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 569–576, 2013. 7

[61] Charan D Prakash and Lina J Karam. It gan do better: Gan-based detection of objects on images with varying quality. *IEEE Transactions on Image Processing*, 30:9220–9230, 2021. 3

[62] Zhi Qiao, Yu Zhou, Dongbao Yang, Yucan Zhou, and Weiping Wang. Seed: Semantics enhanced encoder-decoder framework for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13528–13537, 2020. 3

[63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 6, 7, 8

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3

[65] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[66] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[67] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 8

[68] Mahmood Sharif, Lujo Bauer, and Michael K Reiter. On the suitability of lp-norms for creating and preventing adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1605–1613, 2018. 3

[69] Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1429–1434. IEEE, 2017. 7

[70] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[72] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 3

[73] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3

[74] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015. 3

[75] Junyi Wang, Weimin Tan, Xuejing Niu, and Bo Yan. Rdgan: Retinex decomposition based adversarial learning for low-light enhancement. In *2019 IEEE international conference on multimedia and expo (ICME)*, pages 1186–1191. IEEE, 2019. 2

[76] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1457–1464. IEEE, 2011. 7

[77] Wen Wang, Jing Zhang, Wei Zhai, Yang Cao, and Dacheng Tao. Robust object detection via adversarial novel style exploration. *IEEE Transactions on Image Processing*, 31:1949–1962, 2022. 3

[78] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2604–2612, 2022. 2, 6, 7, 8

[79] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 3

[80] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3

[81] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 3

[82] Daniel Watson, Jonathan Ho, Mohammad Norouzi, and William Chan. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802*, 2021. 8

[83] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 2, 5, 7

[84] Kaixuan Wei, Ying Fu, Jiaolong Yang, and Hua Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2758–2767, 2020. 2, 8

[85] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *IEEE TPAMI*, 44(11):8520–8537, 2021. 8

[86] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. *arXiv preprint arXiv:2203.04306*, 2022. 3

[87] Minglong Xue, Palaiahnakote Shivakumara, Chao Zhang, Yao Xiao, Tong Lu, Umapada Pal, Daniel Lopresti, and Zhibo Yang. Arbitrarily-oriented text detection in low light natural scene images. *IEEE Transactions on Multimedia*, 23:2706–2720, 2020. 3, 7

[88] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4049, 2014. 3

[89] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12113–12122, 2020. 3

[90] Yu Yuan, Jiaqi Wu, Lindong Wang, Zhongliang Jing, Henry Leung, Shuyuan Zhu, and Han Pan. Learning to kindle the starlight. *arXiv preprint arXiv:2211.09206*, 2022. 2, 3

[91] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 3

[92] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 3, 5, 7

[93] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 7

[94] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129(4):1013–1037, 2021. 2, 6, 7

[95] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1632–1640, 2019. 2, 7

[96] Shangchen Zhou, Chongyi Li, and Chen Change Loy. Lednet: Joint low-light enhancement and deblurring in the dark. In *European Conference on Computer Vision*, pages 573–589. Springer, 2022. 2

[97] Dennis G Zill. *Advanced engineering mathematics*. Jones & Bartlett Learning, 2020. 5