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

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

1. Teaser Figure

Images used in the teaser figure are from the SVTP dataset at brightness level 0.4 and random Poisson-Gaussian noise level 0.25.

2. Architecture

2.1. Network design

One can get reasonable recovery of low-light signal by scaling up low-light images. However, in scaling up low-light images, the noise is also amplified, and this is apparent in the scaled inputs visualized in Figure 4 of the main paper. We wanted to reconstruct fine details without noise amplification from a single image, so we opt to use a generative model.

We tested ADM [5] as an alternative to DDPM and found instabilities in training. We also tested a training patch resolution of 64×64 and found that it worked comparably, with slightly longer training times. We choose not to train a GAN such as CycleGAN [29] or Pix2Pix [6] because there is no large paired dataset of low-light/well-lit pairs, which would make training a GAN especially unstable.

2.2. Network conditioning

We use a U-Net [16] as the base of our DDPM. We use the U-Net as implemented in [19], which consists of 3 "down-blocks" and 3 "up-blocks" with skip connections between them. The network uses positional embeddings, 4 residual blocks per resolution, and per-resolution multipliers of [2,2,2]. The network has a base 128 number of channels, and a dropout factor of 0.10. We use the weighting of the L2 loss as prescribed by Karras et al. [8] and apply a fixed scalar weight to the perceptual component (LPIPS [26]) of our custom loss function.

Figure 1. **Inference without exposure and white balancing constraints.** Reconstructions show patch-to-patch inconsistencies in exposure levels and white balancing if we perform inference on individual patches and stitch them together or we do not perform additional ILVR conditioning in DiD.

We show examples of patch-to-patch inconsistencies observed without proper conditioning in Figure 1. Using a multi-scale approach with ILVR [3], we can mitigate these issues to reconstruct a coherent image. Using ILVR [3] at every denoising step led to blurring, but applying ILVR to 6 of the 18 steps was sufficient. For our loss, we empirically found that $\lambda = 5$ worked well. Before training, we apply EDM [8] preconditioning.

3. Training and Inference

3.1. Low-light datasets

It is challenging to find large real low-light training datasets. Multiple works have demonstrated accurate noise modeling for low-light [1, 24], but it remains difficult to model the loss of scene content and color in dim lighting. We opt to use the LOL dataset because it remains one of the most popular choices for low-light training [28], allowing for easier comparison against SOTA.

3.2. Data preprocessing

For tail-normalization, the exact root number and division number may vary from dataset to dataset. We find that

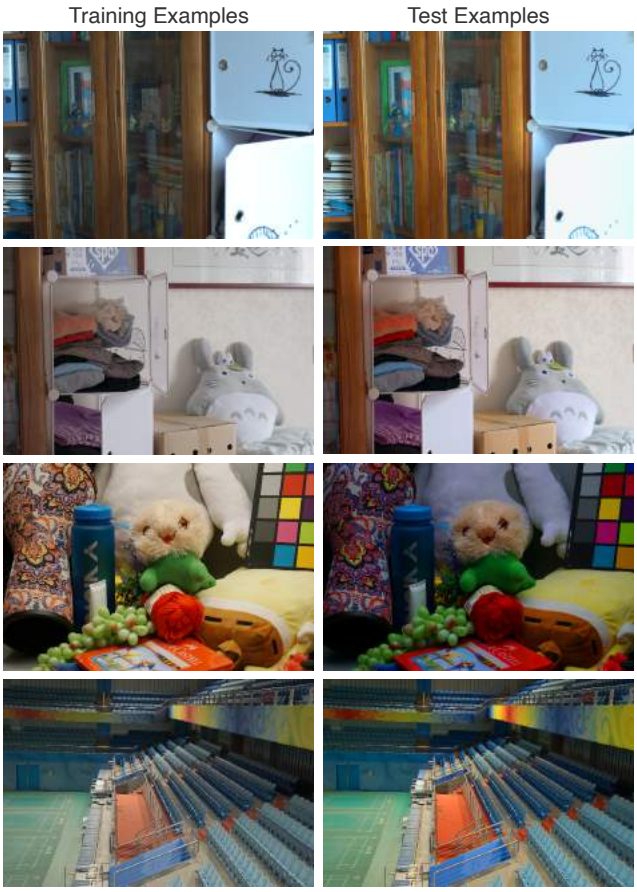

Figure 2. **LOL dataset inconsistencies.** There is overlap in scenes between the LOL training and test datasets, and the well-lit test images are significantly more contrasted than the well-lit training set images.

our choice of fourth root and dividing by two after z-scoring was suitable for the LOL [22], Seeing in the Dark [2], and a modified Seeing in the Dark [25] dataset.

Among low-light datasets, LOL is the most popular to train and test on after custom datasets as found in a survey of low-light reconstruction methods [28]. However, there is significant overlap in scenes from the train and test set, and for unknown reasons, the test set ground truths have their color contrast raised, as seen in Figure 2. This contrast raise makes it challenging to get an accurate sense of performance. We report quantitative results on the LOL test set for comparison sake, but believe that our image quality is reflective of realistic coloring as shown in the training set.

For LOL, we perform preprocessing before training. First, we center crop the image to be 256×256. We then convert the image from sRGB space to linear space. We perform data normalization using a mean and standard deviation found in linear space on a random sample of 30 images. After tail-normalizing our data, we then train with images in the range $[-1, 1]$. Upon inference, we unnormalize the data and convert the image from linear space back

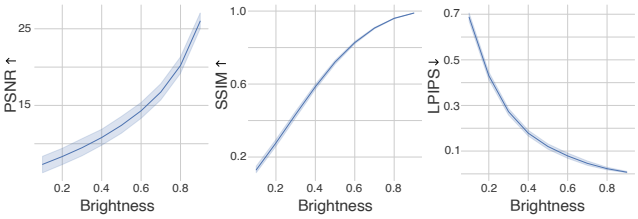

Figure 3. **Metrics are sensitive to exposure levels.** There is a significant drop in performance once exposure levels begin to change, even though content is the same. The brightness represented here is the scaled V value in the HSV-converted image.

to sRGB space for visualization. We compute all metrics in sRGB space. We note that metrics are sensitive to different exposure levels despite having the same content. We show this sensitivity by scaling down the brightness of the LOL test dataset and computing PSNR, SSIM, and LPIPS across different brightness levels (Fig. 3).

3.3. Inference details

For inference, we apply 18 sampling steps. For $s = 0$, we do not apply ILVR. For $s = 1, 2, 3$, we apply ILVR to the first 6 steps, using the low-frequency content from the previous scale's prediction. To filter low-frequency content, we downsample and upsample respectively, using bilinear interpolation with anti-aliasing. We also tested numerous different filters (Lanczos, cubic, and nearest) and found no significant difference in performance. We tested using more inference steps ($N = 100$) and found only minor changes in performance (PSNR: $+0.120$, SSIM: -0.002, LPIPS: -0.011).

4. Experiments

4.1. Baseline methods

LLFlow [21] is non-deterministic in theory and deterministic in practice. The method uses a fixed latent feature, which leads to a deterministic result. By changing the latent feature, one can get different results due to the one-to-one mapping of normalizing flows. However, because the results from different latents are not perceptually obvious, we maintain a fixed latent feature. This choice may differ in cases where there is more training data. For DDRM [9], we scale up the brightness of LOL images by a factor of 6 to produce a brighter image (with amplified noise), and denoise the image using DDRM pretrained on ImageNet. We train an LDM [15] from scratch using LOL and use 200 steps as prescribed.

We show more qualitative results from the LOL test dataset in Figure 4. DiD requires longer inference times than LLFlow on average due to the number of inference steps in the reverse diffusion process. The same can be said

about LDM. As faster sampling methods are being developed, as mentioned in our main paper, we believe the inference time for diffusion models can only be improved while maintaining better quality reconstructions than those from LLFlow.

4.2. Ablation studies

We clarify the term **model-to-scales** ratio. A 4:1 model-to-scales ratio means that we trained 4 models. Each model is trained on 1 scale. An example of the 4 models using a 4:1 to ratio is as follows:

- Model A is trained on 32×32 images that are downsampled versions of the 256×256 low-light measurement.
- Model B is trained on 32×32 patches that are taken from a 64×64 image (which is a downsampled version of the 256×256 low-light measurement).
- Model C is trained on 32×32 patches that are taken from a 128×128 image (which is a downsampled version of the 256×256 low-light measurement).
- Model D is trained on 32×32 patches that are taken from the 256×256 low-light measurement.

A 2:1 ratio means we trained 2 models. Each model is trained on 1 scale. An example of the 2 models using a 2:1 to ratio is as follows:

- Model A is trained on 32×32 images that are downsampled versions of the 256×256 low-light measurement.
- Model B is trained on 32×32 patches that are taken from the 256×256 low-light measurement.

A 1:2 ratio means we trained 1 model with 2 scales. The model is trained on 32×32 patches that either are entire images from downsampling the low-light measurement from 256×256 to 32×32 or are 32×32 patches extracted from the original 256×256 low-light measurement.

We provide additional ablations highlighted in Table 1 and show qualitative results for top-performing ablations in Figure 5. All ablations use ILVR [3] during inference unless specified otherwise. For ablations, we report metrics computed on a randomly selected reconstruction rather than the best of 10 reconstructions. We include an ablation study in which we attempt to refine the predictions in pixel space with a lightweight CNN. This refinement network has 3 Conv2D+LeakyRELU layers with the following channel sizes $[3, 128, 3]$. We use an L2 loss and Adam optimizer for 10,000 iterations. This CNN operates as a deterministic network to improve predictions. We train the CNN on 256×256 predictions from a pretrained DiD and compare the reconstruction to 256×256 ground truth images. However, we find that because the LOL test dataset has a significant distribution shift from its training dataset, there is an upper limit to how much the CNN can improve results. We observe comparable SSIM (+0.06), worse PSNR (-0.47), and comparable LPIPS (+0.01). Since the performance here was overall comparable, we instead report our original method DiD as part of our core contribution without any additional trainable parameters.

5. Scene Text Recognition

We simulate low light in scene text recognition by converting the images from RGB to HSV. We then scale the V channel by a factor less than one (in our simulations, we use 0.4 or 0.5), following [11, 27] in simulating images under differing light conditions. We follow the noise model from Mildenhall et al. [12], which model Poisson-Gaussian noise as a Gaussian with zero-mean and signal-dependent variances. We then convert the image back to RGB and add Poisson-Gaussian noise with a specified standard deviation for the Gaussian distribution and signal-dependent variance for the Poisson distribution. We test the following datasets which display a wide range of capture quality:

- **IIIT5k-Words (IIIT5k) [13]** which contains 3000 test images, most of which are of acceptable quality.
- **ICDAR2013 (IC13) [7]** which consists of 1015 images for testing. The ICDAR 2013 and 2015 datasets are similar in text regularity and conditions.
- **Street View Text (SVT) [20]** which consists of 647 images, many of which are severely degraded by blur, noise, and low resolution.
- **SVT-Perspective (SVTP) [14]** which contains 645 images, with most suffering from heavy perspective distortion.

For each dataset, we sample 30 images to find the mean and standard deviation needed for tail-normalization. For the text processing, we additionally scale our recovered image by 3. Since our method recovers an arbitrary exposure level without noise, scaling the image should not amplify any noise. We show the performance of each method on individual datasets (a decomposition of Figure 5 from our main paper) in Figure 6. We also show more qualitative results of different brightness and noise levels in Figure 7.

6. Reconstruction on Other Datasets

Many real low-light datasets are task datasets with no well-lit ground truth (Dark Zurich [17], ACDC [18], Nighttime Driving [4], CODaN [10]), so we cannot provide quantitative results on reconstruction performance.

Table 1. **Results from all ablation studies. Models/scales** refers to the number of trained models and the number of scales for which each model is trained. **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 an upsampled scale 0 prediction to the conditioning input. Highlighted in blue are ablations which were not included in our main paper. We highlight the best and second best results using **bold** and underlined text, respectively.

ID	Models/scales	Noise	LPIPS	Data	Cond.	PSNR↑	SSIM↑	LPIPS↓
A	1 : 4	✗	✗	✗	✗	16.26	0.57	0.48
B	1 : 4	✗	✓	✓	✗	19.56	0.74	0.35
C	1 : 4	✓	✓	✗	✗	16.94	0.63	0.46
D	1 : 4	✓	✓	✓	✗	17.62	0.74	0.31
E	4 : 1	✓	✗	✓	✗	19.63	0.80	**0.14**
F	1 : 2	✓	✓	✓	✗	17.49	0.72	0.33
G	1 : 2	✓	✓	✓	✓	18.37	0.73	0.33
H	2 : 1	✓	✓	✓	✓	19.35	0.72	0.31
I	1 : 4	✓	✗	✓	✗	17.78	0.74	0.31
J	2 : 1	✓	✓	✓	✗	19.32	0.72	0.32
K	DiD (with CNN)	✓	✓	✓	✓	<u>20.53</u>	**0.88**	<u>0.15</u>
L	DiD (no ILVR)	✓	✓	✓	✓	17.78	0.72	0.36
M	DiD	✓	✓	✓	✓	**21.00**	<u>0.82</u>	**0.14**

Of the real low-light task datasets, only DarkFace [23] has been used for qualitative evaluation by 2 of 8 baselines (Zero-DCE and RUAS). We test our LOL-trained model on DarkFace (Fig. 8), and found DiD to be highly robust against unseen, real test data, while LLFlow leaves an unrealistic red tint on images.

Our method could also be applied for other high-level downstream tasks such as segmentation and classification. However, our contributions are primarily in reconstructing high-frequency details, of which are not completely necessary for succeeding at segmentation and classification tasks. We focus on instead on a task that requires high-frequency details, and thus shows the strengths of diffusion models.

References

[1] 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. 1

[2] 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

[3] 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. 1, 3

[4] Dengxin Dai and Luc Van Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 3819–3824. IEEE, 2018. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1

[7] 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. 3

[8] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. 1

[9] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 2

[10] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot domain adaptation with a physics prior. 2021. 3

[11] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7):2175–2193, 2021. 3

[12] 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. 3

[13] Anand Mishra, Karteek Alahari, and CV Jawahar. Scene text recognition using higher order language priors. In *British Machine Vision Conference (BMVC)*. BMVA, 2012. 3

[14] Trung Quy Phan, Palaiahnakote Shivakumara, Shangxuan Tian, and Chew Lim Tan. Recognizing text with perspective distortion in natural scenes. In *Proceedings of the IEEE In-*

ternational Conference on Computer Vision, pages 569–576, 2013. 3

[15] 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. 2, 9

[16] 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. 1

[17] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3139–3153, 2020. 3

[18] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 3

[19] 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. 1

[20] 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. 3

[21] 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, 9

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

[23] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018. 4

[24] 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. 1

[25] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2281–2290, 2020. 2

[26] 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. 1

[27] Yu Zhang, Xiaoguang Di, Bin Zhang, Ruihang Ji, and Chunhui Wang. Better than reference in low-light image enhancement: conditional re-enhancement network. *IEEE Transactions on Image Processing*, 31:759–772, 2021. 3

[28] Shen Zheng, Yiling Ma, Jinqian Pan, Changjie Lu, and Gaurav Gupta. Low-light image and video enhancement: A comprehensive survey and beyond. *arXiv preprint arXiv:2212.10772*, 2022. 1, 2

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1

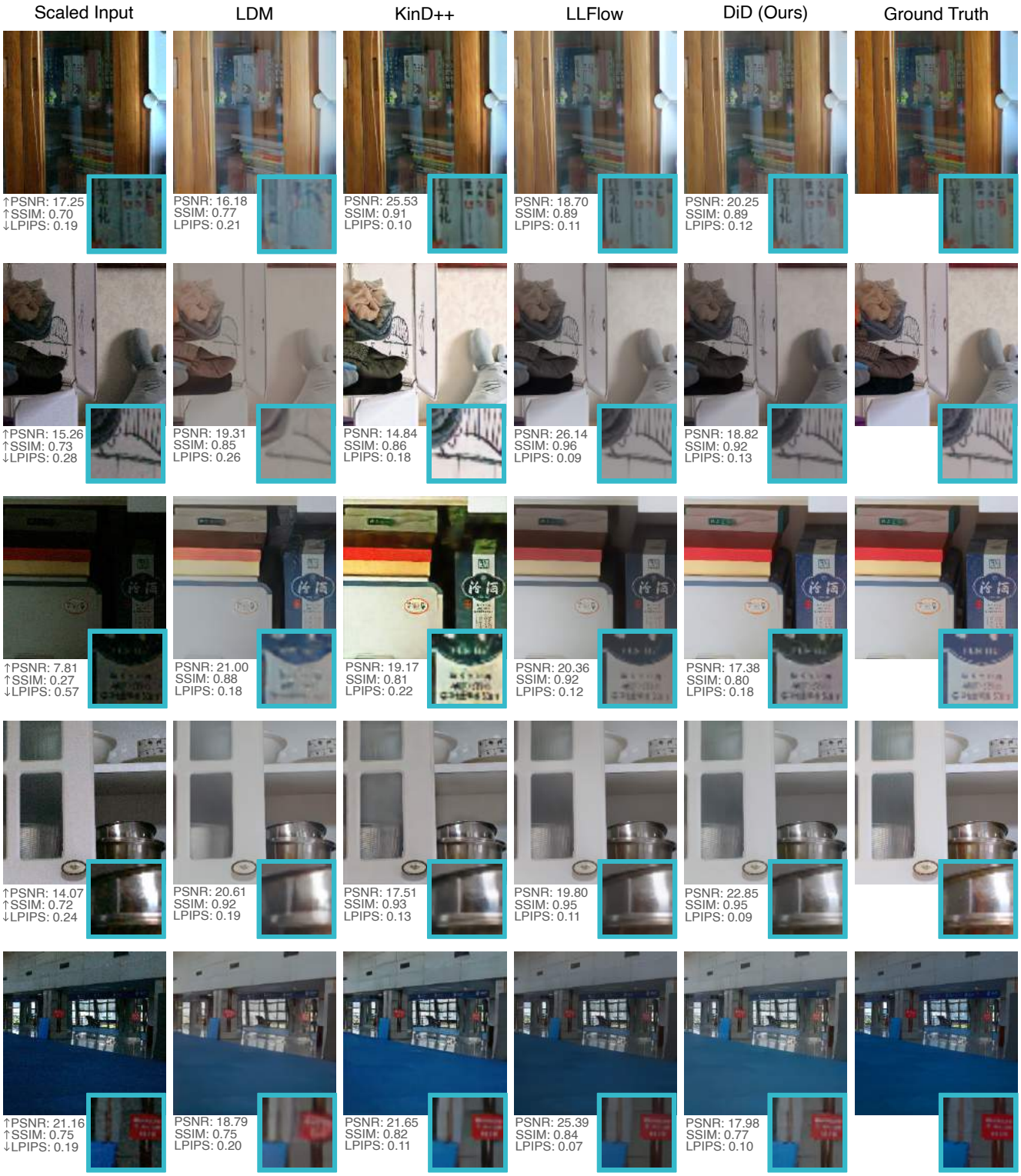

Figure 4. **Qualitative results of baselines from more of the LOL test dataset.** We show results from top-performing low-light baselines. DiD reconstruction is competitive with reconstructions from other methods. We scale the input by a factor of 5 for visualization.

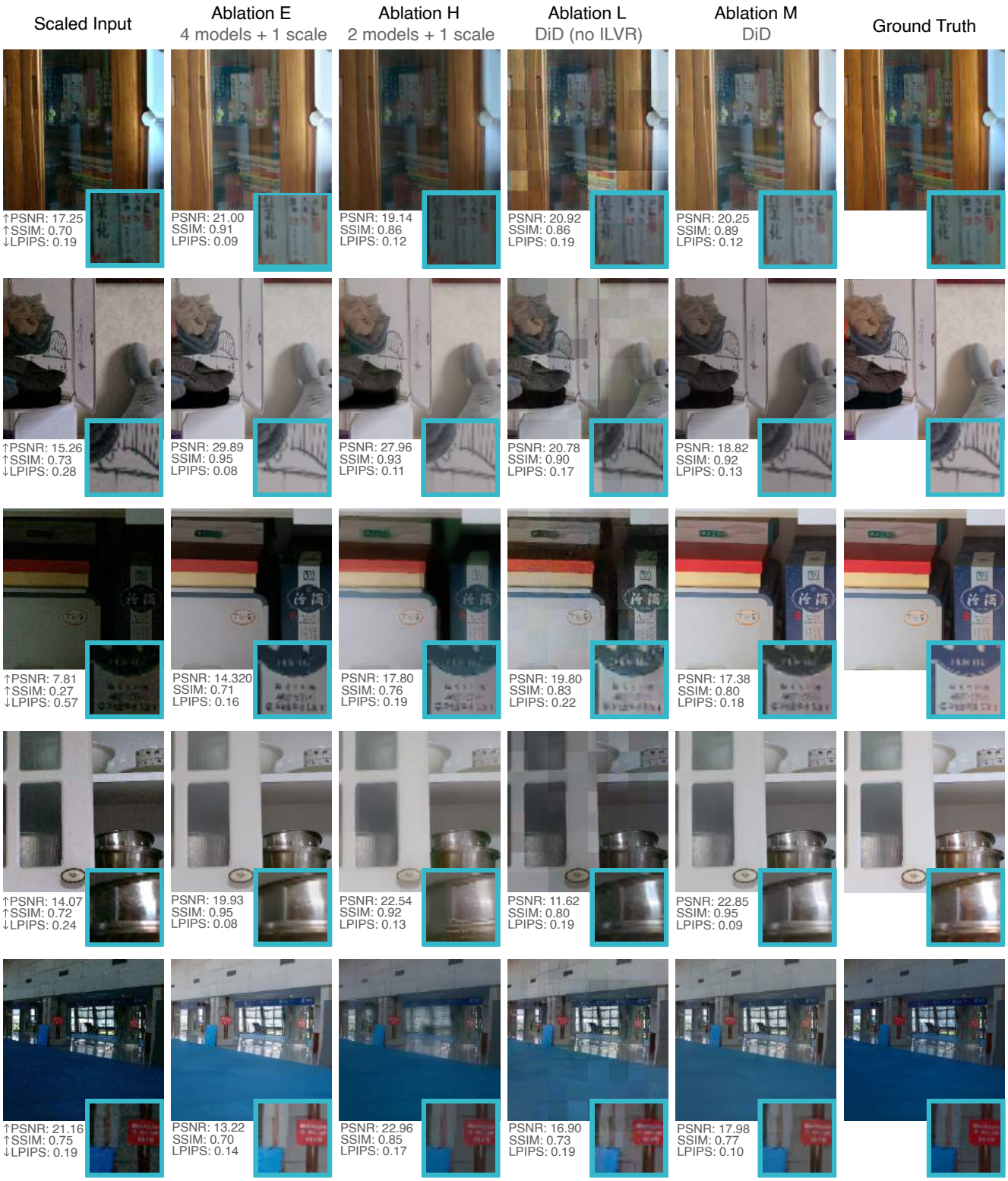

Figure 5. **Qualitative results of ablations of the LOL test dataset.** We show results from top-performing ablations as described Table 1. The combination of all described components, DiD performs the best robustly across images. We scale the input by a factor of 5 for visualization.

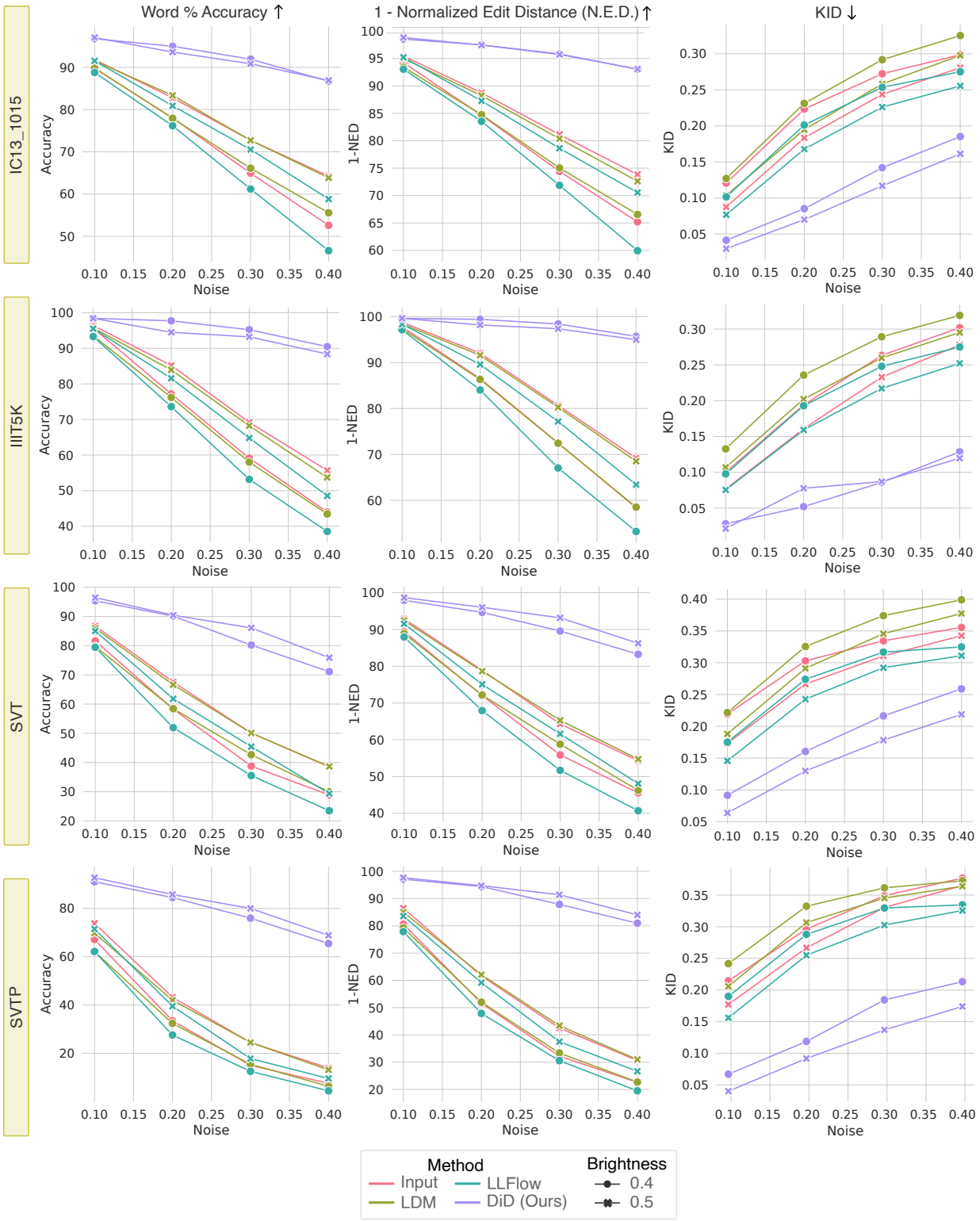

Figure 6. **Quantitative performance on STR datasets.** We show performances of each method on each individual dataset at two levels of brightness and a range of Poisson-Gaussian noise levels using text recognition metrics (Word Accuracy and 1-Normalized Edit Distance) and KID. DiD performs robustly against noisy and dark conditions and exceeds in all these metrics.

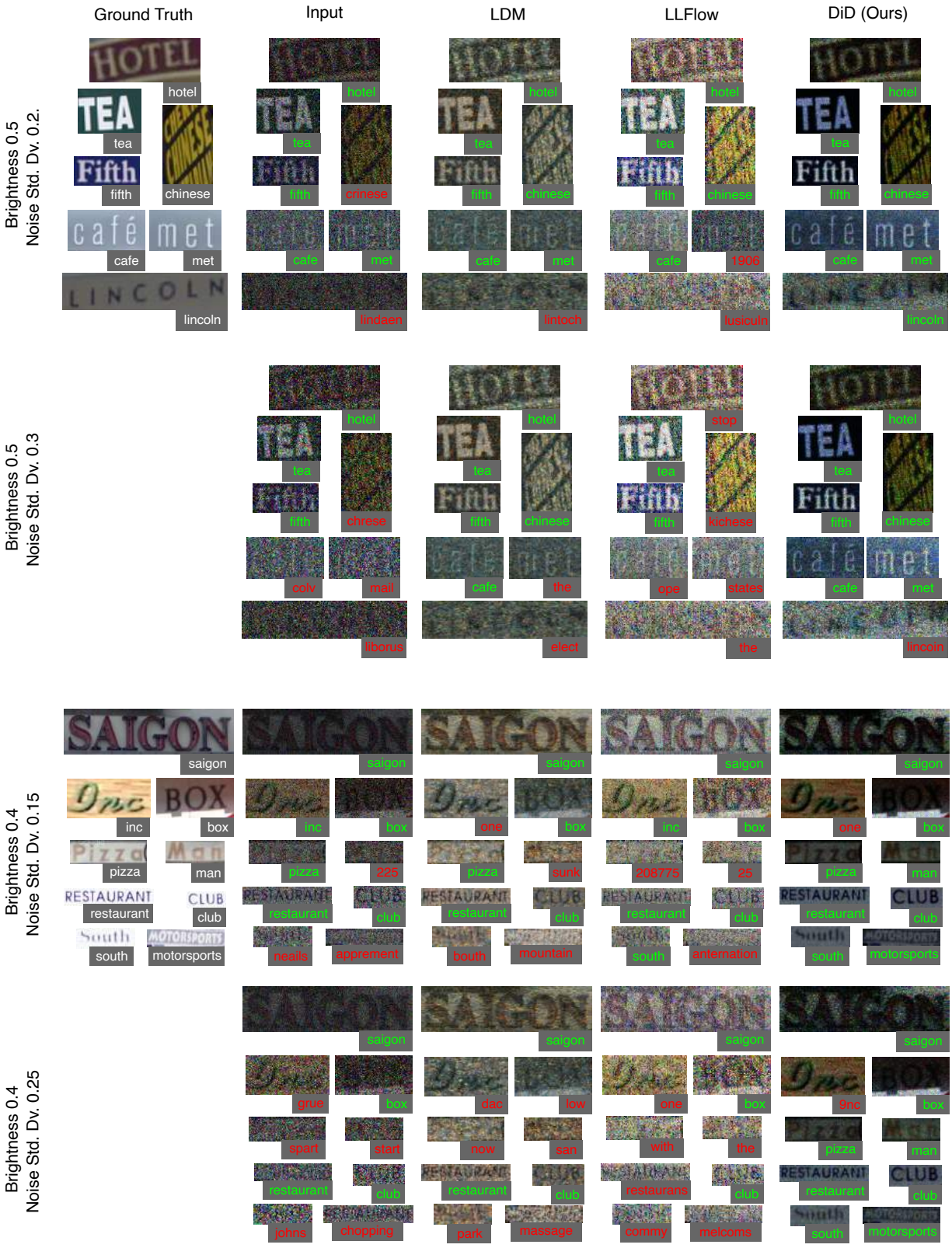

Figure 7. **Qualitative results of STR performance on SVT dataset.** We show results of LDM [15], LLFlow [21], and DiD on different examples in one of the four STR datasets. DiD is able to recover edges and high-frequency detail better in noisy and dark conditions to permit more accurate text recognition predictions than other methods can.

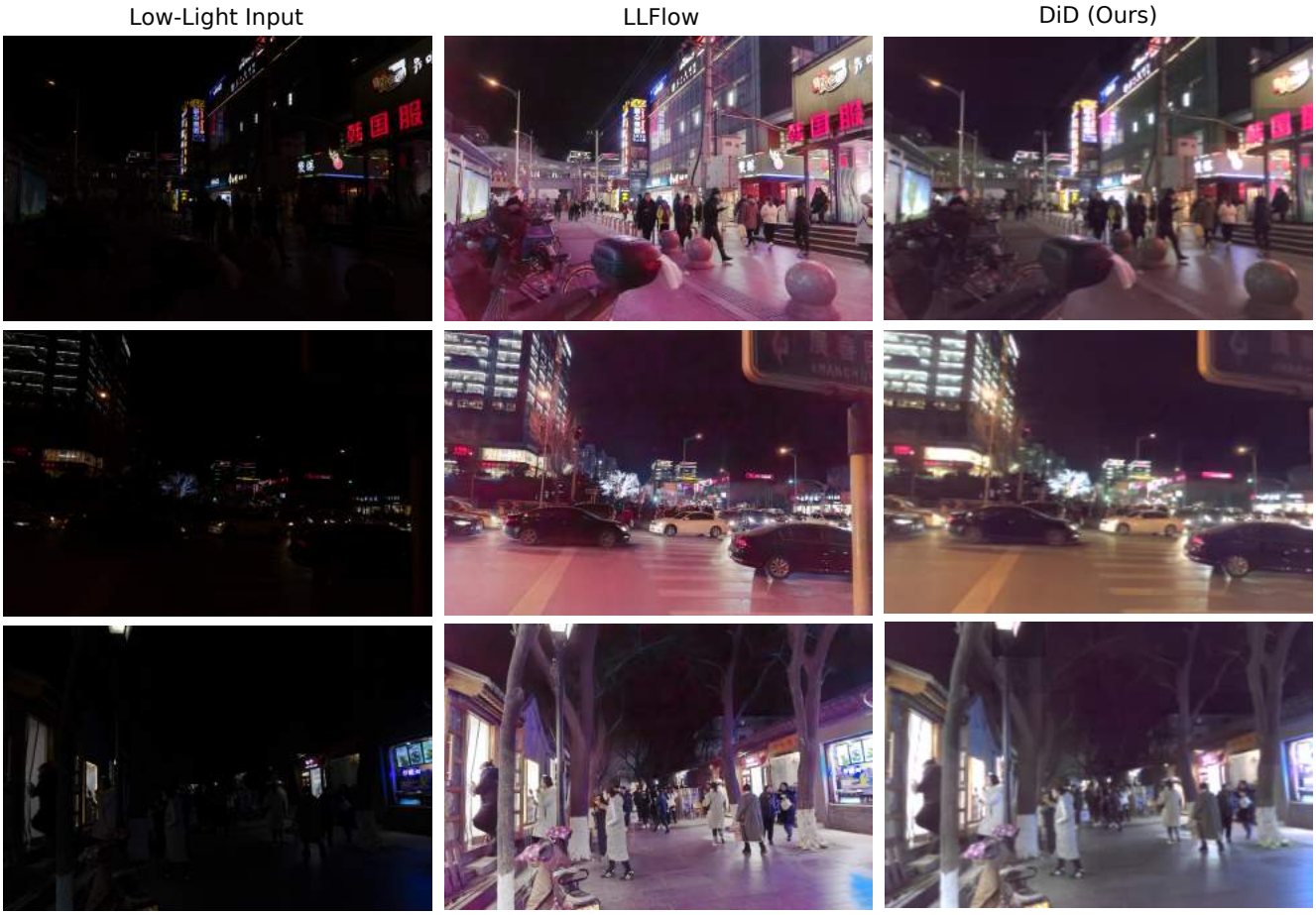

Figure 8. **Reconstruction of DarkFace data, a real low-light task dataset.** DiD provides a realistic reconstruction of real low-light images, while LLFlow provides an unrealistic reddish tint. Both reconstructions could be used for face recognition, but DiD provides more aesthetically pleasing reconstructions.