ExposureDiffusion: Learning to Expose for Low-light Image Enhancement Supplementary Material

Yufei Wang[1], Yi Yu[1], Wenhan Yang[2], Lanqing Guo[1], Lap-Pui Chau[3], Alex C. Kot[1], Bihan Wen[1]*
[1]Nanyang Technological University [2]Peng Cheng Laboratory
[3]The Hong Kong Polytechnic University
{yufei001, yuyi0010, lanqing001, eackot, bihan.wen}@ntu.edu.sg
yangwh@pcl.ac.cn lap-pui.chau@polyu.edu.hk

1. Detail of proofs

The detailed proof of Eq. (7) in the main paper is as follows where Jensen's inequality is used

$$
\begin{aligned}
&\mathcal{D}_{KL}(p_\Theta(X_{0:T})||q(X_{0:T})) \\
&= \mathbb{E}_{p_\Theta(X_{0:T})}[\log \frac{p_\Theta(X_{0:T})}{q(X_{0:T})}] \\
&= \mathbb{E}_{p_\Theta(X_{0:T})}[\log \frac{p_\Theta(X_{0:T})}{\mathbb{E}_{q(X_{ref})}[q(X_{0:T}|X_{ref})]}] \\
&\leq \mathbb{E}_{q(X_{ref})}[\mathbb{E}_{p_\Theta(X_{0:T})}[\log \frac{p_\Theta(X_{0:T})}{q(X_{0:T}|X_{ref})}]] \\
&= \mathbb{E}_{q(X_{ref})}[\mathbb{E}_{p_\Theta(X_{0:T})}[\log \frac{p_\Theta(X_T)\prod_{t=1}^{T} p_\Theta(X_{t-1}|X_t)}{q(X_T|X_{ref})\prod_{t=1}^{T} q(X_{t-1}|X_t, X_{ref})}]] \\
&= \mathbb{E}_{q(X_{ref})}[\mathbb{E}_{p_\Theta(X_{0:T})}[\frac{p_\Theta(X_T)}{q(X_T|X_{ref})} + \sum_{t=1}^{T} \log \frac{p_\Theta(X_{t-1}|X_t)}{q(X_{t-1}|X_t, X_{ref})}]] \\
&= \mathbb{E}_{q(X_{ref})}[\mathcal{D}_{KL}(p_\Theta(X_T)||q(X_T|X_{ref})) \\
&\qquad + \sum_{t=1}^{T} \mathbb{E}_{p_\Theta(X_t)}[\mathcal{D}_{KL}(p_\Theta(X_{t-1}|X_t)||q(X_{t-1}|X_t, X_{ref}))]].
\end{aligned}
$$

Since modeling the ground-truth exposure process requires scene irradiation, we need to use a new variable X_{ref} , that is, a ground-truth image ideally without noise, in our derivation. However, X_0 may not be noise-free, so an extra step is conducted to further remove the noise in X_0, resulting in $\hat{X}_{ref} = F_\Theta(X_0)$, as illustrated in algorithms presented in the main paper

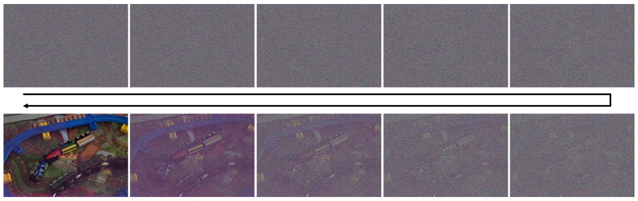

Figure 1: The reverse (denoise) process of the vanilla conditional diffusion model [3] combined with the sampling strategy from DDIM [4]. The image restoration process starts with pure noise and gradually removes noise in it. The noise in each middle step is expected to obey Gaussian distribution so that the low-light image is not any step in it.

For the image reconstruction loss, we aim to minimize the KL divergence between $p_\Theta(X_{t-1}|X_t)$ and $q(X_{t-1}|X_t, X_{ref})$ as follows

$$
\begin{aligned}
&\mathcal{D}_{KL}(p_\Theta(X_{t-1}|X_t)||q(X_{t-1}|X_t, X_{ref})) \\
&= \int p_\Theta(X|X_t) \log \frac{p_\Theta(X|X_t)}{q(X|X_t, X_{ref})} dX \\
&= \int p_\Theta(X|X_t)[X \log \frac{F_\Theta(X_t)}{X_{t-1}} + X_{t-1} - F_\Theta(X_t)] dX_{t-1} \\
&= F_\Theta(X|X_t) \log \frac{F_\Theta(X_t)}{X_{t-1}} + X_{t-1} - F_\Theta(X_t).
\end{aligned}
$$

We do not observe an obvious difference of the converged performance between using L1 divergence and KL divergence. The reason may be that even for the long exposure images, *e.g.*, the reference images we used in the SID dataset, they are still not noise free. Specifically, compared to L1 loss, KL loss may be more susceptible to the influence of noise with smaller values in non-ideal situations. For simplicity, we utilize L1 loss as a substitution. Following previous works that find using the same weighting for losses from different steps achieves slightly better performance, we keep the weight of the reconstruction loss for each step the same.

*Corresponding author.

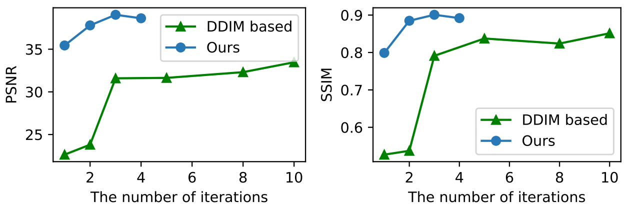

Figure 2: The convergence of the performance under a different number of inference steps. The performances are evaluated on the ×100 task of SID [1] and models are trained with P+G noise model. The diffusion model and the proposed model have a similar number of parameters and FLOPs as shown in Table 1 (b). The proposed method can get converged much faster than DDIM, leading to smaller inference overhead.

2. Comparison with diffusion models

Experiment setup. To further demonstrate the superiority of the method compared with vanilla diffusion models, we compare the proposed method with a conditional diffusion model adopting the condition strategy of [3] and sampling strategy of DDIM [4]. Specifically, similar to [3], the low-light image is concatenated as a part of the input. The original backbone of [4] includes more than $100M$ parameters and $2T$ FLOPs for a single step of a 512×512 image patch, which is not feasible for raw images with much higher resolution, *e.g.*, 4246×2840. Therefore, we slightly adjust the number of channels and blocks to obtain a similar model size to the commonly used backbone in raw low-light image enhancement tasks [5, 2]. The quantitative results are reported in Table 1. Our method outperforms the conditional diffusion model by a large margin when using a model with similar size.

Visualization of vanilla diffusion models. The visualization of the vanilla conditional diffusion models is illustrated in Fig. 3 and 1. Specifically, Fig. 3 is the inverse process that gradually removes noise from a pure noise image. For most of its middle steps, the signal-noise ratio is even lower than that of the input, causing extra inference steps. Besides, for each middle step, the noise is expected to obey the Gaussian distribution, which is different from the real noise distribution. Besides, the estimated clean image of each step is shown in Fig. 1. From the results, we can find while the quality of the reconstructed image gradually improved, there is still intense residual noise even in images of the last few steps. We conjecture the reason is that the noise distribution of the inference process is different from the training one due to cumulative error so the noise can not be effectively removed in the last few steps. We find that the needed number of inference steps are closely related to the noise level, *e.g.*, ISO and exposure time. An adaptive algorithm for deciding the number of inference steps may be left for future work.

The convergence. The comparison of the convergence speed for inference between the proposed method and the vanilla one can be seen in Fig. 2. The vanilla diffusion model needs a relatively large number (*e.g.*, around 10 steps) of inference steps to get converged since they need to denoise from pure noise, which hinders potential applications. The proposed method can get converged quickly in a few steps for the cases the input is not very dark, *e.g.*, ×100 task, and the improvement mainly from the proposed training mechanism. On the contrary, for the harder cases, iterative refinement can significantly improve the performance as shown in Fig. 6 in the main paper. For the convergence of training, we find that the vanilla diffusion-based method requires a larger batch size and more epochs, *e.g.*, we train the vanilla conditional diffusion models for 10000 epochs with a batch size of 64 on four RTX A5000s. While the proposed model can be well converged and trained on only one RTX A5000 with a batch size of 1 and epoch of 300.

Quantitative results. As we can see in Table 1, the diffusion models are very sensitive to the capacity of the model size, *e.g.*, the larger vanilla diffusion model can achieve much better performance than that of a smaller one. Besides, although the model "Diffusion-based-1" has more parameters and FLOPs than the backbone we used, it achieves much worse performance than the proposed method. The reason may be that the vanilla conditional diffusion models need more network capacity to learn the denoising of Gaussian noise with different noise levels.

3. Experiment details

For the detailed settings of experiments, we mainly follow the settings of ELD [5]. Specifically, we adopt the released noise parameters, and the basic noise model $P + G$ from [5]. We implement their proposed noise model $P + G^* + r + u$ and the training on-the-fly manner by ourselves and achieve closing results with the reported in the paper. For the experiments based on PMN [2], *e.g.*, the results in Table 4 in the main paper, we follow their training strategy, *e.g.*, the same batch size, data augmentation strategy, and learning rate. However, PMN [2] adapts a slightly different evaluation strategy with ELD [5], *e.g.*, different evaluation set. To unify the results, we adopt the same evaluation strategy as ELD [5] for all experiments. For the schedule of λ_t, we adopt the linear scheduler, *e.g.*, the exposure time of each step for the ×300 task is $[\lambda_T, 100 \cdot \lambda_T, 200 \cdot \lambda_T, 300 \cdot \lambda_T]$.

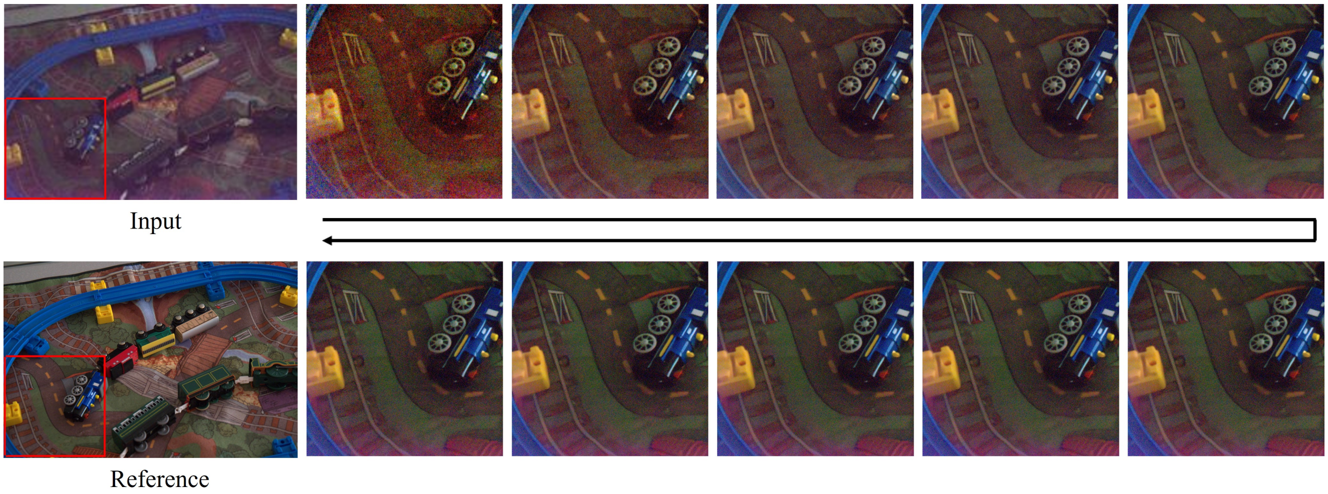

Figure 3: The diffusion process of the vanilla conditional diffusion model [3] combined with the sampling strategy from DDIM [4]. The image restoration process starts with pure noise and gradually removes noise in it.

	Model	PSNR / SSIM
	UNet	**38.88 / 0.901**
×100	Diffusion-based-1	33.47 / 0.851
	Diffusion-based-2	25.13 / 0.539
	UNet	**36.02 / 0.832**
×250	Diffusion-based-1	31.98 / 0.823
	Diffusion-based-2	24.79 / 0.529
	UNet	**34.59 / 0.798**
×300	Diffusion-based-1	31.39 / 0.807
	Diffusion-based-2	22.65 / 0.527

(a) Performance with different models.

Model	Parameters	FLOPs
Ours (UNet-based)	7.762M	55.17G
Diffusion-based-1	8.619M	367.15G
Diffusion-based-2	1.472M	63.07G

(b) Computational cost of each model.

Table 1: Performance of models w/ and w/o the proposed method under different noise models and backbones on SID [1] dataset. The models are trained on P+G noise model.

4. More visualization results

More results on SID [1] and ELD [5] are shown in Fig. 4, Fig. 5, and Fig. 6. The non-deep based method and the deep model trained without reference images are likely to suffer serious color distortion. The proposed method achieves the best perceptual quality overall, *i.e.*, less color distortion and better details even compared with P+G which are trained under the same noise model and backbone network.

References

[1] 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, 4, 5

[2] Hansen Feng, Lizhi Wang, Yuzhi Wang, and Hua Huang. Learnability enhancement for low-light raw denoising: Where paired real data meets noise modeling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1436–1444, 2022. 2

[3] 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. 1, 2, 3

[4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3

[5] 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, 3, 6

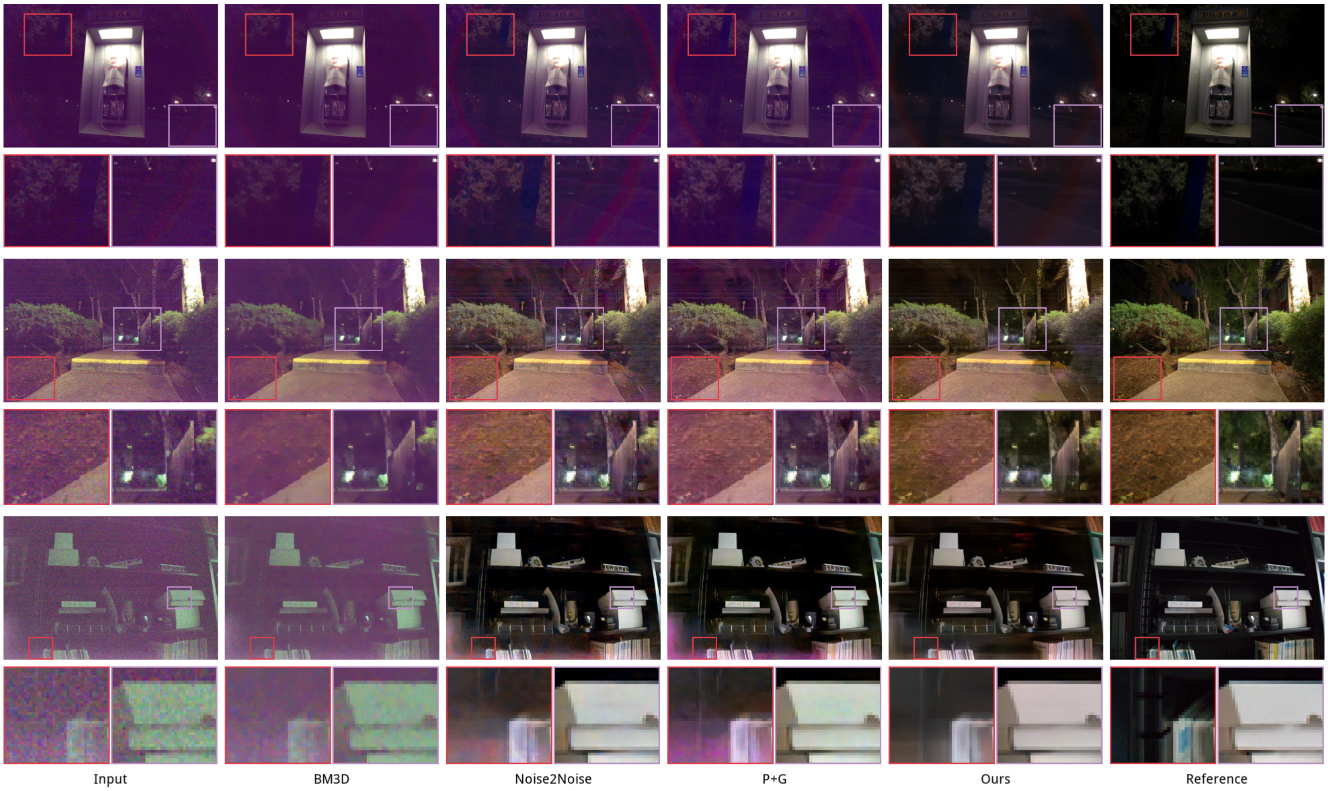

Figure 4: Visual comparison of low-light image enhancement results on SID [1]. We use the same backbone and noise model as P+G. All the images are passed through ISP using the same white balance for better visualization.

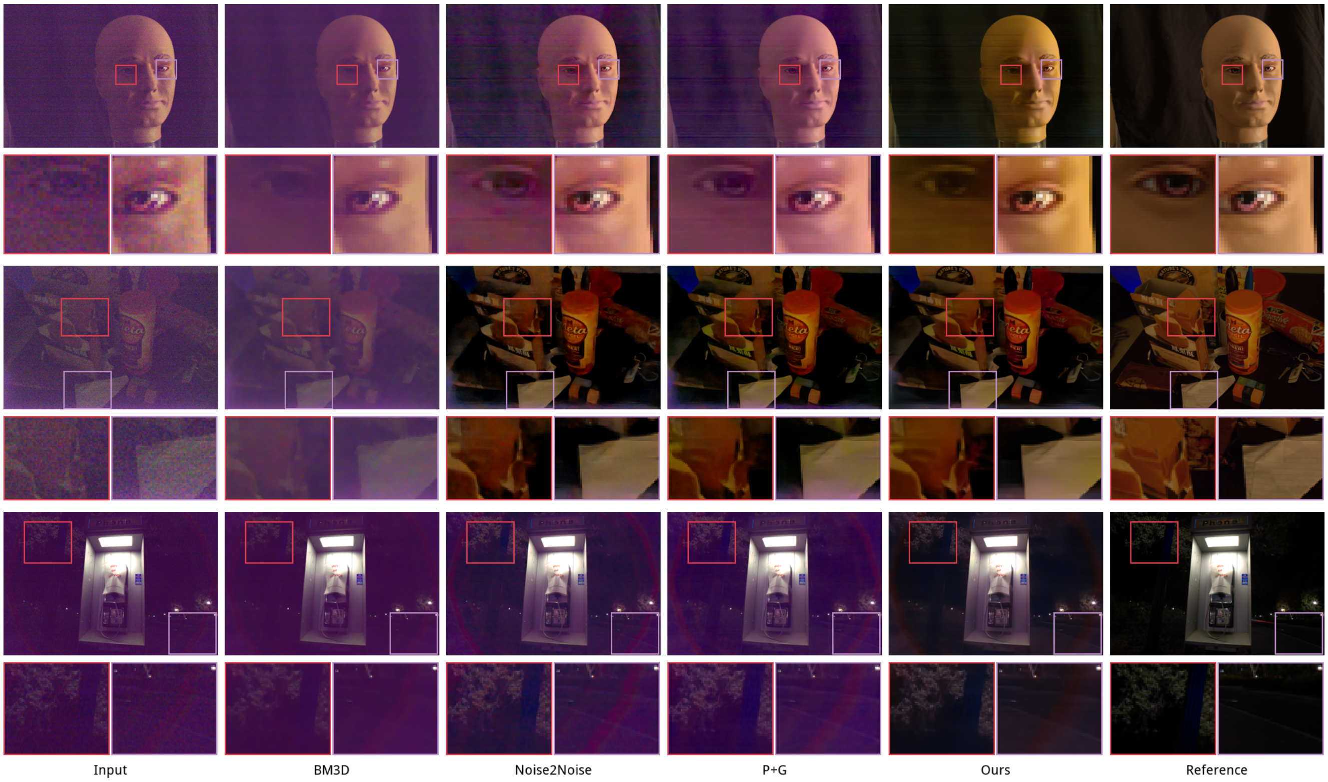

Figure 5: Visual comparison of low-light image enhancement results on SID [1]. We use the same backbone and noise model as P+G. All the images are passed through ISP using the same white balance for better visualization.

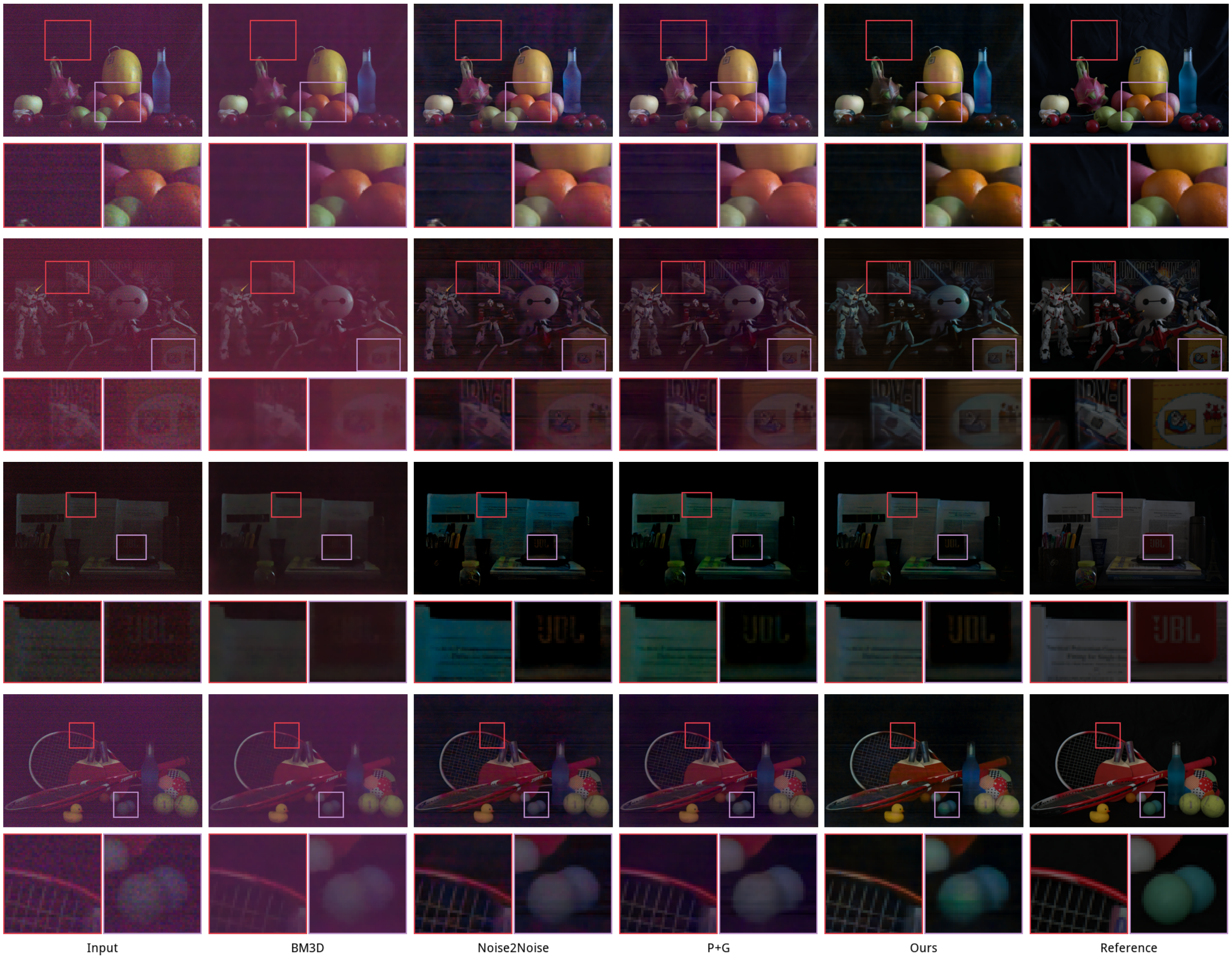

Figure 6: Visual comparison of low-light image enhancement results on ELD [5]. We use the same backbone and noise model as P+G. All the images are passed through ISP using the same white balance for better visualization.