

Supplementary Materials: JoReS-Diff: Joint Retinex and Semantic Priors in Diffusion Model for Low-light Image Enhancement

Anonymous Authors

1 OVERVIEW

In this document, we describe the architecture and training details of the proposed JoReS-Diff in Section 2. We provide additional ablation study in Section 4. We present additional visual comparisons with existing SOTA methods on real-world datasets, including LOL, LOL-v2-real, LOL-v2-synthetic, UHD-LL, MIT-Adobe-FiveK, and ISTD in Section 3. Finally, we present limitations and future works in Section 5.

2 ADDITIONAL IMPLEMENTATION DETAILS

2.1 Pre-Trained Decomposition Net

DNet adopts a lightweight UNet-like network to learn the decomposition mapping based on the assumption of Retinex theory. It first takes both the low/normal-light images $I, \hat{I} \in \mathbb{R}^{W \times H \times 3}$ as input and extract the features $I_f, \hat{I}_f \in \mathbb{R}^{W \times H \times C}$. Then I_f, \hat{I}_f are processed through convolutional encoder and decoder, and two output heads are followed to map the features to reflectance R, \hat{R} and illumination I, \hat{I} . The assumption of Retinex theory indicates the similarity between R and \hat{R} , the smoothness of I, \hat{I} and the mutual reconstruction ability of reflectance with various illumination. Inspired by the training strategy in [2], we reasonably utilize the constant reflectance loss, smooth illumination loss and reconstruction loss to pre-train DNet. The assumption of Retinex theory indicates the similarity between R and \hat{R} , the smoothness of I, \hat{I} and the mutual reconstruction ability of reflectance with various illumination [2]. Therefore, we first introduce the constant reflectance loss \mathcal{L}_R and constrain the similarity of R and \hat{R} as:

$$\mathcal{L}_R = \|R - \hat{R}\|_1, \quad (1)$$

Then we adopt the smooth illumination loss \mathcal{L}_L to minimize the gradient and reserve the textures as:

$$\mathcal{L}_L = \|\nabla L \cdot \exp(-w\nabla I)\|_1 + \|\nabla \hat{L} \cdot \exp(-w\nabla \hat{I})\|_1, \quad (2)$$

where ∇ denotes the derivative operator and w is the weight term controlling the magnitude trade-off between gradient suppression and texture preservation. Furthermore, we conduct the reconstruction loss \mathcal{L}_{mr} as:

$$\mathcal{L}_{mr} = \sum_{r=R, \hat{R}} \|r \cdot L - I\|_1 + \|r \cdot \hat{L} - \hat{I}\|_1, \quad (3)$$

Thus, we pre-train DNet by the overall loss:

$$\mathcal{L}_{DNet} = \lambda_R \mathcal{L}_R + \lambda_L \mathcal{L}_L + \lambda_{mr} \mathcal{L}_{mr}. \quad (4)$$

The proposed JoReS-Diff uses the pre-trained DNet to obtain Retinex-based priors. The advantage of only extracting Retinex components is that, while the architecture of DNet is simple and lightweight, the quality of the produced components could be favorable. To conduct the experiments on benchmark datasets, the DNet is pre-trained on corresponding datasets with the patch and batch size of 128 and 16 for 500 epochs. Furthermore, the weights of

the DNet are fixed during the training stage to exploit the prepared Retinex-based priors.

2.2 Details of Adjustment Net

ANet is trainable with the whole JoReS-Diff. The network architecture of ANet is similar to DNet, while exists some modifications for better adaptation of the adjustment task. For maintaining the feature distribution of Retinex components, ANet adopts skip connections between the encoder and the decoder. Furthermore, each layer in ANet outputs the feature map with the same number of embedding channels, which reduces the parameters of the network and the extra computational cost. The specific implementation details can be found in the codes.

The original purpose of ANet is to provide adjusted Retinex components. However, we find that it could produce under-optimized components if the inputs are low-light and noisy images in early steps, which compromises the reliability of the prior and finally causes unexpected outputs. Moreover, the intermediate features also contain the learned adjustment mapping and are included in the Retinex-based conditions as $\hat{c}_t = [R'_t, L'_t, F_t]$, which is described in Section 3.2.1 in main paper. The decoder of ANet contains three layers with different resolutions and the same number of channels. The resolutions of the streams are 1, 1/2, and 1/4, which generate three features with corresponding resolutions ($F_0 \in \mathbb{R}^{H \times W \times C}$, $F_1 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$, and $F_2 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$, where $C = 64$) respectively. The final output layer of ANet mixes the three features by upsampling and concatenating and produces the reflectance and illumination maps with the sizes of $R'_t \in \mathbb{R}^{H \times W \times 3}$ and $L'_t \in \mathbb{R}^{H \times W \times 1}$. Finally, we take three multi-scale features and the predicted maps as Retinex-based priors to guide the enhancement process, the former is utilized to optimize the image feature by the FRCM module and the latter provides image-level refinement by the IRCM module. More details are shown in the code.

2.3 Supplementary Derivation of Eq. (16) in main paper

In Section 3.2.2 in main paper in the main paper, we reformulate Eq. (6) in main paper into a residual refinement manner based on the analysis in Section 3.2 in main paper and propose IRCM for refining the estimated \hat{x}_0 . To briefly revisit the analysis, we first present the Retinex theory and describe the details and shortcomings of the idea model, which indicates that the decomposed maps are more suitable for acting as guidance. Then we propose to preserve the original information by introducing the low-light input and regarding the reflectance \hat{R} as auxiliary guidance. Finally, we supplement the illumination \hat{L} and obtain the adjusted term $\hat{c} = \mathcal{F}_{\mathcal{A}}(R, L)$ for sufficient guidance.

Eq. (6) in main paper describes the original Retinex model as follows:

$$\begin{aligned}\hat{R} &= \log(I) - \log(L), \\ \hat{I} &= \mathcal{T}(\exp(\hat{R})),\end{aligned}\quad (5)$$

where $\mathcal{G}(\cdot)$ and $\mathcal{T}(\cdot)$ denote the convolution with the Gaussian surround function and the linear transformation function. Inspired by the analysis, we reformulate the Retinex model by treating the \hat{R} and \hat{I} as residual terms and substitute the physical operators with convolutional layers $W(\cdot)$ as follows:

$$\begin{aligned}\Delta\hat{R} &= W(I) - W(L), \\ \Delta I &= W(\hat{R} + \Delta\hat{R}), \\ \hat{I} &= I + \Delta I,\end{aligned}\quad (6)$$

where \hat{I}' is the final result. To fully exploit the residual terms, we adopt scaling and shifting operations $\mathcal{F}_T(\cdot|\cdot)$ instead of the simple addition and subtraction as follows:

$$\begin{aligned}\Delta\hat{R} &= \mathcal{F}_T(W(L)|W(I)), \\ \Delta I &= \mathcal{F}_T(W(\hat{R})|W(\Delta\hat{R})), \\ \hat{I} &= \mathcal{F}_T(W(I)|W(\Delta I)).\end{aligned}\quad (7)$$

As described in Section 3.2.2 in main paper, we obtain the \hat{x}_0 , R_t' and L_t' from the UNet and ANet, which correspond to the I , \hat{R} and L , respectively. Furthermore, we also integrate the output of FRCM $F_{\hat{x}_0}'$ to ensure the preservation of multi-scale information of the ANet for better contents and details. Therefore, we achieve the refinement of the estimated \hat{x}_0 as follows:

$$\begin{aligned}\Delta R_t' &= \mathcal{F}_T(W(L_t')|W(\hat{x}_0)), \\ \Delta\hat{x}_0 &= \mathcal{F}_T(W(R_t')|W(\Delta R_t')), \\ \hat{x}_0' &= \mathcal{F}_T(W(\hat{x}_0)|W(\Delta\hat{x}_0)) + W(F_{\hat{x}_0}'),\end{aligned}\quad (8)$$

3 MORE VISUAL COMPARISONS

Visual Comparison With Other LLIE Methods. As shown in Figs. 1 to 8, we give more visual results of our JoReS-Diff and other baseline methods on LOL, LOL-v2 and UHD-LL datasets as the supplement of the visualization in the main paper. We also presents the visual results on MIT-Adobe-FiveK and ISTD datasets to support the quantitative comparison in the main paper. Consequently, we can see that our method consistently produces more natural results and achieves superior performance over the baseline methods in various scenes, especially compared with other diffusion-based methods [1, 4], sufficiently demonstrating the effectiveness of the novel Retinex-based condition strategy. Notably, you may zoom in the figures for better visibility since the images in the UHD-LL dataset have high resolutions.

Visual Comparison Between Ablated Settings. As shown in Figs. 10 and 11, we provide more visual comparisons for investigating the contribution of the proposed ANet. The “w/o ANet” and “w/ ANet” denote whether the reflectance and illumination maps are adjusted by the ANet, respectively. We show a series of decomposed maps to illustrate the effects of the ANet at different steps. It is clear that the reflectance maps derived without ANet contain more artifacts and noise, and the illumination maps adjusted by the ANet exhibit more pleasing brightness. The visual comparisons

sufficiently demonstrate the effectiveness of the ANet. Notably, the reflectance maps in Step 8 contain visually unsatisfactory parts. Although the severe noise in the original reflectance maps led to the failed situations, the quality of the maps in the following steps is still favorable (at steps 6,4,2). Therefore, the diffusion model still performs well since the final results only depend on the previous step. Furthermore, we provide the visual comparison on MIT-Adobe-FiveK for investigating the effectiveness of the semantic prior. As shown in Fig. 9, the images output by the model without semantic prior exhibits conspicuous color shift and unclear boundaries.

4 SUPPLEMENTARY EXPERIMENTS

According to the analysis in Section 3.2.1 in main paper, the multi-scale features are crucial for fully exploiting the learned mapping and we propose FRCM to incorporate the beneficial features. Therefore, as illustrated in Section 3.2.2 in main paper, the RNet consists of the FRCM and IRCM. We set the number of the FRCM and IRCM to 1 respectively as our baseline and produce the results on the benchmark datasets and other ablation studies. To further investigate the effects of different numbers of the FRCMs and IRCMs in RNet, we conduct experiments on the LOL dataset with different settings. As shown in Table 1, we set the number of FRCMs and IRCMs to 0,1,2 respectively to illustrate the effects of the different architectures of RNet. The comparison between without any RCM and the baseline shows the effectiveness of our design. Furthermore, both without the FRCM and the IRCM obtain the performance decrease compared with the baseline, which is also exhibited in Table 5 in the main paper. Notably, in the bottom half of the table, we add more modules to the baseline, while the metrics are worse than the baseline instead of becoming more promising. The performance decrease is especially severe when adding extra FRCMs, which shows that reusing the multi-scale features is harmful to the iterative enhancement. Thus, we utilize one FRCM and one IRCM in RNet to achieve the best performance.

Table 1: Ablation study on the LOL dataset for investigating the different numbers of the FRCMs and IRCMs.

Settings		PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
FRCM	IRCM			
0	0	26.891	0.871	0.117
0	1	27.344	0.875	0.112
1	0	27.262	0.872	0.119
1	1	27.626	0.884	0.090
2	1	26.371	0.861	0.124
1	2	27.232	0.878	0.102
2	2	26.334	0.859	0.127

5 LIMITATIONS AND FUTURE WORKS

In this section, we discuss the limitations of our work and suggest the potential future research directions of diffusion-based methods for low-light enhancement and other low-level vision tasks.

Limitations. First, while our JoReS-Diff possesses superior enhancement capability thanks to the Retinex-based condition strategy, the entire framework is heavily reliant on the quality of Retinex-based priors provided by DNet. The CNN-based DNet with limited

parameters may cause the loss of content information during the decomposition process. If we obtain the pre-processed conditions with bad quality, the final results will be unsatisfactory. Furthermore, we have to pre-train DNet on the corresponding dataset before we train our JoReS-Diff, which is more complex than the totally end-to-end methods. Thus, the way of obtaining Retinex-based conditions can be improved in an end-to-end manner, and the architectures and the training settings can be well-designed for better decomposition.

Second, the proposed components of our JoReS-Diff (*i.e.*, DNet, ANet, RNet) are preliminary techniques for introducing Retinex-based priors into the diffusion process, which may limit the potential of our method. The architecture of DNet and ANet are simple UNet-like networks without special designs and still have the potential to be optimized. The FRCM and IRCM in RNet only contain several layers to conduct the feature transformation for the consideration of testing speed. Then, the simple manner of integrating multi-scale features in the FRCM may undermine the benefits of feature-level interaction. However, we have tried the Transformer block for feature fusion, which makes the model training and inference times significantly slower. Hence, we can explore both efficient and powerful networks for producing the Retinex-based conditions and more suitable interaction manners to achieve effective feature fusion.

Third, the training cost of the diffusion model is much more than the traditional deep-learning methods. For instance, before 2023, most of the methods without the diffusion model only required less than 10k iterations for training and generally completed the training stage within one day, while the diffusion-based methods need orders of magnitude more iterations (100k for DiffLL [1], 800k for Diff-Retinex [3], 500k for PyDiff [4], 600k for our JoReS-Diff). Furthermore, in the inference stage, the diffusion-based method generally costs more time as well. With the advancement of cameras, image resolution will become increasingly higher, posing new challenges for the exploration of efficient methods, particularly for diffusion-based models. Therefore, we have to rethink the feasibility and efficiency of the diffusion-based method for low-level vision tasks.

Future works. According to the limitations of our JoReS-Diff, our future works can be organized as follows. **(a)** We tend to explore a pre-training strategy to improve the capability of generalizing on various datasets and scenarios and integrate the DNet into the whole JoReS-Diff for an end-to-end training manner. **(b)** For DNet and ANet, we plan to refine the network by reasonably introducing an attention mechanism while maintaining the lightweight architecture and improving the capability of decomposition and adjustment. For FRCM in RNet, we will try more sufficient interaction manners of the multi-scale features for better refinement. **(c)** The long convergence time and the iterative inference of the diffusion model make it impractical. We hope to explore the possibilities of applying the diffusion model to the sub-tasks that consume less time instead of the overall image enhancement task.

REFERENCES

- [1] Hai Jiang, Ao Luo, Songchen Han, Haoqiang Fan, and Shuaicheng Liu. 2023. Low-light image enhancement with wavelet-based diffusion models. *arXiv preprint arXiv:2306.00306* (2023).

- [2] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. 2018. Deep retinex decomposition for low-light enhancement. In *BMVC*.
- [3] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. 2023. Diff-Retinex: Rethinking low-light image enhancement with a generative diffusion model. In *ICCV*. 12302–12311.
- [4] Dewei Zhou, Zongxin Yang, and Yi Yang. 2023. Pyramid diffusion models for low-light image enhancement. *arXiv preprint arXiv:2305.10028* (2023).

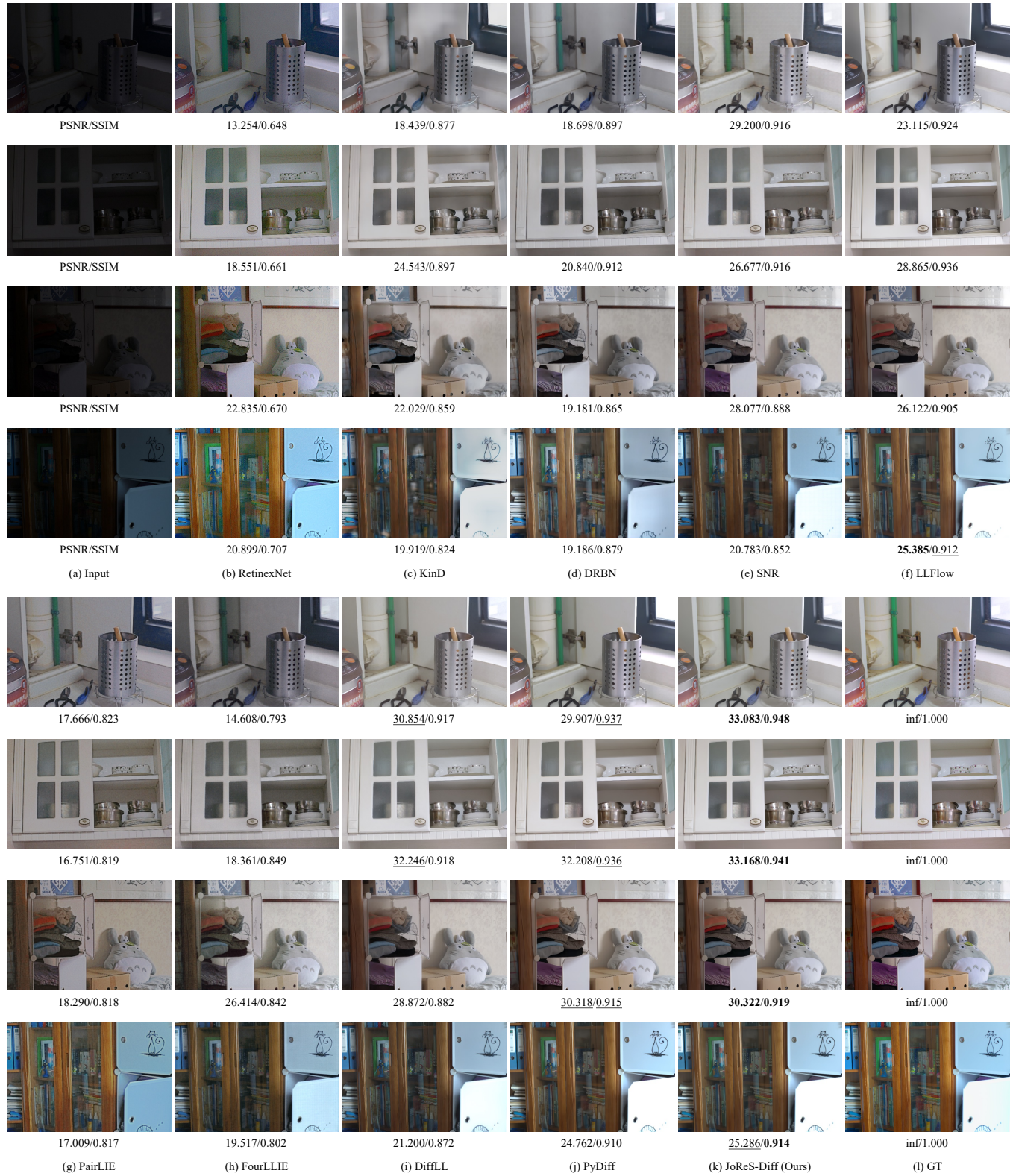


Figure 1: Visual comparison of our JoReS-Diff and the compared LLIE methods on the LOL dataset.

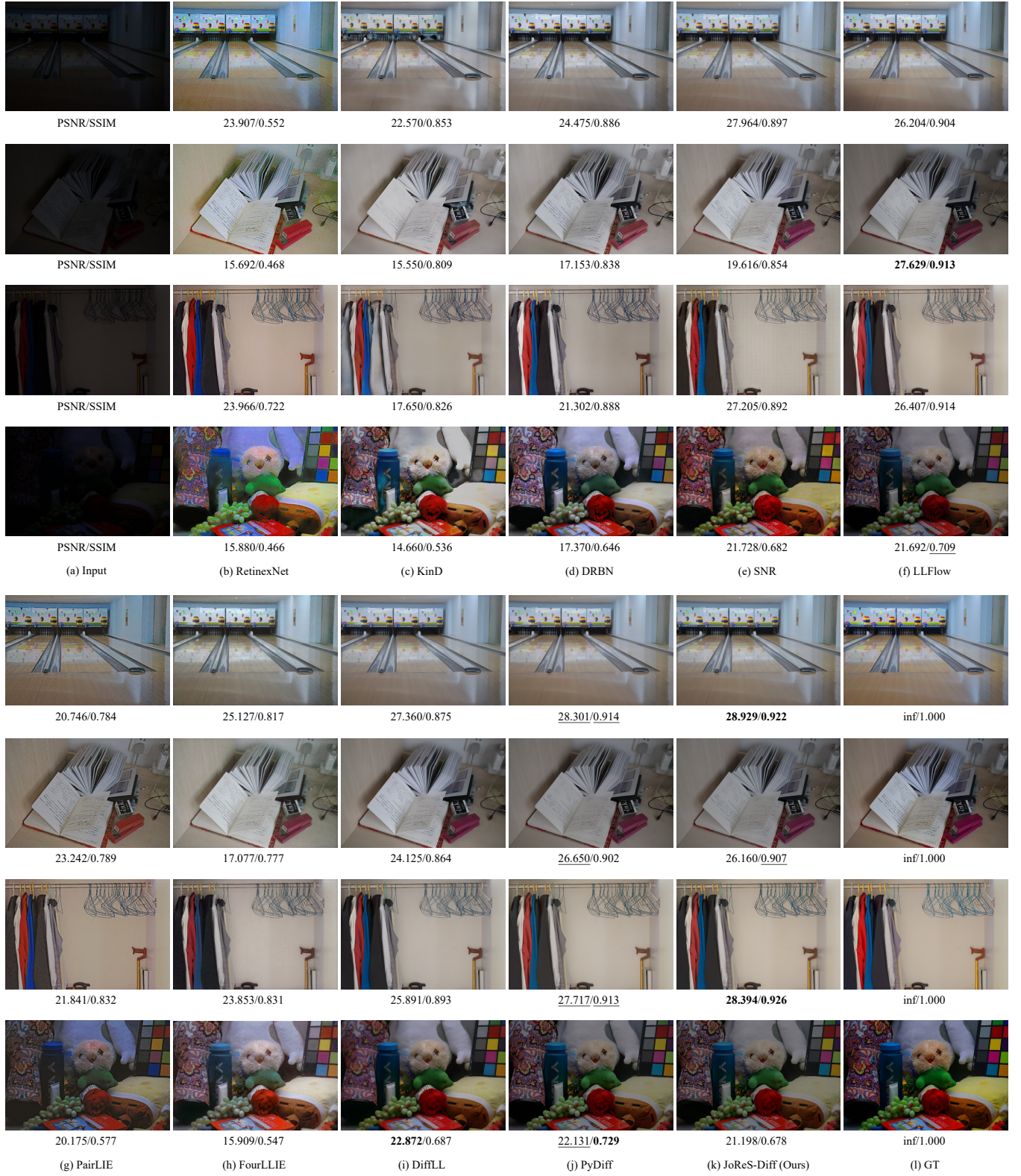


Figure 2: Visual comparison of our JoReS-Diff and the compared LLIE methods on the LOL dataset.

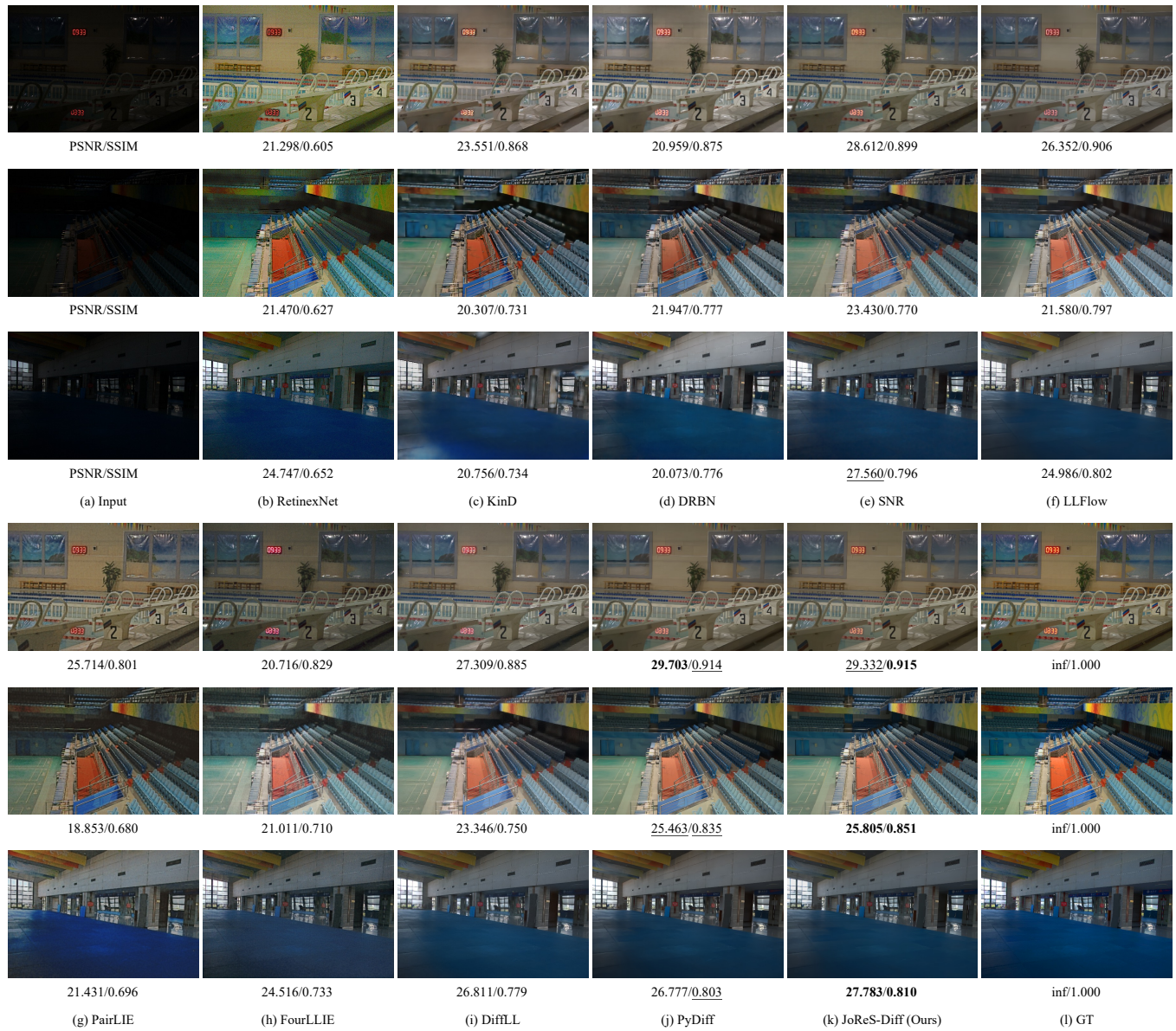


Figure 3: Visual comparison of our JoReS-Diff and the compared LLIE methods on the LOLv2 dataset.

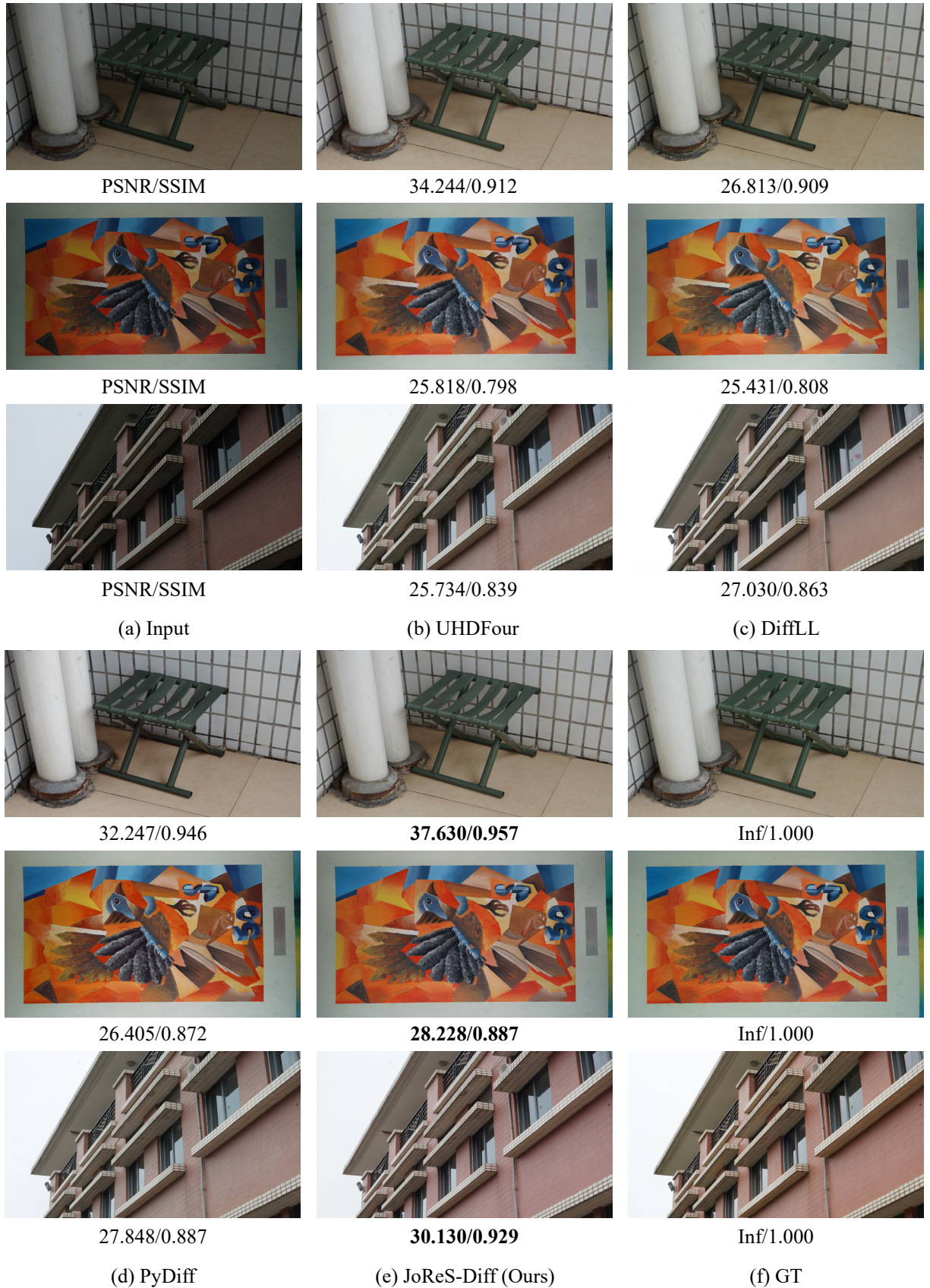


Figure 4: Visual comparison of our JoReS-Diff and the compared LLIE methods on the UHD-LL dataset.



Figure 5: Visual comparison of our JoReS-Diff and the compared LLIE methods on the UHD-LL dataset.



Figure 6: Visual comparison of our JoReS-Diff and the compared LLIE methods on the UHD-LL dataset.

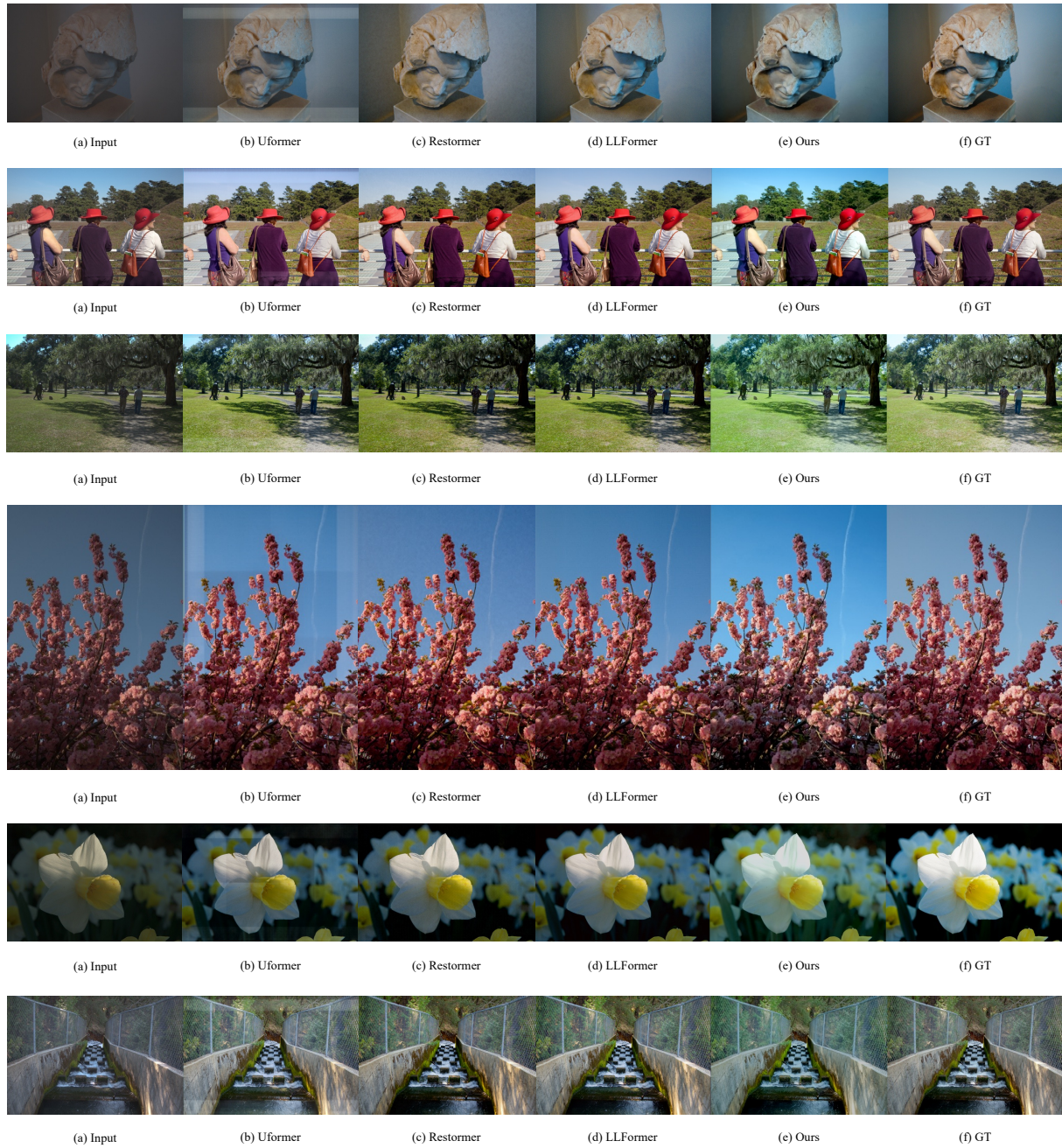
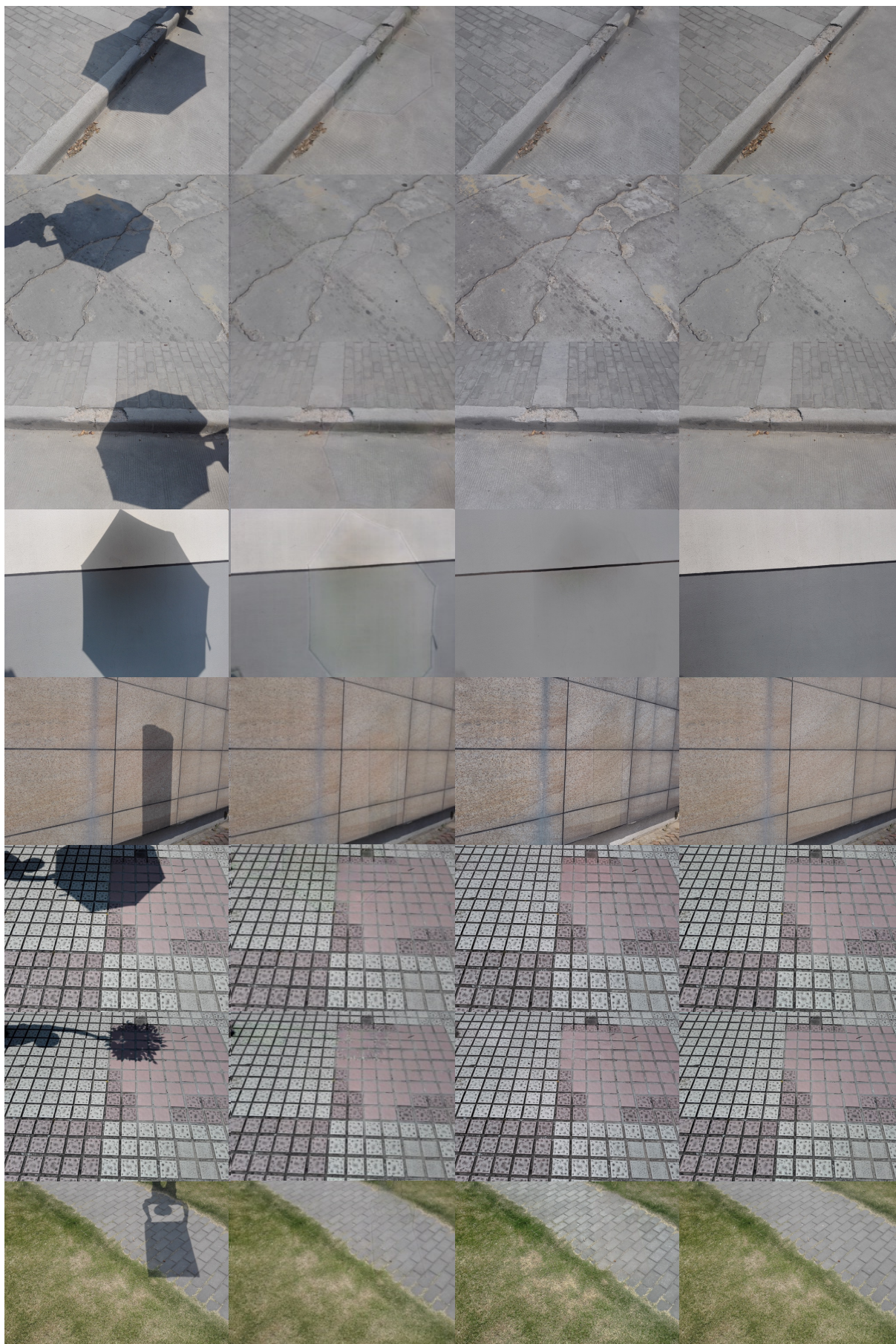


Figure 7: Visual comparison of our JoReS-Diff and the compared methods on the MIT-Adobe-FiveK dataset.



(a) Input

(b) EMNet

(c) Ours

(d) GT

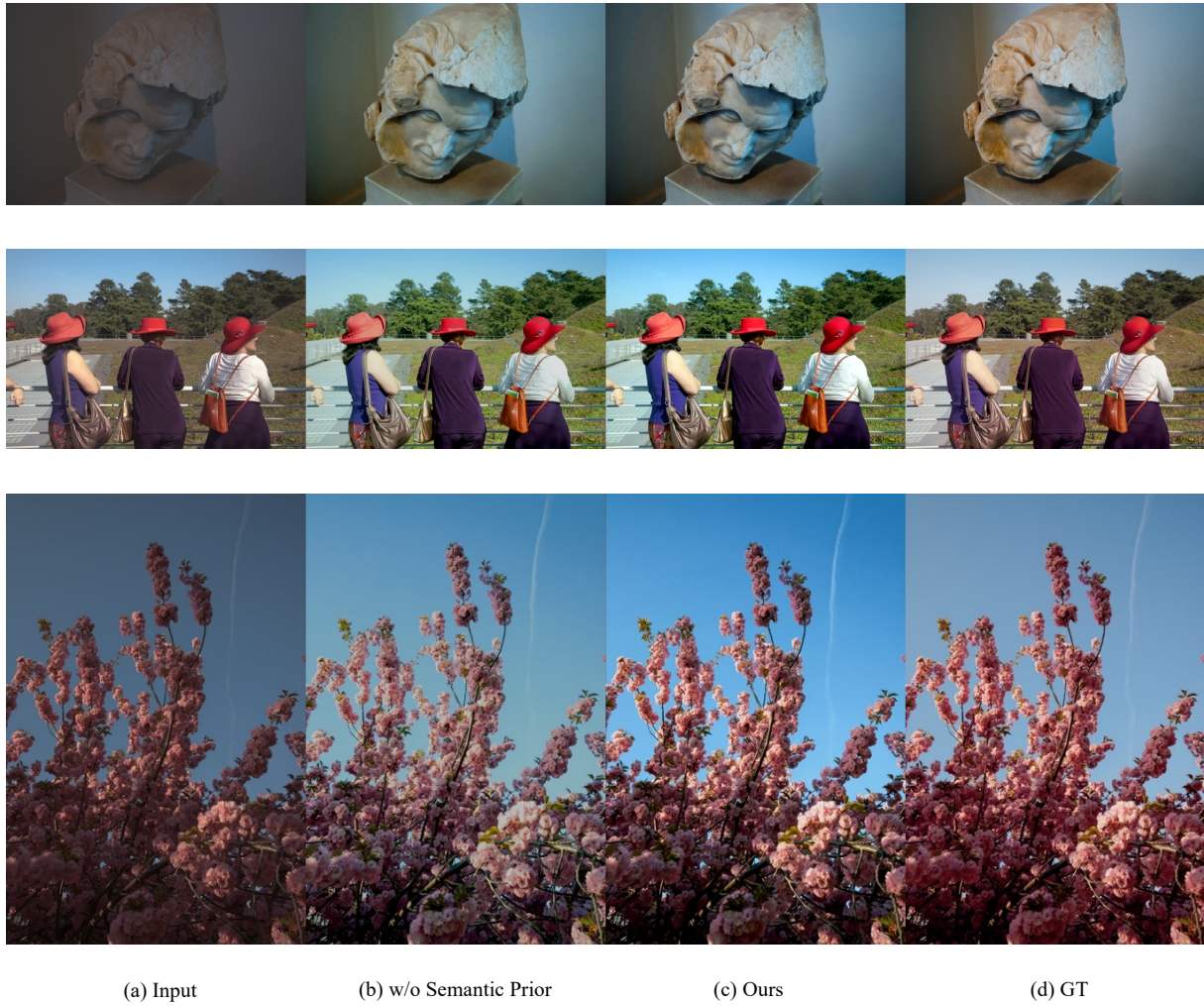


Figure 9: Ablation study on MIT-Adobe-FiveK dataset for investigating the effects of semantic prior.

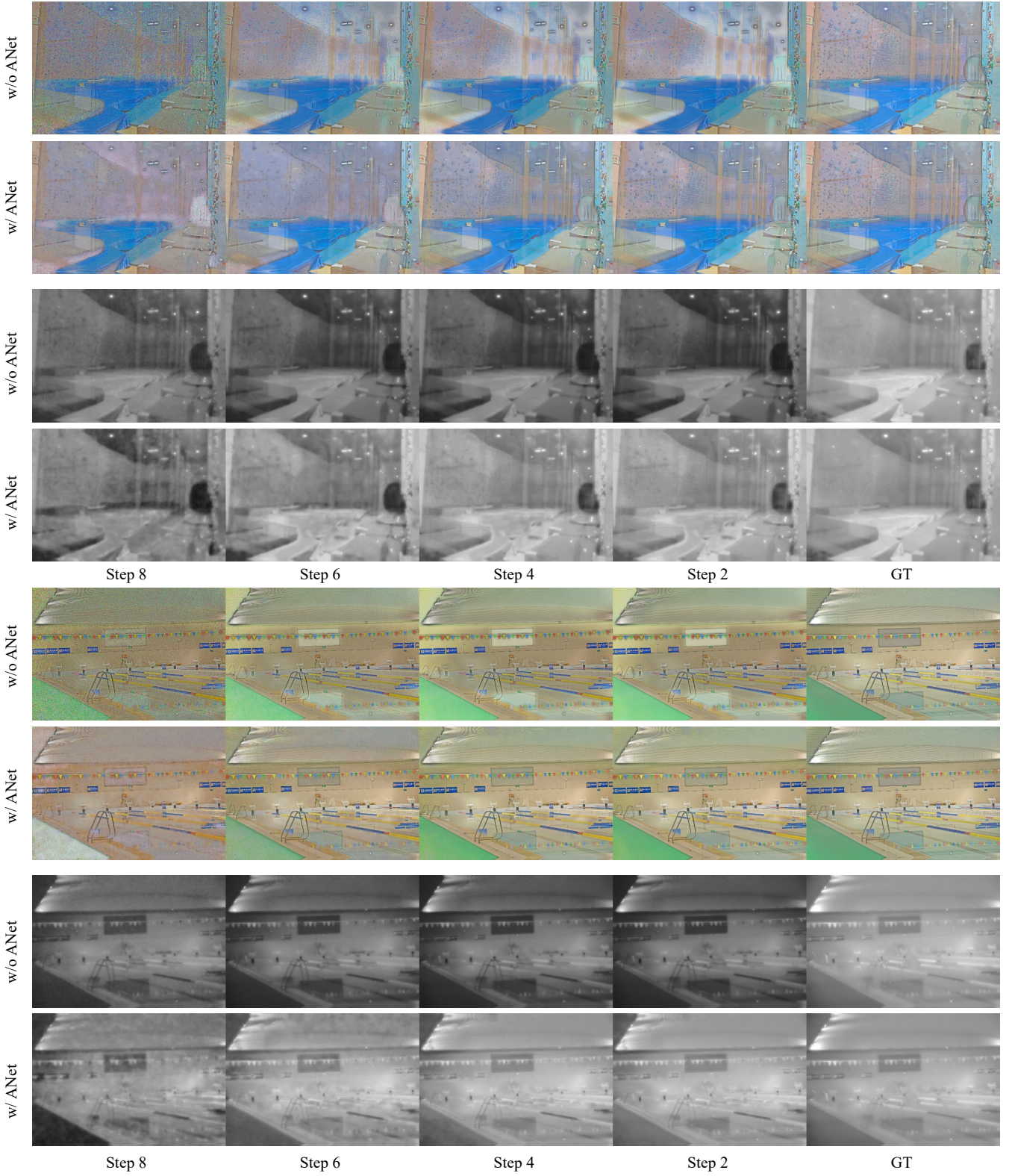


Figure 10: Ablation study on LOLv2 dataset for investigating the effects of ANet. The reflectance (rows 1,2,5,6) and illumination (rows 3,4,7,8) at selected steps are shown from top to bottom. GT denotes the decomposed results of the normal-light image.

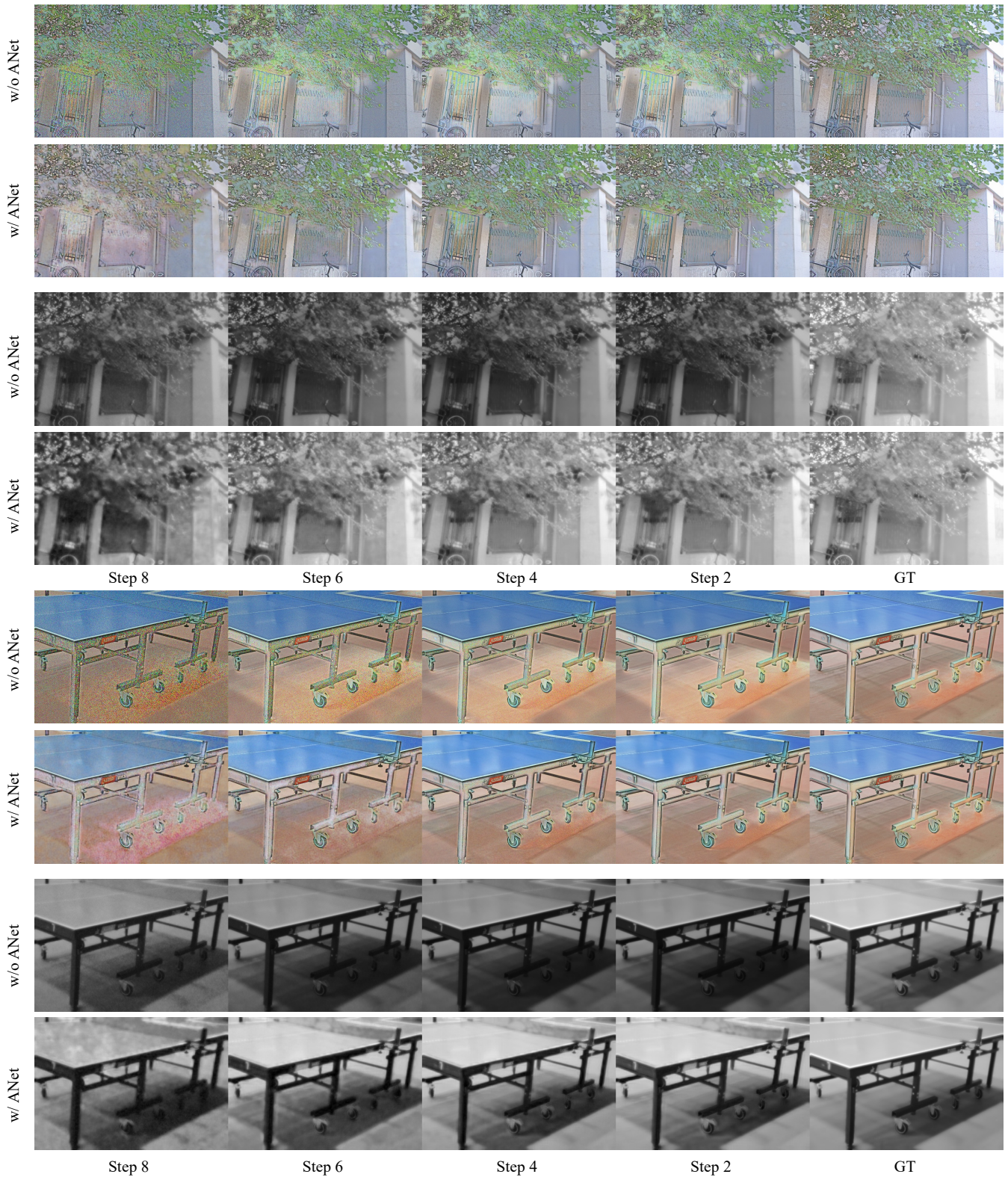


Figure 11: Ablation study on LOLv2 dataset for investigating the effects of ANet. The reflectance (rows 1,2,5,6) and illumination (rows 3,4,7,8) at selected steps are shown from top to bottom. GT denotes the decomposed results of the normal-light image.