Supplementary Materials of "Empowering Low-Light Image Enhancer through Customized Learnable Priors"

Anonymous CVPR submission

Paper ID 1261

This supplementary document is organized as follows:

Sec. 1 provides the discussion about the proposed "Empowering Low-Light Image Enhancer through Customized Learnable Priors" and other Retinex learning-based methods.

Sec. 2 provides the discussion about the necessity of employing networks to learn the illumination and noise priors.

Sec. 3 provides the architecture of the customized learnable illumination prior.

Sec. 4 provides the architecture of the customized learnable noise prior.

Sec. 5 provides the ablation studies about the customized illumination and noise priors.

Sec. 6 provides the ablation studies about the implementation of the customized learnable noise prior.

Sec. 7 provides the ablation studies about the number of unfolding stages K on the Huawei Dataset.

Sec. 8 provides the architecture of illumination adjustment network and reflectance restoration network.

Sec. 9 provides the visualizations of the illumination map incorporated with the customized learnable illumination prior.

Sec. 10 provides more visualization decomposition and enhancement results.

Sec. 11 presents the broader impacts.

1. Discussion

Motivation. Inspired by the powerful learning capability of deep learning, Retinex learning-based methods, *e.g.* RetinexNet [7], KinD [10], and KinD++ [9] have achieved the remarkable progress in low-light image enhancement. However, they heuristically construct diverse network architectures in a black box fashion without considering the intrinsic priors of illumination and reflectance components, resulting in **lacking transparency and interpretability**.

To improve interpretability, a model-driven deep unfolding paradigm, URetinexNet [8], has been proposed. However, it **empirically constructs proximal operator networks in a black-box manner without considering the inherent properties of the illumination and reflectance components** to vaguely deliver the two components priors, leading to **an ambiguous and implicit prior principle**. Therefore, **the existing deep unfolding-based method cannot explicitly extract the customized illumination and reflectance priors while adhering to their intrinsic properties, which is caused by weak interpretable prior operations**. In addition, these Reintex learning-based methods **neglect the inevitable noise pollution and are not optimized in an end-to-end manner**. Motivated by the above analysis, we focus on exploring the potential of the customized learnable illumination and noise priors and further propose a more transparent and interpretable deep unfolding paradigm in an end-to-end manner.

Solution. Inspired by the innate feature representation capability of Masked Autoencoder (MAE), we **customized MAE-based illumination and noise priors** with a masked image modeling strategy. The former is trained from the normal-light image to its illumination map filtered by a bilateral filter, reducing noise without altering the intrinsic structure [1]. The latter aims to learn the histograms of oriented gradients of the normal-light image, which presents the gradient variation while satisfying the lightness independence [2]. Then, we redevelop the two customized learned priors from two perspectives: 1) **structure flow**: embedding the learned illumination prior into the design of the proximal operator in the Retinex decomposition unfolding process; 2) **optimization flow**: redeveloping the learned noise prior as a regularization term to further eliminate the noise by minimizing the gradient presentation difference between the enhanced and normal-light images. Such designs improve the interpretability and representation capability of the model.

Prophesy. To the best of our knowledge, our method is the first time to explore the customized learnable priors for low-light image enhancement task and provides a novel paradigm where the well-learned priors that account for the specific function-

abilities. Compared with the implicit prior, the proposed customized learnable priors possess much more interpretability. In comparison with the explicit prior, our learnable priors have powerful representation ability. We sincerely believe that our solution will spark the realms of the proximal optimal network designs and the development of deep unfolding paradigm.

2. Necessity of employing networks to learn the illumination and noise priors

In this paper, we explore the potential of customized learnable illumination and noise priors on low-light image enhancement. Inspired by the powerful feature representation capability of Masked Autoencoder (MAE), we customize MAE-based illumination and noise priors with customized learning targets. The former is trained from the normal-light image to its illumination map filtered by a bilateral filter, reducing noise without altering the intrinsic structure [1]. The latter aims to learn the histograms of oriented gradients of the normal-light image, which presents the gradient variation while irrelevant to enhanced lightness.

The filtered illumination map and HOG feature is only the processed version of the images, and they are essentially just the images, thus cannot describe the priors of the illumination and noise components. In contrast, benefiting from the masked image modeling strategy that reconstructs the masked patches based on the visible patches, the intermediate representations encode the corresponding properties and priors of the two components. Therefore, we embed the MAE-based customized illumination prior into the unfolding architecture for improving transparency and interpretability and redevelop the MAE-based customized noise prior as a regularization term to constrain the gradient representation consistency.

To prove the necessity of employing networks to learn the priors, we construct the ablation studies on LOL and Huawei dataset. Firstly, We embed the filter illumination map rather than the intermediate representations into the unfolding paradigm, PSNR/SSIM drop from 21.67 / 0.774 to 21.41 / 0.747 on LOL and from 20.31 / 0.658 to 20.19.92 / 0.509 on Huawei. Secondly, we remove the MAE-based noise prior network and directly employ L1 loss on the HOG features of enhanced and GT images, PSNR/SSIM drop from 21.67 / 0.774 to 21.50 / 0.748 on LOL and from 20.31 / 0.658 to 20.04 / 0.625 on Huawei. Thus, it is necessary to employ an MAE-based network for learning HOG.

3. Architecture of the customized learnable illumination prior

Deviated from the original MAE [4], we implement a UNet-like convolution neural network with a customized target feature in the masked image modeling strategy:

- employing the vanilla convolution operator to construct the encoder-decoder architecture;
- implementing diverse lighting conditions by augmenting the input image with gamma transformation;
- dividing the corresponding illumination maps (obtained by the max intensity of RGB channels) into regular non-overlapping regions, randomly sampling a subset of regions, and masking the remaining ones while maintaining the whole structure of the map;
- processing all the regions (both visible and masked regions) through the encoder and decoder to reconstruct the illumination map filtered by bilateral filter in pixels;
- to accentuate, the input of the network is the complete image structure rather than image patches.

Following the U-shaped network [6] designs, the obtained shallow embedding is passed through 3 encoder stages where each stage consists of a stack of a convolution layer with 3×3 kernel size and a ReLU activation layer. In the down-sampling layer, we down-sample the 2D feature maps using 3×3 convolution layer with stride 2. Similarly, in decoder stages, we employ the stack of a convolution layer with 3×3 kernel size and a ReLU activation layer for feature reconstruction over each stage. To assist the recovery process, each stage takes the high-level decoder features concatenated with the same-stage low-level encoder features via skip connections as input. The channel number of the three feature extractors is 16, 32, and 64, respectively. During training of the customized learnable illumination prior, the masked ratio is 75%. After pre-training in the strategy described in Figure 2 of the manuscript, the encoder embraces the powerful feature representation capability of the illumination properties. During the training paradigm of the proposed CUE, the feature of the third encoder stage will be incorporated into the design of $\mathrm{prox}_{\beta\rho_2}$ as described in Figure 3 of the manuscript.

4. Architecture of the customized learnable noise prior

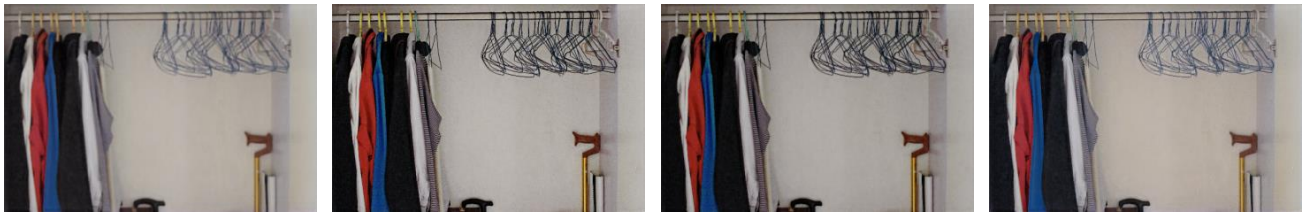
We customize a noise prior with the MAE, where the target feature is the HOG feature of normal-light images:

- obtaining the HOG feature of the whole image;
- dividing the image into regular non-overlapping patches, randomly sampling a subset of patches, and masking the remaining ones;
- flattening the histograms of masked patches and concatenating them into a 1-D vector as the target feature;
- the encoder and decoder are implemented by vanilla vision transformers as in MAE.

Following the MAE [4] designs, the encoder is a ViT [3] that is only applied to visible patches. Similar to a conventional ViT, our encoder embeds patches by a linear projection with positional embeddings and then processes the resulting set via a series of Transformer blocks. However, our encoder only operates on a small subset (25%) of the full set. The input to the lightweight MAE decoder is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens. Each mask token is a shared, learned vector that indicates the presence of a missing patch to be predicted. We add positional embeddings to all tokens in this full set; without this, mask tokens would have no information about their location in the image. The decoder has another series of Transformer blocks. After pre-training in the strategy described in Figure 5 of the manuscript, the encoder of the customized learnable noise prior possesses the powerful gradient representation capability. Thus, we redevelop it as a regularization term to suppress noise as described in the manuscript.

5. Ablation studies about the customized illumination and noise priors

We investigate the effectiveness of the proposed key components, i.e., f_{MAE_L} and f_{MAE_N}. The quantitative and qualitative evaluations are presented in Table 1 and Fig. 1. f_{MAE_L} denotes solving the L sub-problem with the encoder of the pre-trained illumination prior, and f_{MAE_N} denotes employing the gradient representation prior as the regularization. As shown in Fig. 1, removing f_{MAE_L} will cause the illumination map and the enhanced image to be over-smoothed. The enhanced image will suffer from severe noise pollution when discarding f_{MAE_N}. Combining them will achieve a notable performance improvement and a satisfying visual quality.



(a) wo f_{MAE_L} **(b) wo f_{MAE_N}** **(c) wo f_{MAE_L} and f_{MAE_N}** **(d) Ours**

Figure 1. Ablation studies of f_{MAE_L} and f_{MAE_N}.

Table 1. PSNR and SSIM scores of the Ablation studies for f_{MAE_L} and f_{MAE_N} on the LOL dataset.

f_{MAE_L}	f_{MAE_N}	PSNR↑	SSIM↑
		21.05	0.749
✓		21.48	0.752
	✓	21.54	0.758
✓	✓	**21.67**	**0.774**

6. Ablation studies about the implementation of the customized learnable noise prior

We implement an UNet-like customized learnable noise prior to investigate its the effectiveness. The default MAE-based architecture and the UNet-like customized learnable noise priors are denoted as f_{MAE_N} and $f_{MAE_N} - UNet$. The performance on the LOL dataset is presented in Table 2.

Table 2. Ablation studies about the implementation of the customized learnable noise prior on the LOL dataset.

Method	PSNR↑	SSIM↑	NIQE↓
f_{MAE_N}	21.67	0.774	3.776
$f_{MAE_N} - UNet$	21.54	0.768	3.941

7. Ablation studies about the unrolling stage on the Huawei Dataset

We perform ablation studies about K on the Huawei dataset in Table 2, and obtain the same optimal K (=3). The optimal K and performance trend on the Huawei dataset is consistent with that of LOL (see supplementary material). Thus, different datasets will not affect the optimal K.

Table 3. Ablation studies with k stage on Hauwei.

Stage (K)	1	2	3	4	5	6
PSNR	19.86	19.93	20.31	20.28	20.21	21.07
SSIM	0.626	0.637	0.658	0.651	0.644	0.638

8. Architecture of illumination adjustment network and reflectance restoration network

The architecture of the illumination adjustment network and reflectance restoration network is similar to the customized learnable illumination prior as described in 3. The channel of their encoder is 12, 24, and 48, respectively. The illumination adjustment network takes the decomposed illumination component, I_l, and an indicator map as the input. The reflectance restoration network takes the decomposed illumination, reflectance, and noise components, $[I_l, I_r, I_n]$ as the input.

9. Visualizations of the illumination map incorporated with the customized learnable illumination prior

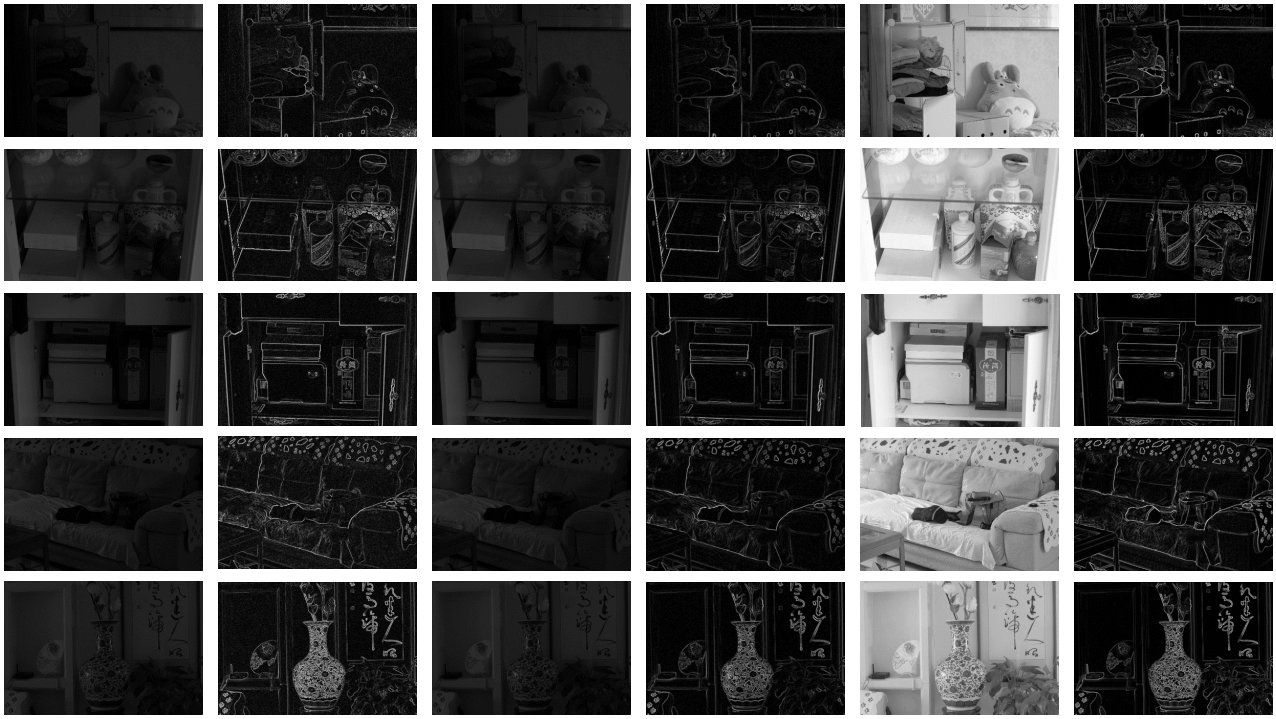
The MAE-based customized learnable illumination prior is trained from the normal-light image to its illumination map filtered by a bilateral filter, reducing noise without altering the intrinsic structure [1]. To verify the effectiveness of the customized learnable illumination prior, we presents the visualization of the illumination map of the low-light image, the illumination map incorporated with the customized learnable illumination prior, and the illumination map of the normal-light image, *i.e.*, $\mathbf{L_l}$, $\hat{\mathbf{L}}_\mathbf{l}$, and $\mathbf{L_n}$. Meanwhile, we utilize the Prewitt filter [5] to extract the gradient of the corresponding illumination map, which demonstrates that $\hat{\mathbf{L}}_\mathbf{l}$ incorporated with the learned illumination prior will reduce the noise while preserving the intrinsic structure.

10. More visualization results of decomposition and enhancement

Due to the limitation of the page length, in this section, we present more visualization results of decomposition and enhancement on LOL dataset in Figure 3 and Figure 4.

11. Broader impacts

Our work explores the potential of the customized learnable illumination and noise priors on low-light image enhancement and further propose a deep unfolding paradigm. Integrating the customized prior designs will improve the transparency and interpretability of neural networks and facilitate the development of AI in real-world applications. However, the efficacy of our method may raise potential concerns when it is improperly used. For example, the safety of the applications of our method in real-world applications may not be guaranteed. We will investigate the robustness and effectiveness of our method in broader real-world applications.



(a) L_l (b) Gradient of (a) (c) $\hat{L}_l$ (d) Gradient of (c) (e) L_n (f) Gradient of (e)

Figure 2. Visualization of the illumination map before and after incorporated with the customized learnable illumination prior. $\mathbf{L_l}$, $\hat{\mathbf{L}}_\mathbf{l}$, and $\mathbf{L_n}$ indicates the illumination map of the low-light image, the illumination map incorporated with the customized learnable illumination prior, and the illumination map of the normal-light image. The visualization demonstrates that $\hat{\mathbf{L}}_\mathbf{l}$ incorporated with the learned illumination prior will reduce the noise while preserving the intrinsic structure of $\mathbf{L_l}$. Please zoom in for better visualization.

References

[1] Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. *ACM Transactions on Graphics (TOG)*, 26(3):103–es, 2007. 1, 2, 4

[2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2, 3

[5] Raman Maini and Himanshu Aggarwal. Study and comparison of various image edge detection techniques. *International journal of image processing (IJIP)*, 3(1):1–11, 2009. 4

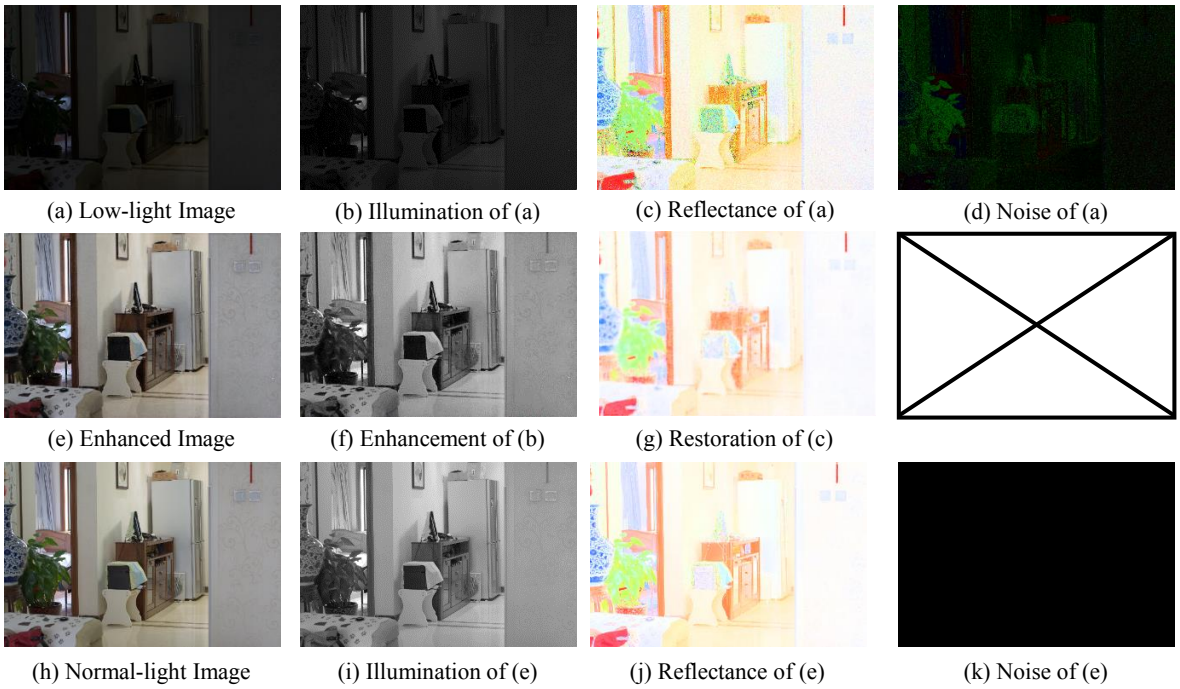
[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[7] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep Retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1

[8] Wenhui Wu, Jian Weng, Pingping Zhang, Xu Wang, Wenhan Yang, and Jianmin Jiang. Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, June 2022. 1

[9] 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. 1

[10] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the ACM International Conference on Multimedia*, pages 1632–1640, 2019. 1

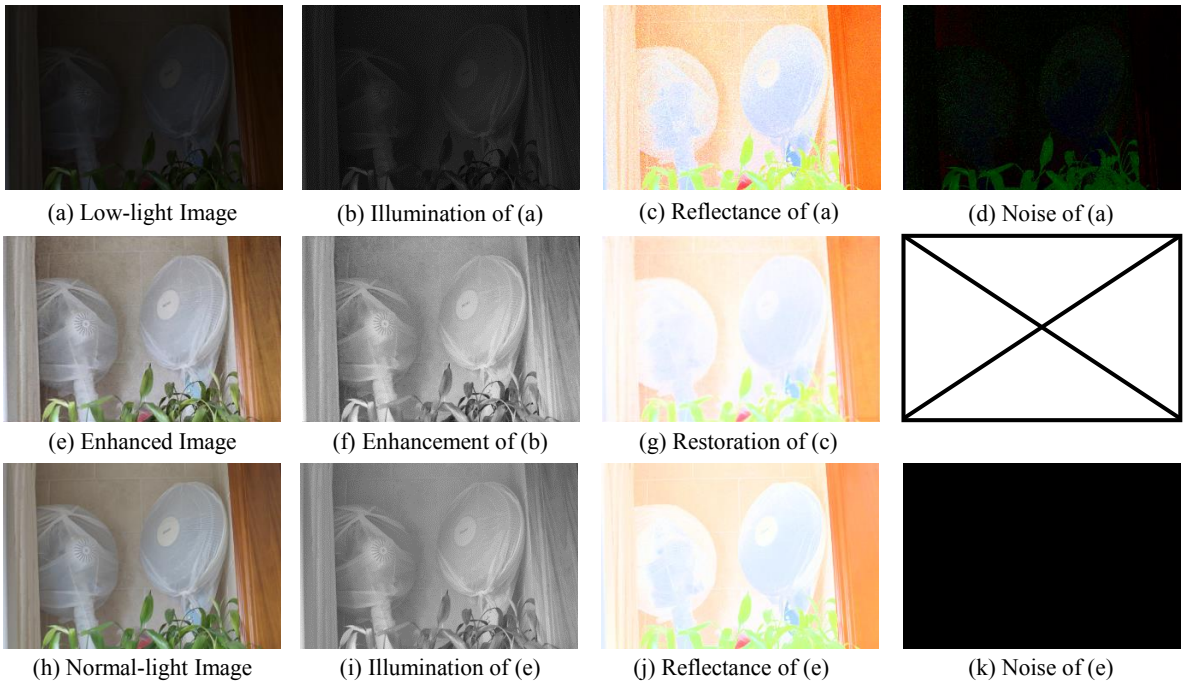


(a) Low-light Image (b) Illumination of (a) (c) Reflectance of (a) (d) Noise of (a)

(e) Enhanced Image (f) Enhancement of (b) (g) Restoration of (c)

(h) Normal-light Image (i) Illumination of (e) (j) Reflectance of (e) (k) Noise of (e)

Figure 3. Visualization results of decomposition and enhancement on the LOL dataset. Please zoom in for better visualization.



(a) Low-light Image (b) Illumination of (a) (c) Reflectance of (a) (d) Noise of (a)

(e) Enhanced Image (f) Enhancement of (b) (g) Restoration of (c)

(h) Normal-light Image (i) Illumination of (e) (j) Reflectance of (e) (k) Noise of (e)

Figure 4. Visualization results of decomposition and enhancement on the LOL dataset. Please zoom in for better visualization.