Supplementary Materials for: Learning Semantic-aware Knowledge Guidance for Low-Light Image Enhancement

Yuhui Wu[1], Chen Pan[1], Guoqing Wang[1]*, Yang Yang[1], Jiwei Wei[1], Chongyi Li[2], Heng Tao Shen[1]
[1]University of Electronic Science and Technology of China, China
[2]S-Lab, Nanyang Technological University, Singapore
wuyuhui132@gmail.com; panchen0103@163.com; gqwang0420@hotmail.com;
dlyyang@gmail.com; mathematic6@gmail.com; chongyi.li@ntu.edu.sg; shenhengtao@hotmail.com

1. Overview

In this document, we describe the architecture and training details of the proposed SKF in Sec. 2. We present additional visual comparisons with existing SOTA methods on real-world datasets, including LOL, LOL-v2, MEF, LIME, NPE and DICM, in Sec. 3. We provide additional ablation study in Sec. 4. Finally, we present limitations and future works in Sec. 6.

2. Additional Implementation Details

2.1. Pre-Trained Semantic Segmentation Network

The proposed SKF uses a pre-trained semantic segmentation network as the SKB to obtain semantic priors. We selected HRNetV2-W48 [6] due to the portability of encoder-decoder architecture and superior performance. The HRNetV2-W48 used as SKB is pre-trained on PASCAL-Context dataset [5] with the input size of 480×480. The PASCAL-Context dataset includes 4,998 scene images for training and 5,105 images for testing with 59 semantic labels and 1 background label. Furthermore, the weights of the HRNetV2-W48 are fixed during training stage to exploit the generative semantic priors.

Additionally, we find that the HRNetV2-W48 could produce under-optimized segmentation labels if the inputs are low-light images, which compromises the reliability of the prior and finally causes unexpected outputs. Hence, for each baseline method, we adopt specific data preprocessing to make the segmentation result more accurate, which is described in Sec. 2.2. The main body of HRNetV2-W48 contains four stages with four parallel convolution streams. The resolutions of the streams are 1/4, 1/8, 1/16, and 1/32, which generate four features with corresponding resolutions($F_0 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, $F_1 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$, $F_2 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 4C}$ and $F_3 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 8C}$, where $C = 48$) respectively. The representation head of HRNetV2 mixes the four features by upsampling and 1×1 convolutions and produce the prediction of size $S \in \mathbb{R}^{H\times W\times 59}$. Finally, we take four multi-scale features and the prediction as semantic priors to guide the enhancement process, the former are utilized to optimize the image feature by SE modules while the latter provides crucial guidance to SCH loss and SA loss.

*Corresponding author.

2.2. Details of Applying SKF to Baseline Methods

In this section, we provide more details of the implementation of each method with SKF, including the specific design of adding SE modules into baseline methods and the detailed experimental settings. The baseline methods are RetinexNet [8], KinD [13], DRBN [10], KinD++ [12], HWMNet [1], SNR-LLIE-Net [9] and LLFlow [7]. All the original hyperparameters and experimental settings are consistent with baseline methods. For RetinexNet-SKF, KinD-SKF and KinD++-SKF, we directly utilize the output of Decomposition Net as the refined image, like ISSR. The Decomposition Net outputs two components, which are reflectance map and illumination map. Based on Retinex theory, the former is close to the normal-light image, which can be utilized as a better input comparing to low-light image. Furthermore, we also process the input image by Decomposition Net in training stage of HWMNet-SKF, SNR-LLIE-Net-SKF and LLFlow-SKF to guarantee the accuracy of semantic map. For DRBN-SKF, we select output of the third recurrence as the refined image. We ensure the reliability of the semantic priors by providing refined image to SKB and enable the semantic-guided Enhancement Net to learn a proper map between low-light and normal-light image.

RetinexNet-SKF, KinD-SKF and KinD++-SKF. First, we group the RetinexNet, KinD and KinD++ together since they are Retinex-based methods with similar architectures, which have multiple subnets and adopt multi-stage training strategy. The subnets of these methods have to be trained in order, while only the last subnet use normal-light im-

age to supervise the parameter update. Thus we only apply SKF to the methods to train the last subnet, while the other are trained by the original settings. Additionally, they are Tensorflow-based methods and we reproduce them by PyTorch, which may cause the different results.

Considering the memory cost and feature compatibility, we reasonably design the RetinexNet-SKF, KinD-SKF and KinD++-SKF. For RetinexNet, we locate two SE modules before the second last and the last decoder layers and utilize F_0 and S respectively. For KinD, we locate three SE modules before the last three decoder layers and utilize F_0, F_1 and F_2 respectively. For KinD++, we locate only one SE module before the last decoder layer and utilize S.

We apply SC loss and SA loss to the stage of training the last subnet as well. Then we experimentally set the $\lambda_{sc} = 1$ and $\lambda_{sa} = 1$ as the standard setting of RetinexNet-SKF, KinD-SKF and KinD++-SKF.

DRBN-SKF. The DRBN is a deep recursive band network, which is trained via two stages. The results on LOL-v2 dataset is achieved by the first stage called recursive band learning. In this stage, the input image is enhanced by four cascade recursive subnets called recurrence. Each recurrence is encoder-decoder architecture with three various scales, which is suitable for locating SE modules. Then we apply our SKF to the last recurrence. The input of segmentation net is the output of the third recurrence. We locate three SE modules before every three decoder layers and utilize S' (prediction before $Softmax$), F_0 and F_1 respectively. As for training settings, we set $\lambda_{sc} = 1$ and $\lambda_{sa} = 1$.

SNR-LLIE-Net. SNR-LLIE-Net can be divided into three parts: encoder, SNR-guided hidden layers, decoder. To maintain the original contribution of the method, we locate SE modules in the decoder to optimize the multi-scale features. Specifically, we locate three SE modules before the last three decoder layers and utilize F_0, F_1 and F_2 as semantic priors respectively. As for training settings, we first manipulate the SC loss and SA loss to the similar scale of the original loss and set $\lambda_{sc} = 0.1$ and $\lambda_{sa} = 0.1$.

HWMNet-SKF. The HWMNet is designed based on M-Net [4], which is a UNet-like network architecture. The encoder and decoder of HWMNet both have four various scales. Then we locate three SE modules before the last three decoder layers and utilize F_0, F_1 and F_2 respectively. As for training settings, we set $\lambda_{sc} = 0.1$ and $\lambda_{sa} = 0.1$.

LLFlow-S-SKF and LLFlow-L-SKF. LLFlow consists of a conditional encoder to extract the illumination invariant color map and an invertible network that learns a distribution of normally exposed images conditioned on a low-light one. The conditional encoder has three output features with different scales, which are inputs of the invertible network. We locate the SE modules between the conditional encoder and the invertible network. The SE modules utilize F_0, F_1 and F_2 to enhance the three features output by conditional network. The LLFlow is trained as a normal flow, which outputs a invariant color map instead of a normal-light image. Hence we do not apply SA loss to LLFlow, while the SC loss is still available. The SC loss is reasonably applied to the color map, which also includes color information. We set $\lambda_{sc} = 1$.

More details are shown in the code.

2.3. Additional Details of the Results

In Table. 1 in main paper, we train the methods on LOL and LOL-v2 datasets to obtain the results on each dataset respectively. Note that we obtain these numbers of baseline methods and other SOTA methods either from the respective papers or by running the respective public code. The results on MEF, NPE, LIME and DICM in Table. 2 in main paper are also obtained by methods all trained on LOL dataset in order to fair comparison. Finally, we conduct all the ablation studies in Table. 3 to 5 in main paper on LOL dataset as well.

3. More Visual Comparisons

As shown in Figs. 1 to 4 in this supplementary materials, we give more visual results of methods with our SKF and baseline methods on LOL/LOL-v2, MEF, NPE, LIME and DICM datasets as the supplement of the visualization in the main paper. We can see that methods with our SKF consistently produce more natural results and achieve superior performance over the baseline methods in various scenes.

4. More Ablation Analysis

As illustrates in Sec. 2.1, the HRNetV2-W48 has four output features available to be semantic priors. For the experiments on benchmark datasets, as shown in main paper, each baseline method utilizes specific number of SE modules to deal with the same number of semantic features. Hence, we conduct experiments to investigate the varying number of SE modules. We choose HWMNet because it has a comprehensive UNet-like architecture with four different scales in decoder and the end-to-end training strategy. As shown in Table. 1, we select number of modules from 1 to 4. SE-1 denotes that only one module is located before the last layer and SE-4 denotes four modules are located before all the four layers respectively. The results illustrate that stacking more SE modules can achieve better performance but consume more memory. Thus, we utilize three SE modules in HWMNet-SKF and design other methods with SKF in a similar way.

In Fig. 6 and Table. 3 in main paper, we report quantitative results and qualitative results respectively to investigating the contribution of every three key components of our SKF, *i.e.*, SC loss, SA loss and SE module. More qualitative results are shown in Fig. 5 to further complement the

Table 1. Ablation study of HWMNet-SKF for investigating the number and different inputs of SE modules. I+S denotes the semantic-aware manner in this paper, I+I denotes two inputs of SE module are both image features.

Arch	PSNR ↑	SSIM ↑	LPIPS ↓	Param ↓
SE-0	24.240	0.852	0.114	66.56
SE-1	24.792	0.854	0.111	67.11
SE-2	24.891	0.857	0.109	67.82
SE-3	25.123	**0.860**	0.108	68.77
SE-4	**25.155**	**0.860**	**0.105**	70.35

ablation study of our SKF.

5. More Computational Efficiency Analysis

In this section, we provide more computational efficiency analysis for evaluating practicality. The computational cost of each model is show in Table. 2. We evaluate the cost using one image with a size 400×600.

Table 2. Computational Efficiency Analysis.

Method	RetinexNet	**+SKF**	KinD	**+SKF**
GFlops(G)	129.33	157.44	356.72	360.41
Param(M)	0.62	0.71	8.03	8.50
Method	DRBN	**+SKF**	KinD++	**+SKF**
GFlops(G)	38.79	41.37	335.98	344.27
Param(M)	2.21	2.37	9.63	10.21
Method	HWMNet	**+SKF**	SNR-Net	**+SKF**
GFlops(G)	929.93	1074.89	87.26	115.97
Param(M)	66.56	68.77	39.13	39.44
Method	LLFlow-S	**+SKF**	LLFlow-L	**+SKF**
GFlops(G)	128.20	136.14	1048.60	1107.79
Param(M)	4.97	5.26	37.68	38.21

6. Limitations and Future Works

In this section, we discuss the limitations of our work and suggest the potential future research directions of semantic-aware low-light enhancement.

Limitations. First, while our SKF improves the enhancement capability of baseline methods significantly, the entire framework is heavily reliant on the quality of semantic priors provided by SKB. We use the reflectance map as the segmentation net's input, which may result in misclassification due to the divergence between the reflectance map and the normal-light image. Furthermore, we chose HRNet pre-trained on the PASCAL-Context dataset as our SKB because the PASCAL-Context dataset primarily contains indoor scenes, which are also common in LOL/LOL-v2 datasets. However, there is still a gap between PASCAL-Context and LOL/LOL-v2 datasets. Hence, our SKF can generalize across various methods, but the generalizability across various datasets may be limited by the SKB, to be specific, by the scale of semantic segmentation dataset and the performance of semantic segmentation network.

Second, the proposed components of our SKF (*i.e.*, SE module, SCH loss and SA loss) are preliminary techniques for introducing semantic priors, which may limit the potential of our SKF. We design the semantic-aware embedding (SE) module inspired by some well-designed attention blocks [2, 3, 11] instead of specifically designing delicate module for cross-modal interaction between semantic priors and images. The simple interaction manner of our SE module may undermine the fusion process and thus lead to unsatisfactory features. Furthermore, the SCH loss and SA loss simply utilize the semantic maps to obtain regional information and still have potential to be optimized.

Third, we only apply the idea of semantic guidance to LLIE task due to the motivation of optimize the color and details of enhanced image. Actually, similar tasks in the area of low-level vision are suitable to introducing semantic priors as well.

Future works. According to the limitations of the SKF, our future works can be organized as follows. We will explore the improvement of SKB in restoring more semantic priors and providing correct priors when meeting unknown instance. Furthermore, we will investigate whether the SKB could learn from low-light datasets during training stage to avoid the distribution gap. Then, we plan to design cross-modal interaction modules specifically for embedding semantic features and semantic-guided losses to utilize priors in a more reasonable manner. Finally, another valuable direction is to explore the potential of establishing semantic-guided framework, *e.g.*, our SKF, in other low-level vision tasks.

References

[1] Chi-Mao Fan, Tsung-Jung Liu, and Kuan-Hsien Liu. Half wavelet attention on M-Net+ for low-light image enhancement. *arXiv preprint arXiv:2203.01296*, 2022. 1

[2] Hyunyoung Jung, Eunhyeok Park, and Sungjoo Yoo. Fine-grained semantics-aware representation enhancement for self-supervised monocular depth estimation. In *ICCV*, pages 12642–12652, 2021. 3

[3] Huajun Liu, Fuqiang Liu, Xinyi Fan, and Dong Huang. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv preprint arXiv:2107.00782*, 2021. 3

[4] Raghav Mehta and Jayanthi Sivaswamy. M-net: A convolutional neural network for deep brain structure segmentation. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 437–440. IEEE, 2017. 2

[5] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and se-

mantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 1

[6] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 43:3349–3364, 2020. 1

[7] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *AAAI*, volume 36, pages 2604–2612, 2022. 1

[8] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 1

[9] Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. SNR-aware low-light image enhancement. In *CVPR*, pages 17714–17724, 2022. 1

[10] Wenhan Yang, Shiqi Wang, Yuming Fang, Yue Wang, and Jiaying Liu. From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In *CVPR*, pages 3063–3072, 2020. 1

[11] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 3

[12] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *IJCV*, 129:1013–1037, 2021. 1

[13] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACMMM*, pages 1632–1640, 2019. 1

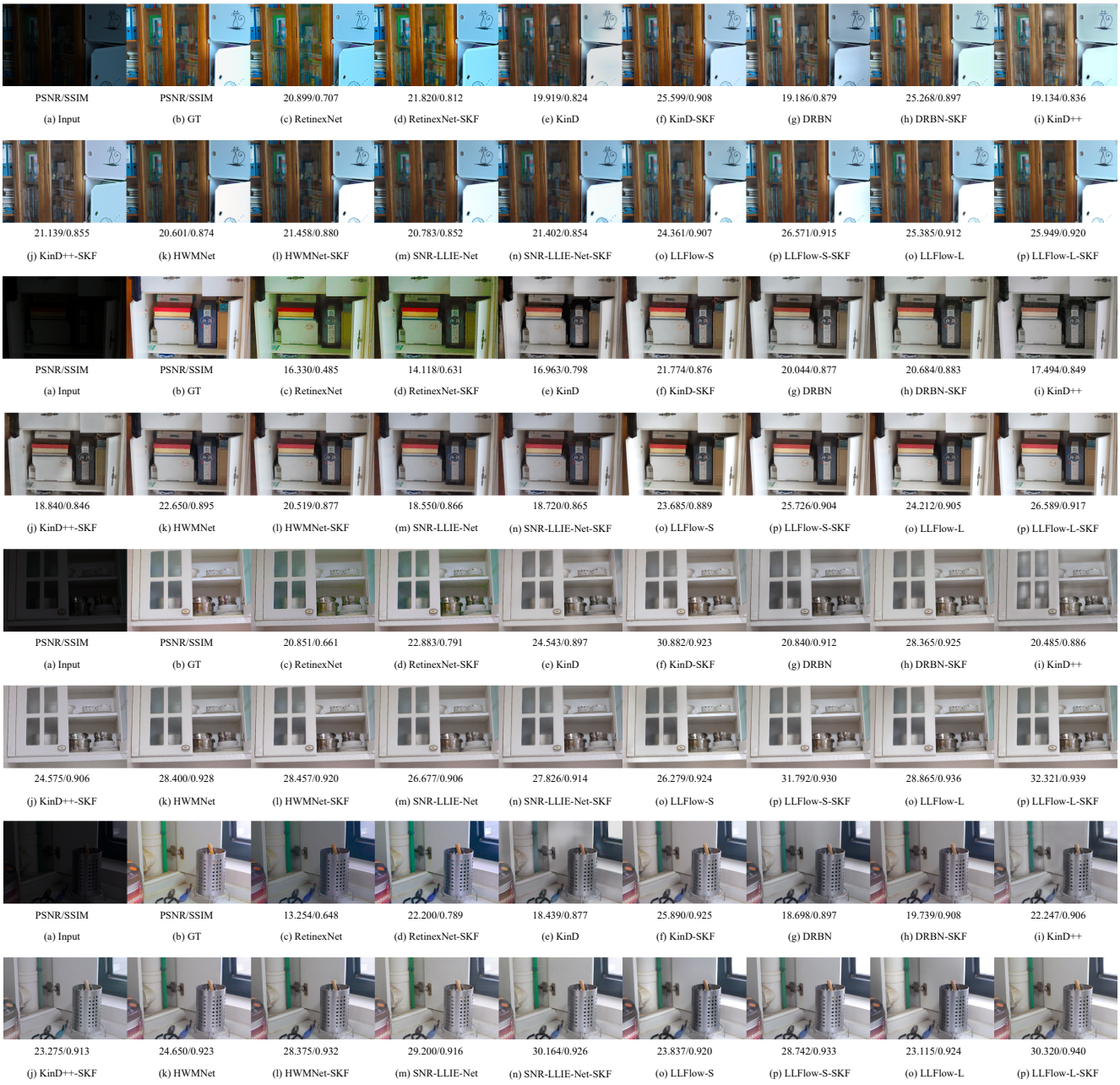

Figure 1. Visual comparison of baseline methods with and without SKF on LOL/LOL-v2 dataset.

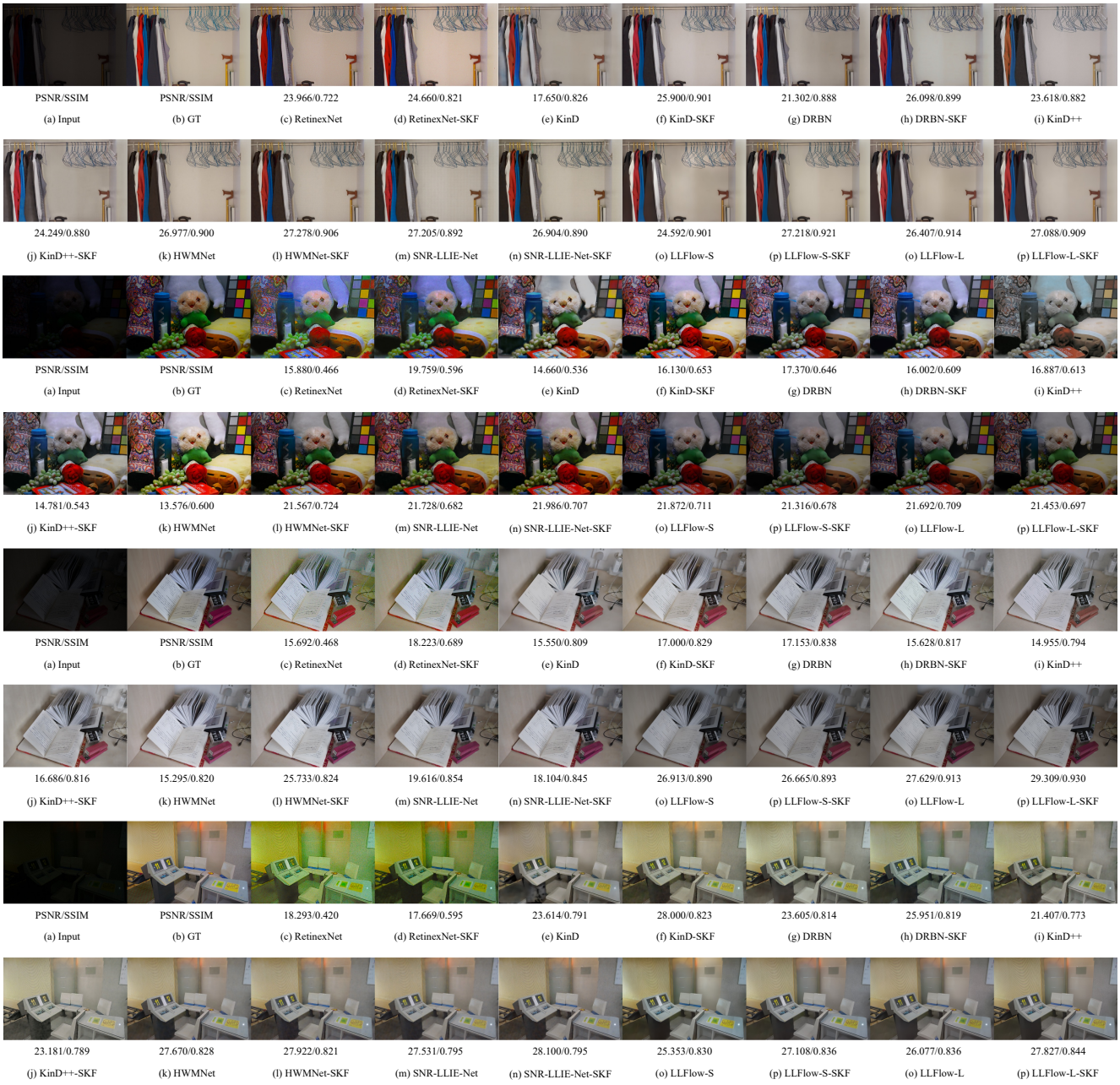

Figure 2. Visual comparison of baseline methods with and without SKF on LOL/LOL-v2 dataset.

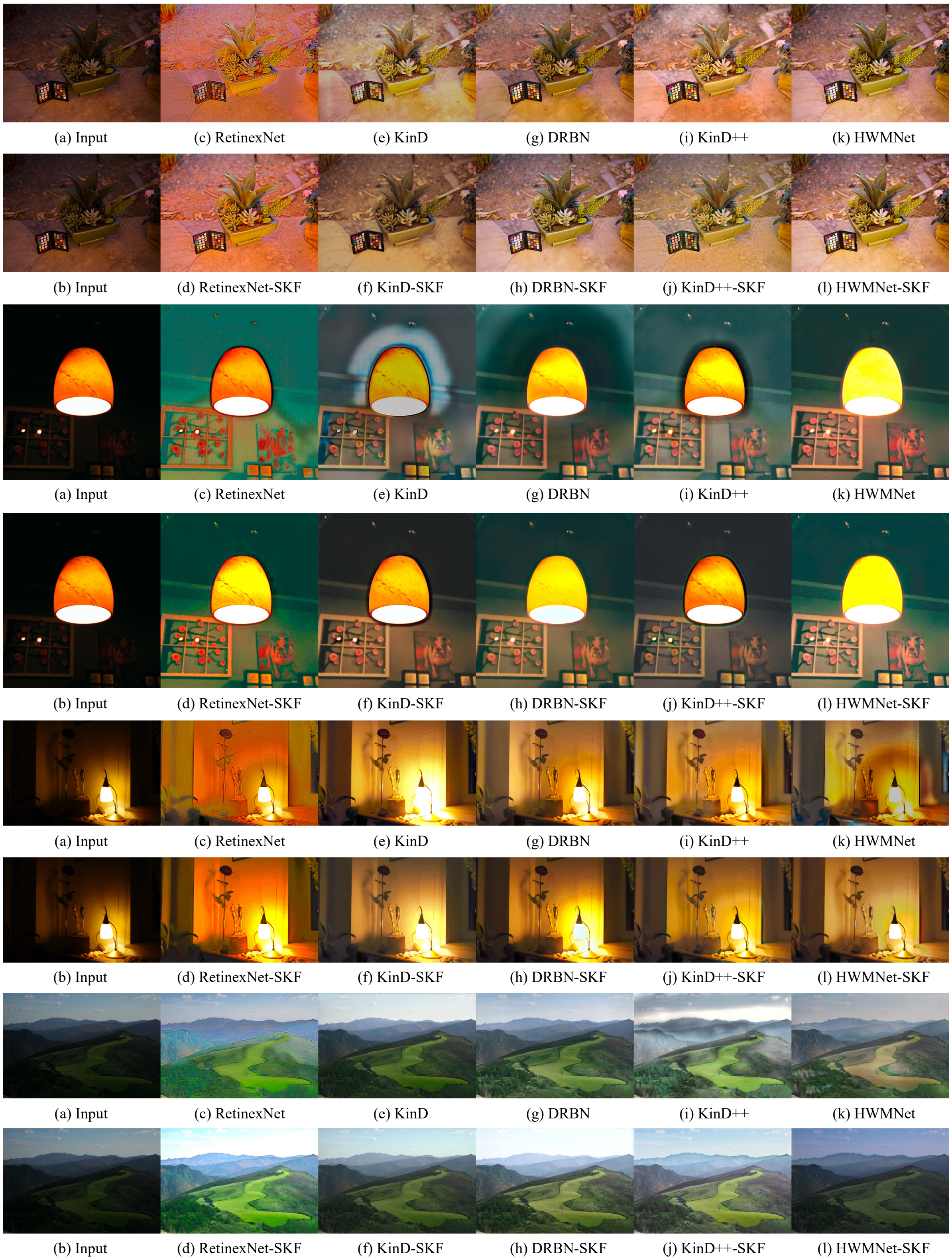

Figure 3. Visual comparison of baseline methods with and without SKF on LIME/MEF dataset.

(a) Input (c) RetinexNet (e) KinD (g) DRBN (i) KinD++ (k) HWMNet

(b) Input (d) RetinexNet-SKF (f) KinD-SKF (h) DRBN-SKF (j) KinD++-SKF (l) HWMNet-SKF

(a) Input (c) RetinexNet (e) KinD (g) DRBN (i) KinD++ (k) HWMNet

(b) Input (d) RetinexNet-SKF (f) KinD-SKF (h) DRBN-SKF (j) KinD++-SKF (l) HWMNet-SKF

(a) Input (c) RetinexNet (e) KinD (g) DRBN (i) KinD++ (k) HWMNet

(b) Input (d) RetinexNet-SKF (f) KinD-SKF (h) DRBN-SKF (j) KinD++-SKF (l) HWMNet-SKF

(a) Input (c) RetinexNet (e) KinD (g) DRBN (i) KinD++ (k) HWMNet

(b) Input (d) RetinexNet-SKF (f) KinD-SKF (h) DRBN-SKF (j) KinD++-SKF (l) HWMNet-SKF

Figure 4. Visual comparison of baseline methods with and without SKF on NPE/DICM dataset.

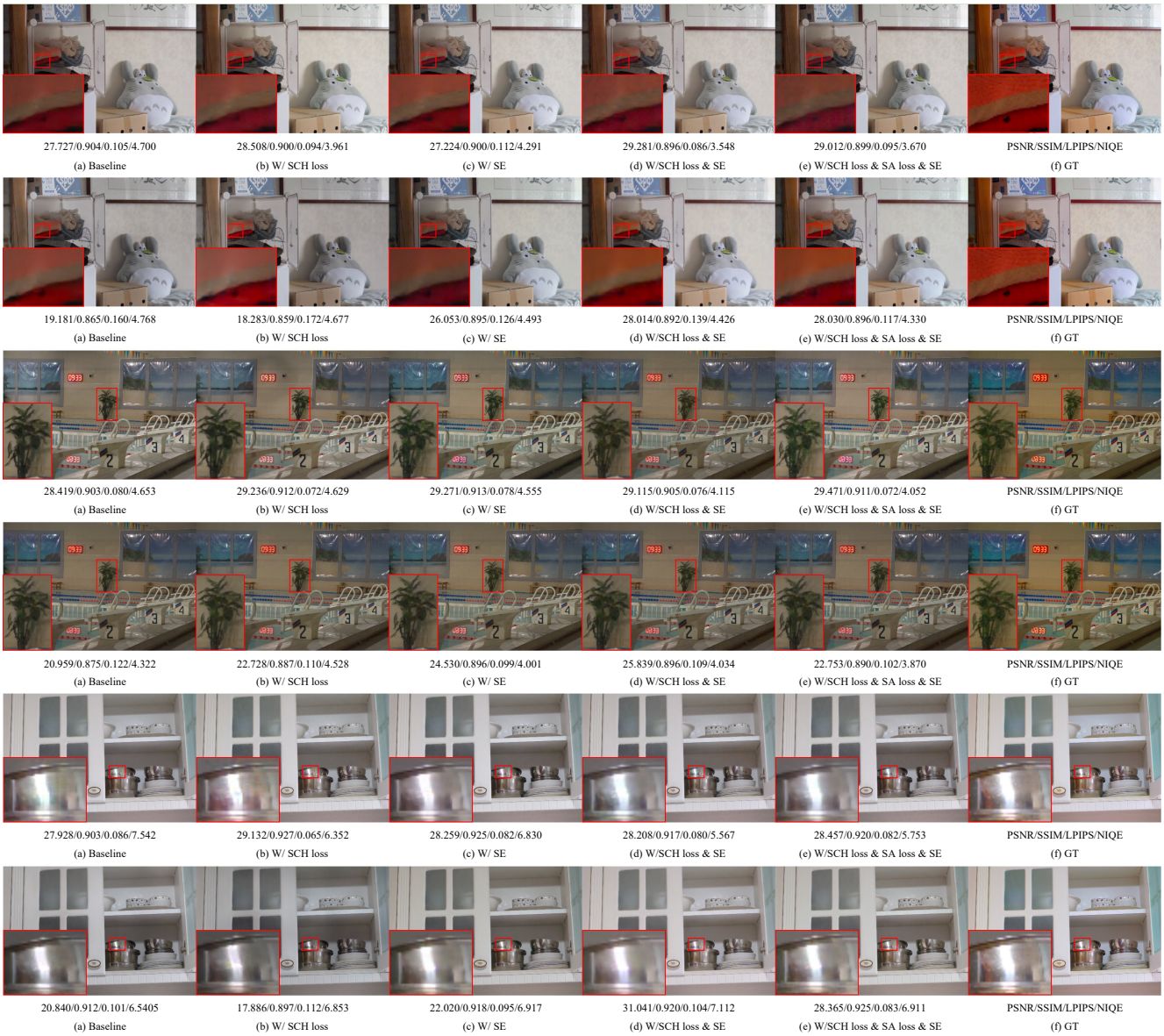

Figure 5. Ablation study of DRBN-SKF and HWMNet-SKF. Baseline, baseline with SC loss, baseline with SE module, baseline with SC loss and SE module, baseline with SC loss, SA loss and SE module, groundtruth are shown from left to right. For each image, top and bottom rows are results of HWMNet-SKF and DRBN-SKF respectively.