SELF-SUPERVISED LEARNING ON A LIGHTWEIGHT LOW-LIGHT IMAGE ENHANCEMENT MODEL WITH CURVE REFINEMENT

Wanyu Wu[1], *Wei Wang*[1,*], *Kui Jiang*[2], *Xin Xu*[1], *Ruimin Hu*[2]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology
[2]School of Computer Science, Wuhan University

ABSTRACT

Deep learning networks with deeper layers become a trend for their good performance but lacks the potential for real-time mobile deployment. Another challenge for paired training networks is the limited generalization capacity caused by the sample bias. To overcome these two challenges, we propose a lightweight self-supervised low-light image enhancement method, that trains with low light images only. Specifically, our method consists of a low-resolution dense CNN network stream and a full-resolution guidance stream, responsible for image-to-curve transformation with refinement and spatial guidance fusion, respectively. Then, a new self-supervised loss function is introduced to measure the restored patch-based color deviations among color channels. Experimental results show that our method gives competitive performance to the full-supervised approaches.

Index Terms— lightweight self-supervised network, real-time low-light image enhancement, image-to-curve transformation

1. INTRODUCTION

The unpleasing lighting conditions, such as low lighting conditions, set up a barrier for capturing high-quality images/videos, further influence the performance of high-level vision tasks, e.g. object detection and person recognition. Therefore, low-light image enhancement (LLIE) has become a prominent topic with increasing attention.

Existing LLIE methods can be commonly divided into three categories: full-supervision, semi-supervision, and self-supervision. Since there is a strong constraint of paired samples, full-supervised methods have become the mainstream for LLIE in recent years for their impressive results. For example, the work [1, 2] introduced full-supervised methods that directly learn the translations from the input to the corresponding output image. Recently, several methods utilize the retinex theory to achieve better performance, since the enhancement of the image components is executed on separate sub-networks introduced in [3, 4]. The derivative works include adding noise decomposition module by RRDNet [5] or multi-scale luminance attention module [6]. In [7–9], different regularizers were designed for refining the direct reflections component and obtain good performance. For varying luminance conditions, methods in [10, 11] attempted to build edge-enhancement module and multi-path fusion for exposure fusion. However, current full-supervised methods mainly rely on synthetically paired images, thus lack generalization for real scenes.

To release the high reliance on synthetic data pairs, EnlightenGAN [12] firstly introduced the unpaired training on LLIE with a global-local discriminator, and some semi-supervised methods were also introduced. Lv *et al.* [13] fused different enhanced intermediate results to generate exposure correction results. Yang *et al.* [14] modeled the linear band representation instead of the restored image in diverse luminance. However, these methods adopt the full-size image transform, which is computationally expensive in pixelated reconstruction. Considering further generalization improvement and decreasing the influence from sample bias, self-supervised methods were proposed. For example, Zhang *et al.* [15] proposed a new Retinex model, that matches the maximum channel of reflectance to that of the input, but fails in the low light cases with severe information distortion. Recently, self-supervised image-to-curve models [16–18] gained widespread attention. Zhang *et al.* [16] assembled a specific "S" curve model like gamma correction, while [17, 18] estimated a image-specific curve and the piece-wise version for nonlinear luminance adjustment, respectively.

However, current self-supervised methods still face several difficulties: 1) Most of the methods process full-size images, thus cannot achieve a fast inference speed. 2) Even the semi or self-supervised methods do not require the ground truth, normal exposure samples are still essential for training and restrict the model generalization. To train with low-light images only and accelerate computational speed, we propose a lightweight image-to-curve transformation method. The main contributions of our work are threefold:

- Lightweight self-supervised network for real-time low-light image enhancement.
- DenseNet based image-to-curve transformation with low-light inputs only.

This work was supported by the Natural Science Foundation of China (U1803262).

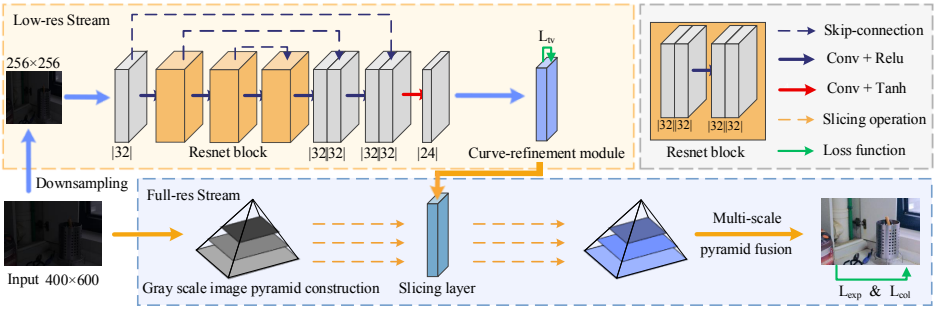

Fig. 1. The architecture of the proposed method. The low-resolution stream transforms the down-sampled input into two bilateral grid volumes of the mapping curves. The full resolution stream uses the input to construct a grayscale Gaussian pyramid as the latter guidance map in the slicing layer, then to perform multi-scale fusion to obtain the full-resolution output.

- New self-supervised loss function on luminance-color deviation measurement.

2. PROPOSED METHOD

In this paper, a new self-supervised convolutional network-based method is proposed and trains with low light images only, whose architecture is shown in Fig. 1.

2.1. Why image-to-curve formulation

Image-to-curve mapping is an effective LLIE solution, when lacking of paired data in diverse real-world lighting condition. Since the low-light image is regarded as an inverse version of the hazy image [19], the radiance scattering model is also applicable to the low-light condition, in terms of a virtual dark surface to indicate the dark transmission map. Modified from the direct scattering in dehazing model [20, 21] with atmospheric extinction coefficient α, the direct reflection I_d of pixel X at coordinates (x, y) in the low-light condition is:

$$I_d(X) = J(X)e^{-\alpha(\tau D(x)+D(y))} \tag{1}$$

where $D(x)$ is the distance, denoting the light line travels from a light source to the object, and $D(y)$ is the distance of the reflectance from the object to the camera. The $\tau \in [0, 1]$ refers to the dark attenuation coefficient to represent the virtual dark surface. The scatted light I_b is defined with the scattered distance $D(ry)$,in which r is the partition ratio of $D(y)$ to form the scatter angle θ:

$$I_b(X) = k \int_0^1 \frac{e^{-a[\tau[D(ry)+1]}}{[D(ry)]^2} dr \tag{2}$$

where the term k is an effective constant for the absorbed radiance intensity. Especially, it is not a closed-form solution for equation (2) in integral but it is a smooth function with a numerical solution independent of physical parameters.

Since $I(x, y) = I_d(x, y) + I_b(x, y)$, the prediction of the restored image $J(x, y)$ is formed in the log domain:

$$J_c(\hat{x}, y) = \omega_c \sum_c F(\hat{x}, y) + b_c \tag{3}$$

here the $F(x, y) = \int_0^1 \frac{e^{-a[\tau[D(ry)+1]}}{[D(ry)]^2} dr$ is the integration function of the pixel X. This equation (3) represents the restored image $J_c(\hat{x}, y)$ is a linear prediction of $F_c(\hat{x}, y)$ in color channels c, which is a curve estimation function with pixel wise fusion. Therefore, this prediction is feasible for curve mapping in the low resolution stream of our method.

2.2. Low-resolution stream for Image-to-curve

The low-resolution stream is a convolutional network to scale the image-to-curve mapping function $F_c(x, y)$, whose operators contain limited intensity variations in a local patch, in order to separate the intensities from positions. Inspired by the polynomial function in [5], $F_c(x, y)$ is formed by iteratively applying the curve estimation for n times:

$$F_c^n(x, y) = F_c^{n-1}(x, y) + \omega_F F_c^{n-1}(x, y)(1 - F_c^{n-1}(x, y)) \tag{4}$$

in which ω_F is the curve parameter matrix.

The low-resolution stream is responsible for image-to-curve mapping by a convolutional network, consisting of three residual blocks and four convolutional layers, with symmetric convolutional layers or residual blocks connected in a hopping layer. Each convolutional layer consists of 32 convolutional kernels of size 3×3 and steps size 1, and each residual block is also composed of two convolutional layers. The last output layer uses the tanh function, while the remaining ones use ReLU as the activation function. As such, the curve parameter map of 24 channels is processed into eight corresponding reference curves in the penultimate layer, which is followed by a curve-refinement module.

In the curve-refinement module, the 24 mapping matrices are reshaped and concatenated into two 12 channels mapping matrices, which handles the mapping to the resultant image in the dynamic range of $[0, 0.5)$ and $[0.5, 1]$, respectively. These two grid cells form a continuous mapping function with a wider range than single curve formation approaches. Using the bilateral grid in [22], the curve-refined output has been unrolled with a $16 \times 16 \times 8$ bilateral grid with 2 gird cells, and each contains 12 digits.

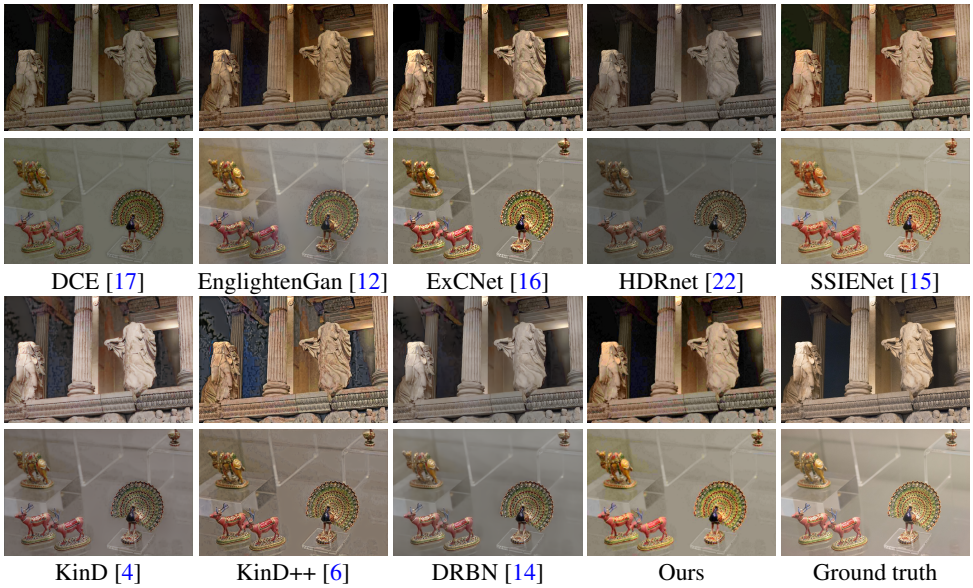

Fig. 2. Results of the proposed method and existing SOTA methods in LLIE. Our method trains with the low-light images only.

2.3. Full-resolution stream for multi-scale reconstruction

The full-resolution stream has three procedures, building a full-resolution gray-scale pyramid, slicing the bilateral grid per pyramid level, and multi-scale fusion for optimal curve estimation.

In the slicing layer similar to [22], the pyramid levels act as the $2D$ guide maps. Then, a data-dependent lookup matching is executed in the bilateral grid with non-parameter trilinearly interpolation. Especially, our two grid cell is transformed into two $3D$ full-resolution volumes, in which range falloff is not affected by small range variations and stays discontinuous. This procedure ensures our locally smooth curve prediction. Then, the pyramid's first layer is up-sampled and superimposed with the second one at weights of 0.8 and 1.2. This step is repeated again to obtain the final output.

Notably, our model is trained with low-light images only. Owing to the difficulties in measuring the enhancement extent of the input image, we propose a self-supervised color loss function introduced in the next subsection.

2.4. Loss function

In our network, the total loss function is a weighted sum of three self-regularized loss functions with the corresponding weights, and we empirically set the three weights W_{col}, W_{exp} and W_{tv} to 10, 30 and 200, respectively.

$$L_{total} = w_{col}L_{col} + w_{exp}L_{exp} + w_{tv}L_{tv} \tag{5}$$

Illuminance-color deviation loss Aiming to effectively suppress the generation of color bias, we propose a new exposure-guided color loss function L_{col} with the prior knowledge that the direction of image luminance vector variation is similar to that of color vector, thus our loss function measuring the cosine similarity between the color tensor and the luminance tensor in the local area, which can be expressed by the specific formula:

$$L_{col} = \frac{1}{M}\sum_{k=1}^{M} \frac{\sum_{i=1}^{n}(I_i \cdot C_i)}{\sqrt{\sum_{i=1}^{n}(I_i)^2}\cdot\sqrt{\sum_{i=1}^{n}(C_i)^2}} \tag{6}$$

where M represents the number of non-overlapping local regions of size 16×16, n is the dimension of the tensor, and I and C are the luminance tensor and color tensor with the shape of $[1, h, w]$, respectively. The color tensor is obtained by averaging the intensities of the three RGB channels, while the luminance tensor intensities are the grayscale values of the input image.

Exposure control loss follows the definition in L_{exp} [17], and adjusts the exposure level according to the local distance between the average intensity and a good exposure level E:

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M} |Y_k - E| \tag{7}$$

Illumination smoothness loss L_{tv} preserves a monotonic relationship among adjacent pixels by controlling the smoothing on curve parameter matrices A:

$$L_{tv} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c}\left(|\nabla_x A_n^c| + |\nabla_y A_n^c|\right)^2, c=\{R, G, B\} \tag{8}$$

3. EXPERIMENT

We build our network on Pytorch and train for 999 epochs with a mini-batch size of 6 on an NVidia TITAN Xp GPU. The Adam optimizer is used with an initial learning rate of $1e^{-3}$ and $5e^{-4}$ after 500 epochs.

3.1. Visual comparisons

The visual comparison is carried out on the SICE [23] and LOL-V1 dataset [3], to evaluate the performance of the proposed method with 8 SOTA methods. Two challenging cases of lighting and transparent object are shown in Fig. 2. In

Table 1. Objective evaluations on the LOL-V1 dataset. Red and blue are the best and second best results respectively.

Datasets		Method	SSIM↑	PSNR↑	FSIM↑	GMSD↓	NIQE↓	Params (M)↓	Times(s)↓	GFLOPs(G)↓(256 × 256)
Paired	240 synthetic and 460 pairs in LOL-V1	KinD++ [6]	0.9013	24.3569	0.9224	0.0960	3.6504	8.275	0.392	371.27
	689 image pairs in LOL-V2	DRBN [14]	0.9079	21.0993	0.9526	0.0770	4.1718	0.577	2.561	28.47
	485 image pairs in LOL-V1	KinD [4]	0.7598	19.8388	0.8209	0.1592	3.3723	8.16	0.059	23.24
	485 image pairs in LOL-V1	HDRNet [22]	0.7349	16.6052	0.8985	0.1166	4.7806	0.482	0.008	0.05
Multi-exposure	914 low-light and 1016 normal-light images	EnlightenGAN [12]	0.7756	19.7530	0.9121	0.1038	2.9434	8.637	0.057	16.58
	3022 multi-exposure images in SICE Part1	Zero-DCE [17]	0.6552	16.9637	0.8556	0.1661	4.4446	0.079	0.010	5.21
Low light only	485 low-light images in LOL-V1	SSIENet [15]	0.8201	22.5313	0.9004	0.1145	2.9454	0.682	0.124	29.46
	485 low-light images in LOL-V1	OURS	0.8102	20.8577	0.8983	0.1244	3.7158	0.094	0.003	5.95

Input (9.42/0.40) | w/o Low-resolution (18.14/0.79) | w/o Full-resolution (13.23/0.64) | w/o L_{col} (19.42/0.80) | w/o L_{exp} (8.00/0.23) | w/o L_{tv} (20.28/0.80) | Ours (20.86/0.81)

Fig. 3. Ablation study of different modules. PSNR/SSIM measurements are given in parentheses.

Table 2. Comparison for single low-light image training.

Method	SSIM↑	PSNR↑	FSIM↑	GMSD↓	NIQE↓
ExCNet [16]	0.6671	17.9791	0.8698	0.1350	4.5777
SSIENet*	0.7183	17.9292	0.7382	0.2336	4.4220
OURS*	0.8032	19.9585	0.8719	0.1287	3.7280

case 1, artifacts in dark shadows on the left are observed in the results from SSIENET [15], KinD [4] and KinD++ [6], caused by the light radiating from the small light source at the right corner. In case 2, the color of the peacock craft is dim im results from Zero-DCE [17], EnglightenGan [12], HDR-net [22], KinD [4], KinD++ [6] and DRBN [14]. On the contrary, the performance of our method is robust in the near-field light and preserves vivid color that closes to the corresponding ground truth.

3.2. Objective evaluations

We compare the proposed method with SOTA methods via evaluations on LOL-V1 dataset [3] in Table 1 to 2, in terms of PSNR, SSIM, GMSD [24], FSIM [25], NIQE [26] and parameter size. Considering the best performance from the fully-supervised methods [14] on SSIM and FSIM evaluations, the self-supervised methods(SSIENet [15] and our approach) use nearly 35 percent of unpaired training data in [14] to achieve less than 9 percent performance degradation. In terms of PSNR, GMSD and NIQE indices, the best performance from fully-supervised and self-supervised approaches is quite close, with less than 2 percent difference. In all, our approach has the shortest running-time and ranks second in parameter size, while ranks fourth after paired-training approaches for PSNR, SSIM evaluations. In Table.2 of training with only one randomly selected low-light image from LOL-V1, our performance ranks first on all evaluation metrics, with the smallest parameter size. Besides, our method requires only 14 percent parameters and 2 percent running-time of SSIENet [15] to achieve more than 90 percent of its performance.

3.3. Ablation study

We adopt ablation studies on the effectiveness of two streams and loss functions of our model, shown in Fig. 3.

Low-resolution stream Replacing our low-resolution stream with a few simple convolutional layers, the results will exist with obvious artifacts, indicating the stability and effectiveness of our low-resolution stream.

Full-resolution stream Discarding the multi-scale pyramid in the full resolution stream would result in unstable luminance, since the fusion on multi-scale feature maps attain better guidance in pyramid levels.

Loss functions Discarding the proposed color loss L_{col} leads to a significant color bias, indicating that the color loss function gives specification on the interaction among three color channels. Removing the exposure loss L_{exp} fails to restore dark regions. Finally, eliminating the loss of illumination smoothness L_{tv} leads to insufficient recovery on edges.

4. CONCLUSION

In this paper, we propose a lightweight self-supervised model for real-time low-light image enhancement, with low light image inputs only. In our model, the low-resolution stream executes the image-to-curve transform with embedded curve refinement. Then, the full-resolution stream slices the transformation curves in $3D$ volumes with the guided grayscale pyramid levels to produce the final restored image via multi-scale fusion. In summary, our method is applicable to a real-time mobile deployment, which is verified by the leading experiment performance on parameter size, running-time. Besides, our approach achieves comparable performance to the full-supervised methods in image enhancement evaluations. Our future research will focus on applying this real-time method on extremely low light images or raw images, such that the gap between the image enhancement and the target tracking task can be gradually eliminated.

5. REFERENCES

[1] Li-Wen Wang, Zhi-Song Liu, Wan-Chi Siu, et al., "Lightening network for low-light image enhancement," *TIP*, vol. 29, pp. 7984–7996, 2020.

[2] Jiaqian Li, Juncheng Li, Faming Fang, et al., "Luminance-aware pyramid network for low-light image enhancement," *TMM*, vol. PP, no. 99, pp. 1–1, 2020.

[3] Chen Wei, Wenjing Wang, Wenhan Yang, et al., "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[4] Yonghua Zhang, Jiawan Zhang, Xiaojie Guo, "Kindling the darkness: A practical low-light image enhancer," in *ACMMM*, 2019, pp. 1632–1640.

[5] Anqi Zhu, Lin Zhang, Ying Shen, et al., "Zero-shot restoration of underexposed images via robust retinex decomposition," in *ICME*, 2020, pp. 1–6.

[6] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, et al., "Beyond brightening low-light images," *IJCV*, vol. 129, no. 4, pp. 1013–1037, 2021.

[7] Wenhan Yang, Wenjing Wang, Haofeng Huang, et al., "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *TIP*, vol. 30, pp. 2072–2086, 2021.

[8] Jun Xu, Yingkun Hou, Dongwei Ren, et al., "Star: A structure and texture aware retinex model," *TIP*, vol. 29, pp. 5022–5037, 2020.

[9] Xutong Ren, Wenhan Yang, Wen-Huang Cheng, et al., "Lr3m: Robust low-light enhancement via low-rank regularized retinex model," *TIP*, vol. 29, pp. 5862–5876, 2020.

[10] Minfeng Zhu, Pingbo Pan, Wei Chen, et al., "Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *AAAI*, 2020, vol. 34, pp. 13106–13113.

[11] Siyuan Li, Qingsha Cheng, Jianguo Zhang, "Deep multi-path low-light image enhancement," in *MIPR*, 2020, pp. 91–96.

[12] Yifan Jiang, Xinyu Gong, Ding Liu, et al., "Enlightengan: Deep light enhancement without paired supervision," *TIP*, vol. 30, pp. 2340–2349, 2021.

[13] Feifan Lv, Bo Liu, Feng Lu, "Fast enhancement for non-uniform illumination images using light-weight cnns," in *ACMMM*, 2020, pp. 1450–1458.

[14] Wenhan Yang, Shiqi Wang, Yuming Fang, et al., "Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality," *TIP*, vol. 30, pp. 3461–3473, 2021.

[15] Yu Zhang, Xiaoguang Di, Bin Zhang, et al., "Self-supervised image enhancement network: Training with low light images only," *arXiv preprint arXiv:2002.11300*, 2020.

[16] Lin Zhang, Lijun Zhang, Xiao Liu, et al., "Zero-shot restoration of back-lit images using deep internal learning," in *ACMMM*, 2019, pp. 1623–1631.

[17] Chunle Guo, Chongyi Li, Jichang Guo, et al., "Zero-reference deep curve estimation for low-light image enhancement," in *CVPR*, 2020, pp. 1780–1789.

[18] Chongyi Li, Chunle Guo, Qiming Ai, et al., "Flexible piecewise curves estimation for photo enhancement," *arXiv preprint arXiv:2010.13412*, 2020.

[19] Xiaojie Guo, Li Yu, Haibin Ling, "Lime: Low-light image enhancement via illumination map estimation," *TIP*, vol. 26, no. 2, pp. 983–993, 2016.

[20] Bo Sun, Ravi Ramamoorthi, Srinivasa Narasimhan, et al., "A practical analytic single scattering model for real time rendering," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 1040–1049, 2005.

[21] Srinivasa G. Narasimhan, Mohit Gupta, Craig Donner, et al., "Acquiring scattering properties of participating media by dilution," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 1003–1012, 2006.

[22] Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, et al., "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, 2017.

[23] Jianrui Cai, Shuhang Gu, Lei Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *TIP*, vol. 27, no. 4, pp. 2049–2062, 2018.

[24] Wufeng Xue, Lei Zhang, Xuanqin Mou, et al., "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *TIP*, vol. 23, no. 2, pp. 684–695, 2014.

[25] Lin Zhang, Lei Zhang, Xuanqin Mou, et al., "Fsim: A feature similarity index for image quality assessment," *TIP*, vol. 20, no. 8, pp. 2378–2386, 2011.

[26] Anish Mittal, Fellow, IEEE, et al., "Making a 'completely blind' image quality analyzer," *SPL*, vol. 20, no. 3, pp. 209–212, 2013.