LOW-LIGHT IMAGE ENHANCEMENT WITH ATTENTION AND MULTI-LEVEL FEATURE FUSION

Lei Wang[1], Guangtao Fu[2], Zhuqing Jiang[1], Guodong Ju[3], Aidong Men[1]

[1]Beijing University of Posts and Telecommunications, Beijing, China
[2]Academy of Broadcasting Science, Beijing, China
[3]GuangDong TUS-TuWei Technology Co.,Ltd

ABSTRACT

Low-light image enhancement has made impressive progress with convolutional neural networks (CNNs). However, most existing CNNs-based networks ignore the importance of feature channels and multi-level features. To address these issues, we propose a novel low-light image enhancement network. First, we establish Feature Extraction Block (FEB) to extract features and Feature Fusion Block (FFB) to fuse multi-level features. Then, we adopt a compact channel attention module to re-define the channel importance of input features at the beginning of each Feature Extraction Block (FEB) and Feature Fusion Block (FFB). Besides, we adopt two types of low-light image datasets for training, concluding synthetic images and real-world images. Experiments show that our new network makes competitive progress for low-light image enhancement compared with the state-of-the-art methods.

Index Terms— Low-light Enhancement, Feature Extraction, Channel Attention, Multi-level Features Fusion, Low-light Dataset

1. INTRODUCTION

Low-light image enhancement aims to improve images with low contrast and illumination. It is widely used in computer vision areas, and its final output images can be used for high-level visual tasks, like object detection, pedestrian re-identification, and autonomous driving.

Traditional low-light image enhancement methods mainly adopt the histogram equalization (HE) technique and Retinex-based method. Now the field of low-level visual tasks makes remarkable progress with a deep convolutional neural network. Same as low-light image enhancement. Liang et al. [1] successfully proposed a novel model based on convolutional neural network and Retinex theory. Eli et al. [2] trained an end-to-end image processing network based on the deep neural model for joint denoising and demosaicing. Although impressive improvements have been made, there remains much space to improve.

First, convolutional feature channels can adjust feature expressions flexibly by giving different weights to feature channels, which is propitious to extract useful features for recovering image contrast and illumination. However, existing methods based on Convolutional Neural Network (CNN) [1] [3] [4] handle the channel relationship equally without considering the distinction between feature channels, which leads to poor feature representation.

Second, in convolutional neural networks, features in deep layers have strong semantic information, while these in shallow layers contain a bigger receptive filed. These features represent different information. Most existing networks do not take advantage of multi-level features or just use skip connections to cancat features [16].

Third, capturing real-world low-light images with ground truth is difficult, so the inputs of most existing methods are synthetic pairs from existing datasets. Especially for networks with traditional technology, their input images are extracted from existing datasets, like Pascal VOC [5] or Microsoft COCO. These synthetic image pairs are easier processed than real-world low-light images because the latter is influenced by many environmental and unknown factors, such as environment illumination and camera shake. The most existing methods adopt only one type of dataset, ignoring the application of synthetic and real-world images.

To address these issues, we propose a novel method for low-light image enhancement by using the latest deep learning technology. As shown in Fig.1, our network consists of four types of modules for image demosaicing and denoising, i.e., RAW image preprocess, Feature Extraction Module, Feature Fusion Module, and sub-pixel image output. We combine convolution layer with attention module to extract image feature, as Feature Extraction Block (FEB). Then Feature Fusion Block (FFB) considers the FEB output, the last FEB up-sampling output and the previous FFB output as its input, making the best of multi-level features. And we adopt channel attention mechanism for obtaining significant feature expressions.

This work is supported by the Project of the National Natural Science Foundation of China No.61671077 and No.61671264.

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00054

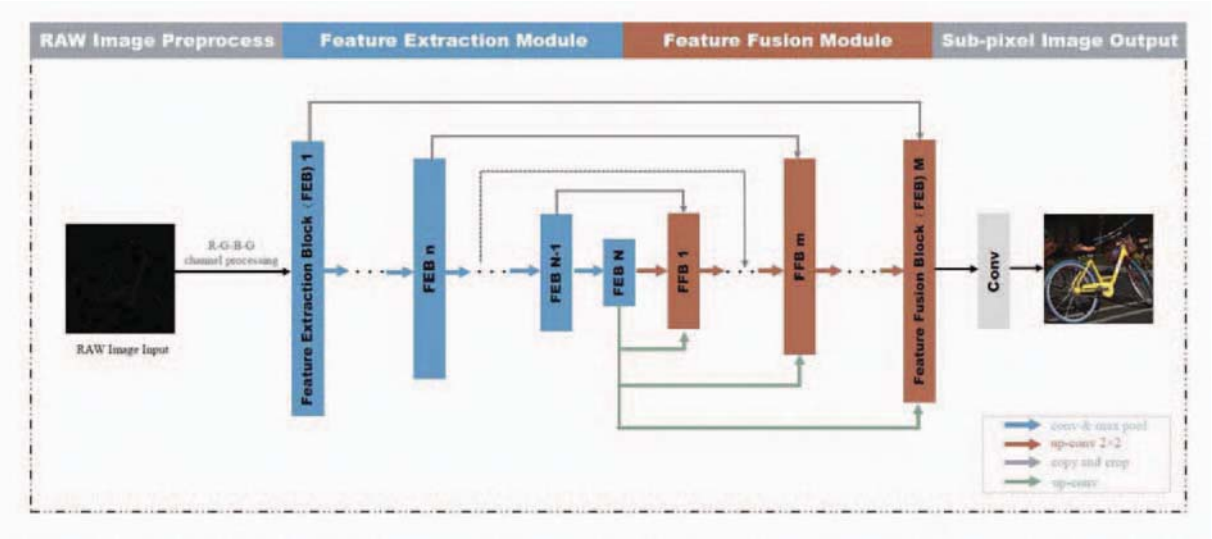


Fig. 1. **Network Architecture.** It consists of four types of modules, RAW image preprocess, Feature Extraction Module (FEM), Feature Fusion Module (FFM) and sub-pixel image output.

Besides, to make our network more adaptable for more types images, we adopt two types of low-light datasets: synthetic image dataset and real-world low-light dataset. Meanwhile, we have done adequate experiments to improve the robustness and stability of our networks. The contributions of our work are summarized as follows:

- *Attention module for feature extraction and feature fusion.* In each Feature Extraction Block (FEB) and Feature Fusion Block (FFB), we fuse a channel attention mechanism to recalibrate feature channel weight of input features.
- *Multi-level feature fusion.* We aggregate three levels of feature maps to improve low-light images. It concludes shallow level features from FEB, deep level features from FFB and a stable feature map extracted from the last feature extraction block. This fusion takes great advantage of multi-level features.
- *Two types of low-light datasets.* In our experiments, we adopt two types of low-light datasets, synthetic image dataset, and real-world low-light dataset. The synthetic dataset is extracted from the existing computer vision datasets, and real-world datasets choose SID and S7ISP for experiments.

2. RELATED WORK

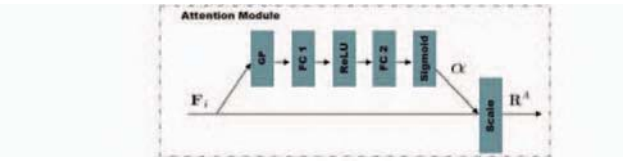
This section briefly introduces existing methods for low-light image enhancement and related techniques we used in our network.

2.1. Image processing

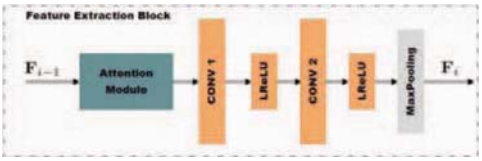
Traditional image processing methods conclude a series of modules such as white balance, demosaicing, denoising, color space conversion, and gamma correction. Image denoising makes great importance in image processing among all modules, so it is a well-developed task in low-level computer vision areas. There have been many approaches proposed using different techniques, such as sparse coding [6] [7], 3D transform-domain filtering (BM3D) [8] and so on. Recently, many approaches based on the deep neural network have been proposed. [9] restores images with a deep convolutional neural network; [10] uses deep network self-encoder technique to achieve adaptive brightening and denoising by training different low-light images. Besides, joint denoising and demosaicing has been taken into consideration using deep networks [11][12]. Unfortunately, most existing methods mentioned above have been evaluated on certain synthetic data.

2.2. Low-light Image Enhancement

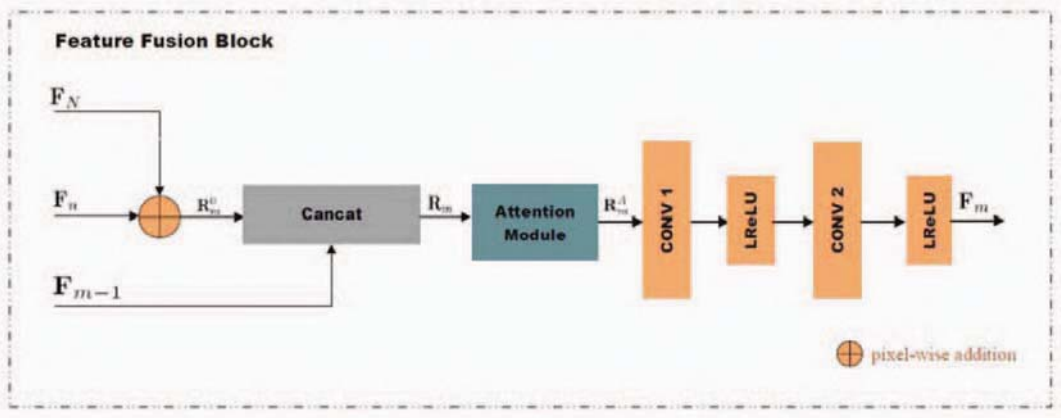
In general, low-light image enhancement methods can be mainly divided into two categories. One is histogram equalization (HE) technique and its variants. The other is built on the Retinex-based model. Histogram equalization balances the histogram of the whole image by increasing image contrast. Another image contrast enhancement method is Gamma Correction, which compresses bright pixels and expands the dark regions. Jobson et al. [13] introduce retinex technique for image enhancement, which can enhance various types of images adaptively. Then other methods based on retinex technique have been proposed, such as [14]. Recently, enhancement algorithms based on defogging and dark channel-first



(a) Attention Module



(b) Feature Extraction Block(FEB)



(c) Feature Fusion Block(FFB)

Fig. 2. Three main Blocks in our Network. (a) The GP and FC denote global pooling and fully connection layer. (b) FFB is used for feature extraction, and we adopt an attention module at the beginning of block. (c) Each FEB block has three input and we still adopt a attention module at the beginning of block. $\oplus$ refers to pixel-wise addition.

methods provide some new research directions, while these methods lack mature theories that lead to exaggerated results.

2.3. Attention mechanisms

Recently, attention mechanisms have been widely used in computer vision areas, especially for object detection and instance segmentation. Experiments show that attention modules improve performance. [15] considers the relation between image channels, so it proposes Squeeze-and-Excitation (SE-Net) adopting a novel "feature recalibration" strategy. Specifically, it learns the importance of each feature channels automatically and then enhances the useful features while suppresses features that are not vital in the current work. SE-Net makes a great improvement for object detection. This module also can be used in our network for feature abstract, and produce a good performance.

2.4. Low-light datasets

There are many datasets for image denoising and other image studies, such as RENOIR dataset, Google HDR+ dataset and Darmstadt Noise Dataset (DND). While the existing datasets fail to reflect the real-world low-light situations, so Chen et al. [16] creates a raw low-light dataset called See-in-the-Dark (SID). It includes 5094 raw short-exposure images and their corresponding long-exposure images. Our approach mainly uses SID and another RAW dataset for low-light image processing experiments.

3. THE PROPOSED METHOD

In this section, we will introduce the proposed network in details, including attention module, two feature related block, and overall network architecture.

3.1. Attention Module

Our method aims to use convolutional network for image joint demosaicing and denoising by obtaining low-light image feature. We connect low-level features with theirs corresponding high-level ones, while the direct connection leads to bad feature fusion results. So, enlightened by Hu et al. [15], we adopt this lightweight channel attention technique in our network, which selectively enhances useful features and suppresses less useful ones by giving feature channels with different weights.

As shown in Fig.2(a), we adopt a global average pooling to extract global channel information $\mathbf{F}_{scale} \in \mathbb{R}^{H \times W \times C}$ from the input $\mathbf{F}_i$. To fully capture channel-wise dependencies, we opt to two fully connection convolution layers (i.e. 1×1 convolution layers) apart with a LReLu activation and

Table 1. Comparisons results. PSNR for our network and other existing low-light image enhancement methods. Bold indicates the best result. Other methods results are taken from [4].

Dataset	Imput Image	SRIE [18]	LIME [19]	MSRCR [14]	SID-Sony	Ours
Synthetic Image	19.80	18.65	22.60	27.65	28.18	**29.68**
S7ISP-PNG	15.53	15.60	18.61	20.27	23.98	**25.18**
SID-sRGB	14.72	12.89	14.69	15.69	17.40	**17.84**
S7ISP	21.08	13.19	21.80	25.16	27.46	**29.22**
SID	12.83	12.98	22.17	25.16	28.78	**29.79**

a sigmoid activation. Among them, the first convolution layer aims to reduce the feature dimension with reduction ratio 4, and the second convolution layer has been set for recover the feature dimension. Final, we get the output of channel attention module (denoted as $\mathbf{R}^A \in \mathbb{R}^{H\times W\times C}$) by rescaling the input features $\mathbf{F}_i$ with channel weight α:

$$\mathbf{R}^A = \alpha \odot \mathbf{F}_i \tag{1}$$

where $\odot$ refers to channel-wise multiplication between channel weight α_i and the feature channel input $\mathbf{F}_i$, $i \in \mathbb{R}^{H\times W}$, $i = 1, 2, ..., C$.

3.2. Feature Extraction Block

This block is set for feature extraction. It is a simple typical convolutional block, consisting of a channel attention module and two convolutional layers, shown in Fig.2(b). The convolutional layer uses $32 \times n$ kernels of size 3×3 (where n means the n-th feature extraction block) , followed by a LReLu nonlinearity and a 2x2 max pooling operation with stride 2 for down-sampling. At each feature extraction block (FEB), we double the number of feature channels for next block calculation.

3.3. Feature Fusion Block

In order to make full use of hierarchical features extracted by Feature Extraction Block (FEB), we propose a new Feature Fusion Block (FFB) different from the feature directly connection in [16]. Feature Fusion Block (FFB) fuse multi-level features by pixel-wise addition and channel connection, because of more semantic information in Deep-level features and same channel number of corresponding convolution layers.

As shown in Fig.1(c), the m-th Feature Fusion Block (FFB) consists of three different level features input, denoted as the n-th FEB output ($\mathbf{F}_n$) ,the last FEB up-sampling output ($\mathbf{F}_N$) and the previous FFB output ($\mathbf{F}_{m-1}$),where N is the number of FEB. For $\mathbf{F}_m$, we apply 2×2 up-convolutional layer to matching feature scales. Meanwhile, for $\mathbf{F}_n$, we just crop and transport to the next module.

For m-th FFB in our network, we denoted $R_m^0 \in \mathbb{R}^{H\times W\times C}$ as the output of pixel-wise addition of $\mathbf{F}_n$ and $\mathbf{F}_N$:

$$R_m^0 = \mathbf{F}_n \oplus \mathbf{F}_N \tag{2}$$

where $\oplus$ means pixel-wise addition. Then we directly connect R_m^0 with F_{m-1} as the final input of m-th feature fusion block (represented as R_m).

Obviously, the input R_m consists of three different level features, which makes bad results for later processing if we directly send it to the FFB. So, we apply the attention module to obtaining a channel-weighted feature representation R_m^A, which is used for next convolutional processing. This process can be easily expressed as

$$R_m^A = M[(\mathbf{F}_n \oplus \mathbf{F}_N) + F_{m-1}] \tag{3}$$

where $M[\cdot]$ represents channel attention operation. Finally, the integration feature R_m^A is processed by two 3×3 convolutional layer (denoted as $S[\cdot]$), each follow by a LReLU nonlinear layer. We obtain the final output of the m-th FFB:

$$F_m = \delta(S[\delta(S\,[R_m^A])]) \tag{4}$$

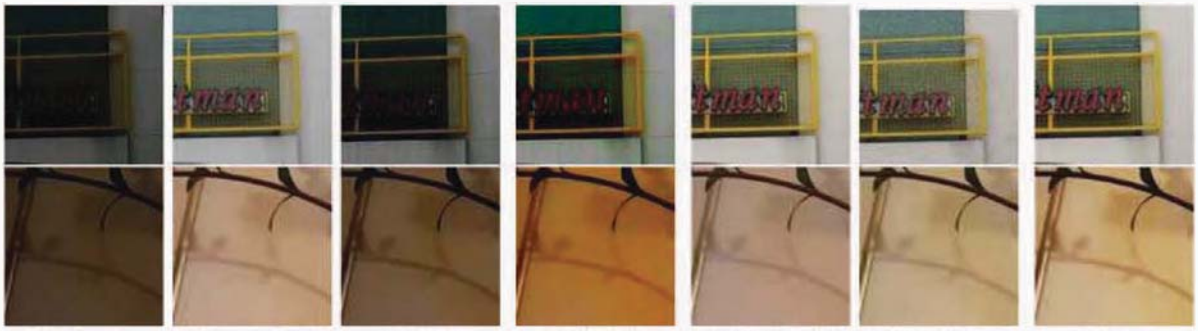
where $\delta(\cdot)$ refers to the functions of LReLu.

3.4. Network Architecture

The proposed network, as shown in Fig.1, consists of four types of module, RAW image preprocess, Feature Extraction Module, Feature Fusion Module, and sub-pixel image output. Inspired by [16], the first module handles the input image to a four-channel R-G-B-G pattern, including black level subtraction and amplification. The main parts of our networks are total N FEBs and M FFBs, aiming for image feature extraction and fusion. After M FFBs, our model output a 12-channel RGB image, and then we use sub-pixel image process mentioned in [17] to get a high-quality output image.

4. EXPERIMENTS

The proposed network is evaluated and compared with the existing low-light image enhancement through extensive experiments. In this section, we formulate four major parts, includ-



(a)Input (b)Ground Truth (c)SRIE (d)LIME (e)MSRCR (f)SID-Sony (g)Ours

Fig. 3. Visual results of different methods on synthetic and real-world images

ing datasets, implementation details, comparisons with the state-of-the-art methods and network ablation experiments.

4.1. Datasets

Different from some existing network, we choose synthetic images and real-world low-light images for experiments. Inspired by[4], we conduct a set of low-light images based on the Pascal VOC [5] as synthetic images. As for real-world low-light images, we mainly use see-in-the-dark dataset (SID)[16] and S7ISP dataset [2]. The former one contains 5094 raw short-exposure images, each with a corresponding long-exposure ground-truth image. The latter one, using a Samsung S7 rear camera, totally includes 110 pairs of raw images and their corresponding PNG images.

4.2. Implementation Details

Taking the trade-off between performance and time consumption, we use five feature extraction blocks (FEBs) and four feature fusion blocks (FFBs) in our network. All convolutional layers in our model have the same kernel size (3×3) and the stride 1. Considering the factors of image dataset and computer memory, we randomly crop a 512×512 patch for training and apply flipping and rotation for data augmentation. Our model is evaluated with Peak Signal to Noise Ratio (PSNR). We totally adopt 4000 epochs for training. The learning rate is set to 10^{-4} and is decreased to 10^{-5} after 2000 epochs. Our method is implemented with pytorch and an NVIDIA GTX 1080ti GPU.

4.3. Comparisons with the state-of-the-art methods

We use the peak signal-to-noise ratio (PSNR) for a quantitative evaluation. A higher PSNR means that the output image is closer to the ground truth. We compare our proposed method with four existing low-light image enhancement methods: SRIE [18], LIME [19], MSRCR [14] and SID-Sony [16]. As shown in Table 1, all the best results are boldfaced, and our network obtains the highest PSNR results both on synthetic and real-world low-light images.

Fig.3 shows the visual comparison of two types of images. Row 1 displays synthetic images, and row 2 shows RAW pattern real-world images respectively. The results of SRIE prefer to black to a certain degree. LIME results have high saturation, and MSRCR appears a little sharpening to the edge of objects. SID-Sony is close to our results, while our network achieves better performance in details.

Table 2. Ablation experiments on attention module and feature fusion.

Attention module	Feature Fusion	PSNR
		28.78
✓		28.89
	✓	29.37
✓	✓	**29.79**

Table 3. Experiments the number of FEB and FFB.

M	N	PSNR
4	3	28.91
5	4	**29.79**
6	5	28.99

4.4. Ablation experiment

To verify the effects of attention module and feature fusion, we conduct four ablation experiments. The results are shown in Table 2. Keeping attention module in our model only has an increase of 0.11. By removing the attention module and holding feature fusion, we get a relatively well result, almost 0.6 improvements. It confirms that multi-level fusion makes a great influence on low-light enhancement. Meanwhile, the row4 in Table 2 shows the attention module after feature fusion is necessary.

Besides, the number of FEB (M) and FFB (N) is vital to our final results, so we take another experiment of the pro-

posed network. As shown in Table 3, we choose three pairs of parameters for training, and it is clear that M=5 and N=4 makes the best performance.

5. CONCLUSION

In this paper, a novel network with multi-level fusion and attention module is proposed for low-light image enhancement. We mainly introduce two parts of our network, FEB and FFB. Both of the two parts adopt the attention module for channel recalibration, and FFB is established for multi-level features fusion. Experiments show that our new network makes competitive progress for low-light image enhancement.

6. REFERENCES

[1] Liang Shen, Zihan Yue, Fan Feng, Quan Chen, Shihao Liu, and Jie Ma, "Msr-net: Low-light image enhancement using deep convolutional network," *arXiv preprint arXiv:1711.02488*, 2017.

[2] Eli Schwartz, Raja Giryes, and Alex M Bronstein, "Deepisp: Learning end-to-end image processing pipeline," *arXiv preprint arXiv:1801.06724*, 2018.

[3] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu, "Gladnet: Low-light enhancement network with global awareness," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 751–755.

[4] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim, "Mbllen: Low-light image/video enhancement using cnns," in *British Machine Vision Conference*.

[5] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.

[6] Elad Michael and Aharon Michal, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Tip*, vol. 15, no. 12, pp. 3736–3745, 2006.

[7] Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman, "Non-local sparse models for image restoration," in *IEEE International Conference on Computer Vision*, 2010.

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian, "Image denoising by sparse 3-d transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007.

[9] L. Xu, J. S. J. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *International Conference on Neural Information Processing Systems*, 2014, pp. 1790–1798.

[10] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[11] Micha?l Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frdo Durand, "Deep joint demosaicking and denoising," *Acm Transactions on Graphics*, vol. 35, no. 6, pp. 1–12, 2016.

[12] Keigo Hirakawa and Thomas W Parks, "Joint demosaicing and denoising," in *IEEE International Conference on Image Processing*, 2005.

[13] D J Jobson, . Rahman, Z., and G A Woodell, "Properties and performance of a center/surround retinex," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 6, no. 3, pp. 451–62, 1997.

[14] D J Jobson, . Rahman, Z., and G A Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image Processing*, vol. 6, no. 7, pp. 965–976, 2002.

[15] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] Chen Chen, Qifeng Chen, Xu Jia, and Vladlen Koltun, "Learning to see in the dark," 2018.

[17] Wenzhe Shi, Jose Caballero, Ferenc Huszr, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Computer Vision Pattern Recognition*, 2016.

[18] Xueyang Fu, Delu Zeng, Huang Yue, Yinghao Liao, Xinghao Ding, and John Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Processing*, vol. 129, no. C, pp. 82–96, 2016.

[19] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Trans Image Process*, vol. 26, no. 2, pp. 982–993, 2017.