Sea Surface Object Recognition Under the Low-Light Environment

Zihan Yue[1,2], Liang Shen[1,2], Wei Lin[1,2], Wu Lv[3], Shihao Liu[1,2], Tao Geng[1,2], and Jie Ma[1,2(✉)]

[1] National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
{yuezihan,shenliang,wlin,shihaoliu,majie}@hust.edu.cn
[2] Guangdong HUST Industrial Technology Research Institute, Guangdong Province Key Lab of Digital Manufacturing Equipment, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
[3] China State Shipbuilding Corporation Systems Engineering Research Institute, Beijing 100094, China
lw_all@126.com

Abstract. Computer vision research usually needs sufficient high-quality images. However, the contrast of images captured at night is always low, which makes many classification tasks difficult. For example, it is hard for an unmanned surface vessel to sense the environment at night. In this paper, we attempt to solve the low-light image classification problem via using deep learning methods. Inspired by the multi-scale Retinex algorithm, we design a novel convolutional structure named dedark block to enhance the low-light image and add it to the CNN architecture for object recognition. By creating different models, we gradually improve our models to close to the result of CNN baseline based high-quality images, and every model is good motivated. Besides the strong learning ability of CNN, the dedark block which is regarded as pre-training also plays an important role in our final model. At last, the result of every model is evaluated on two classical datasets, and all of them achieve great performance.

Keywords: Low-light enhancement · Object recognition
Retinex algorithm

1 Introduction

Image classification has been a classic task in computer vision. Since the breakthrough of convolutional neural network in image classification [1–7], many CNN-based networks have been proposed and the accuracy has dramatically been improved. Unfortunately, all these methods work on the good quality of

L. Shen contributed equally.

Z. Chen et al. (Eds.): ICIRA 2018, LNAI 10985, pp. 346–356, 2018.
https://doi.org/10.1007/978-3-319-97589-4_29

images, whereas the captured images in real-world scenes are usually degraded. For instance, when we get images and videos at night, the low contrast and brightness of images always makes the classification tasks difficult. Considering the fact that in many conditions only dark images can be obtained, a great number of enhancement methods have been put forward to deal with this problem.

In this paper, we propose a novel convolutional neural network structure to solve low-light image classification. To the best of our knowledge, our work is the first end-to-end learning model for low-light object classification. Firstly, inspired by the multi-scale Retinex algorithm [8–10], we design a dedicated convolutional neural network named dadark block and explain the relationship with multi-scale Retinex algorithm. We think dark image enhancement as a machine learning problem. Unlike traditional method, most of the weights are updated by back-propagation [11]. After dedark block layer, the second part in our network is convolutional and fully-connected classification layers. The whole network can be considered as an end-to-end mapping between dark images and their labels. A lot of experiments reveal the advantages of our model in comparison with traditional methods such as histogram transformation.

Overall, the main contributions of our work are as follows: Firstly, based on multi-scale Retinex [8], we propose a special structure named dedark block to enhance the low-light image. Secondly, by concatenating the dedark block with convolutional and fully-connected layers, we consider low-light image classification as an end-to-end machine learning problem between dark images and their labels. Last but not least, in experiments we gradually improve our models to close to the result of classification methods based high-quality images.

2 Related Works

2.1 Image Classification

Recently, convolutional neural networks have made a series of breakthroughs in computer vision. For example, Alexnet [2] achieves a top-5 error rate of 16.4% in ILSVRC-2012, which exceeds the traditional methods in a great extent. Lin *et al.* [3] proposed micro neural networks in every layer with complex structures. By increasing the number of layers in the network, VGG-net [4] shows that the depth of network is very important. And a number of image classification tasks have also greatly benefited from these deep networks. At the same time, the inception architecture of GoogleNet [5] has been proposed in view of computational efficiency and practicality. They increase the width and depth of the network by using a well-designed structure and save the computing resource at the same time. After this, in order to make optimization easier, more ingenious structures have been proposed. For example, Resnet [6] presents a residual structure which makes the training process of networks easier than those used previously. They attempt to learn the functions between the residual and the layer inputs, rather than learn it between different layers directly. Densenet [7] attempts to connect each layer to every other layer in the network. For each layer, all of the preceding layers are used as inputs, and the result of this layer

is also used as input into subsequent layers. Convolutional neural networks usually accumulate low/mid/high-level feature-maps in a multi-layer structure, and deeper levels of features usually correspond to semantic information in an image.

2.2 Low-Light Image Enhancement

Generally speaking, low-light image enhancement can be roughly classified to two categories: histogram-based and Retinex-based models.

One of the traditional methods to lighten the dark image is based on histogram transformation [12], which amplifies the dark image directly. For example, histogram equalization (HE) makes the histogram of the whole image as balanced as possible. Logarithmic transformation is also a good choice to increase the contrast by expanding the dark regions and compressing the bright ones. However, all of these methods try to deal with every pixel in the image individually, which makes the enhanced result inharmonious to some extent. Variational methods such as [13] have been proposed to resolve this problem.

Land [14] introduced the Retinex theory to explain the color perception property. He assumed that the image can be decomposed into reflection and illumination. Single-scale Retinex (SSR) [15] proposed by Jobson *et al.* is based on the center/surround Retinex and regards the reflection as the enhanced image. After this, they put forward Multi-scale Retinex (MSR) [8], which is equal to a weighted sum of several SSR results.

2.3 Combining Low-Level and High-Level Vision Tasks

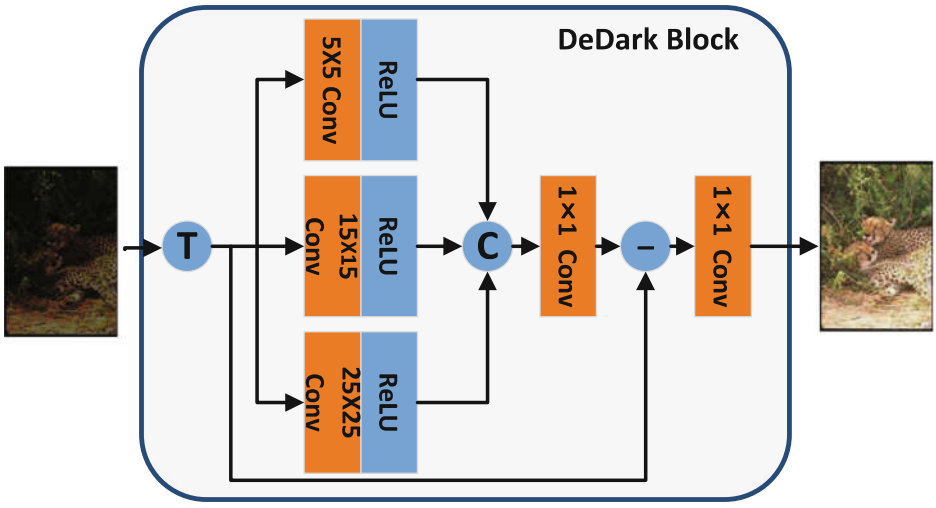
Recently, the breakthroughs of convolutional neural network have made a series of improvements in computer vision areas such as classification [6], detection [16], tracking [17], segmentation [18] and so on. At the same time, deep learning has also played an important role at low-level vision tasks such as super-resolution [19], de-rain [20] and de-haze [21]. However, all of these methods are independent to others. Combining low-level and high-level vision tasks is a novel direction. LR-CNN [22] is a deep structure designed for classification of low-resolution images. Li *et al.* [23] tried to concatenate DehazeNet and Faster R-CNN [16], which achieves great improvement on the object detection performance of hazy images. Liu *et al.* [24] attempted to use the semantic information from those high-level tasks to guide image denoising. Vasiljevic *et al.* [25] showed that network trained on high-quality images usually suffers a significant degradation on performance when it is applied to those blurred images. Specifically, they used the blur images to fine-tune the network that trained on high-quality images, which allows it to regain much of the lost accuracy. Diamond *et al.* [26] proposed an end-to-end structure for both denoising and deblurring, and made classification tasks robust to realistic noise and blur. This structure improves the accuracy of a classification model to a great extent on many challenge conditions.

3 Low-Light Image Classification

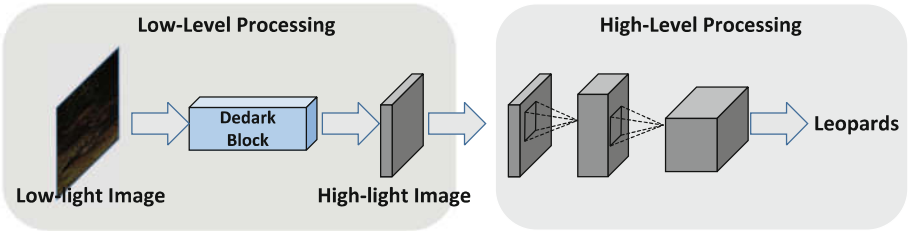
Firstly, we propose a dedicated convolutional neural network and explain the relationship with multi-scale Retinex algorithm. Then we concatenate it with convolutional and fully-connected layers to classify different low-light images.

3.1 Low-Light Image Enhancement Structure

In this section, we elaborate that multi-scale Retinex as an image enhancement method can be considered as a special structure as shown in Fig. 1. Retinex theory assumes that the image can be decomposed into reflection and illumination. Based on this surround Retinex, Single-scale Retinex (SSR) [15] is similar to the difference-of-Gaussian (DOG) function. Further, multi-scale Retinex (MSR) [8]



(a) DeDark block architecture



(b) The overall architecture

Fig. 1. DeDark block architecture and the overall architecture. (a) DeDark block architecture, where 'C', 'T', and '-' represent the concatenation, logarithmic transformation, and element-wise subtraction, respectively. (b) The overall architecture, which is divided into two parts, low-level processing and high-level processing. High-level part can be replaced by these common classification networks such as AlexNet, VGG-net and so on.

is equal to a weighted sum of several SSR results. Mathematically, the formula is as follows:

$$R_{MSR_i} = \log I_i(x,y) - \frac{1}{3}\log I_i(x,y) * \left[\sum_{n=1}^{3} F_n(x,y)\right] \tag{1}$$

Where $F_n(x,y)$ represents the n^{th} surround function.

Noticing the fact that the sum of convolution can be represented as an Inception structure [5], and the subtraction can be considered as a residual structure [6]. Therefore, we can represent the above Eq. 1 by using a special convolutional network named Dedark block, as shown in Fig. 1(a).

More specifically, considering that MSR is consist of three different Gaussian surround functions, the receptive field of three convolutional layers in our model is set to 5×5, 15×15 and 25×25 respectively. After this, we concatenate these 3D tensors to a larger 3D tensor H and let it go through the convolutional layer:

$$H = [H_1, H_2, H_3] \tag{2}$$

$$H_{avg} = H * W_{avg} + b_{avg} \tag{3}$$

Where W_{avg} represents the parameters of a convolutional layer with three output channels (red, green and blue) and the 1×1 receptive field, which plays an average role in three images above. And if we set it to be more concrete:

$$W_{avg} = \begin{bmatrix} \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{bmatrix} \tag{4}$$

Then this layer is exactly equivalent to the average in MSR. After this, a residual structure [6] is used to represent the subtraction between the oral and dark image.

As we can see, our special convolutional network is similar to traditional image enhancement methods such as multi-scale Retinex. One of the biggest different points is that low-light image enhancement in this paper is regarded as a machine learning problem. More details about our training dataset will be explained in Sect. 4.

The goal of this module is to make the enhanced image $f(X)$ and the real image Y as close as possible. For more visualization, dark images and the enhanced results via our dedark block are shown in Fig. 2. As we can see, the low-light images are restored to the original images realistically by our dedark block.

3.2 Classification Layers

The full network of our model is elaborated in Fig. 1(b). As we can see, after image enhancement layers, the second part of our model is convolutional and fully-connected classification layers. As shown in Sect. 2, a number of CNN

frameworks [2–7] have been proposed for object recognition. In this paper we choose two popular structures: AlexNet [2] and ResNet [6]. Both of them consist of many Conv-ReLU-Pool layers followed by several fully-connected layers. In Fig. 1(b), we illustrate the structure of AlexNet for simplicity.

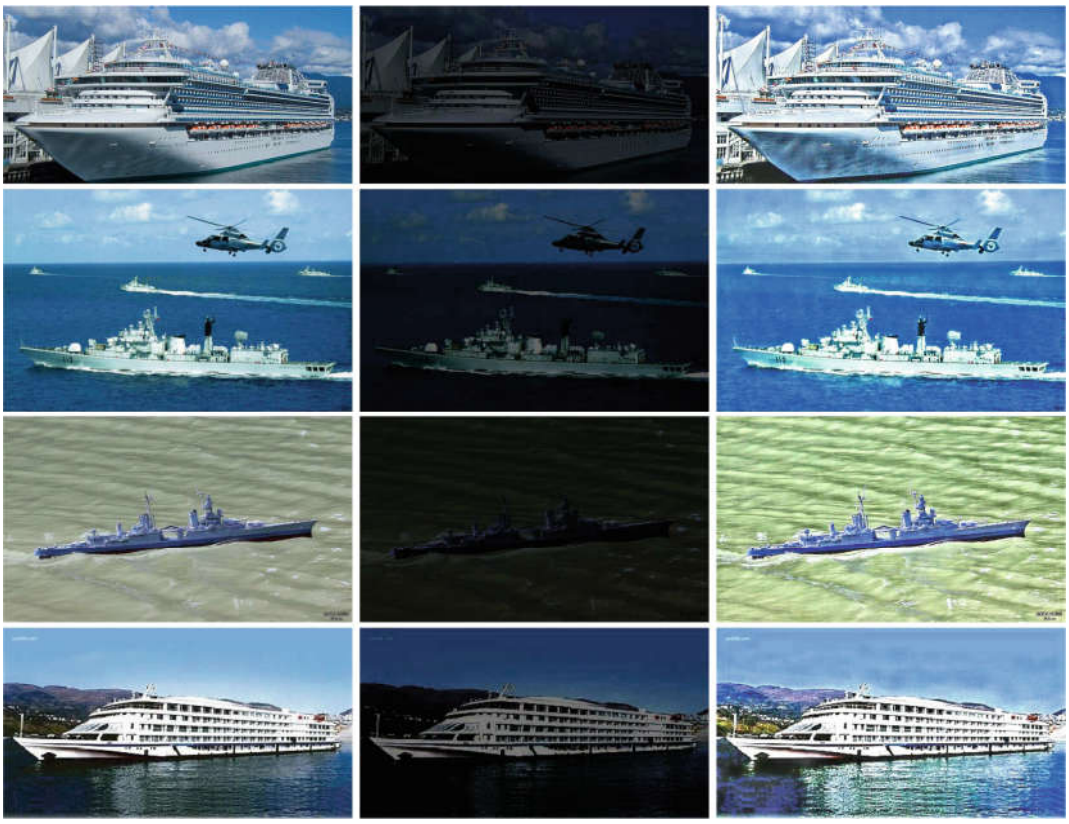
The whole network can be considered as an end-to-end mapping between dark images and their categories. We use the cross entropy loss as the optimization goal. Mathematically, the loss function can be written as:

$$L(\hat{y}, y) = -\frac{1}{C}\sum_{c \in C} y_c \log \hat{y}_c \tag{5}$$

Where C is the set of all possible labels, y is a one-hot ground-truth label vector, $\hat{y}$ is the result of softmax.

4 Experiments

In this section, we construct an image dataset for training the dedark block network. To evaluate the classification accuracy of our model, we gradually improve our models to close to the result of method based high-quality images.



(a) ground truth (b) low-light image (c) Our results

Fig. 2. Image enhancement results via our model. Figure (a): Ground truth images. Figure (b): Dark images generated by our method in Sect. 4.1. Figure (c): Our enhanced results.

4.1 Image Dataset Generation

To the best of our knowledge, our work is the first attempt for low-light image classification. Therefore, no public dataset is available. We generate our own dataset based on two classical datasets: Flower 102 [27] and Caltech 101 [28]. Considering that all the images in the datasets look natural, we need to generate the corresponding low-light images of those high-quality images. Each high-quality image is used to generate a low-light image by two steps. At first, by transforming the image to HSV space, we scale the V(Value) component with a random factor ranging from 0.5 to 1. Then we use gamma transform with parameters ranging from 1 to 3 to darken the image further. Figure 2 gives a high-quality/low-light example. Our dataset is consist of low-light images as inputs and their categories as labels.

4.2 Training Setting

The whole network is consist of two component: Dedark block and classification block. As is mentioned before, the dedark block is made up of an Inception structure and a Residual structure. We use AlexNet and ResNet-8 for classification. All of the experiments are trained on Pytorch software package. A popular optimization algorithm Adam [29] is used in the training process. We start with a learning rate of 0.001 and use weight decay of 0.0001. Experiments reveal that 600 epoches are enough for the model to converge as Fig. 4 shows.

4.3 Comparison with Different Models

In this section we start from high-quality images based model firstly and perform a step-by-step model evolution to close to this upper bound.

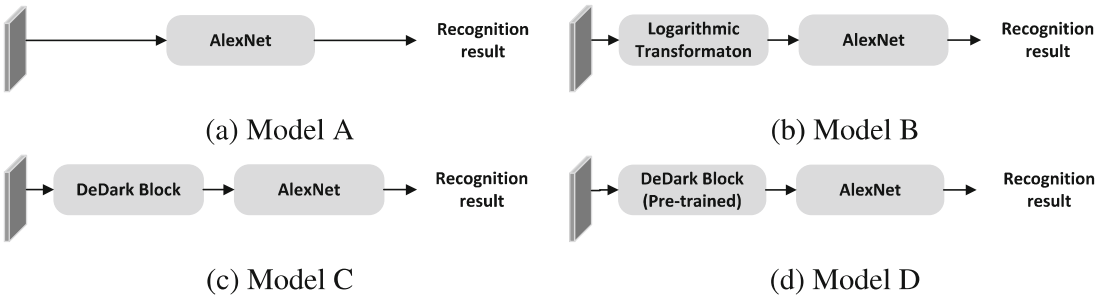


Fig. 3. Different models for the low-light image classification. (a) Model A: Classify the high-quality image directly using traditional CNN models such as Alexnet (which is used for comparation). (b) Model B: Use logarithmic transform to enhance the dark image at first, and then classify it by CNN. (c) Model C: Low-light image classification is considered as an end-to-end network by adding the dedark block. (d) Model D: Pre-training the dedark block between dark and bright images and then fine-tune it on model C.

Table 1. The classification accuracy of different models

Network	Model	Flower 102	Caltech 101
AlexNet	Model A	75.2%	73.4%
	Model B	68.9%	67.3%
	Model C	70.1%	69.1%
	Model D	71.8%	72.8%
Resnet-8	Model A	81.4%	67.1%
	Model B	73.5%	62.8%
	Model C	74.2%	63.4%
	Model D	75.1%	63.5%

Model A. As is shown in Fig. 3(a), the first method is to classify the high-quality image directly using traditional CNN models such as Alexnet [2]. The accuracy of this method is shown in Table 1. There is no doubt that high-quality images will obtain high accuracy. In the rest of this section we try our best to make our model close to this result.

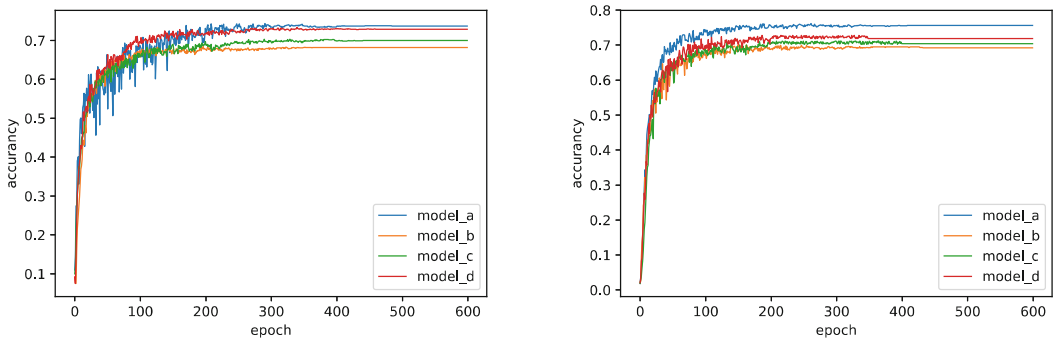
Model B. Actually, most of the details in the dark image have been buried into the background. Considering the fact that low-contrast or low-light condition always makes the classification task difficult, a lot of image enhancement methods have been proposed to solve this problem. One of the most widely used technique is histogram equalization, which makes the histogram of the whole image as balanced as possible. Logarithmic transformation is also a good choice to increase the contrast by expanding the dark regions and compressing the bright ones. In this work we choose the latter for comparison. The whole structure of the model is shown in Fig. 3(b). As we can see, the only difference between model A and model B is the addition of a logarithmic transformation, which makes the image brighter. However, experiments reveal that logarithmic transformation makes a limit improvement in the accuracy of image classification, which is far less than the result of high-quality. Table 1 shows the accuracy gap between them.

Model C. The main drawback of model B is that all the dark images are enhanced by identical logarithmic transformation without the dependence of image distribution, which makes the enhanced result inharmonious to some extent. In Sect. 3 we consider low-light image enhancement as a supervised learning model, which can solve this problem to some extent. The simple structure of this model is shown in Fig. 3(c), while more details have been elaborated in Fig. 1. As we can see, this model can be treated as two parts. The first part lightens the dark image to a bright image, and the second part classifies the bright images. The whole network can be considered as an end-to-end mapping between dark images and their labels. All of the parameters are optimized by back-propagation. As Table 1 shows, comparing to traditional histogram trans-

formation methods, learning-based methods can further improve the accuracy of classification.

Model D. Although the improvements are obtained by an end-to-end mapping between dark images and their labels, we attempt to squeeze the last bit of performance on convolutional neural network. As has been elaborated in Sect. 3, dedark block is equivalent to multi-scale Retinex if suitable weights are selected. However, all of the weights in Model C are initialized randomly, which may make the dedark module plunge into local optimum. A more appropriate consideration is that the dedark block can be pre-trained between dark and bright images. This problem can be regarded as a regression model by using the Frobenius norm as the loss function, as Sect. 3 suggests. After this, the whole network is fine-tuned on these pre-trained weights. The whole procedure of Model D is shown in Fig. 3(d). Test set accuracies in Table 1 show that pre-trained model is superior to Model C. Hence, better performance of Model D is credited to the knowledge about the weights of dark image enhancement, which verifies our motivation to classify low-light image via image enhancement.

Figure 4 shows the convergence curves on different datasets. As we can see, model A always reaches the highest accuracy and model D follows it. By pre-training the dedark block on model C, the effect of image enhancement has been revealed, which is in line with our expectation.



(a) Caltech 101:AlexNet (b) Flower 102:AlexNet

Fig. 4. Convergence curves of different datasets. Figure (a) corresponds to four models trained by Caltech 101 based on AlexNet, Figure (b) corresponds to four models trained by Flower 102 based on AlexNet. As we can see, model A always reaches the highest accuracy and model D follows it.

5 Conclusion

In this paper, we propose the low-light image enhancement structure based on Retinex theory, and combine it with traditional classification network. Comparing to the CNN baseline trained by high-quality images, we gradually improve our models to close to the result of methods based high-quality images, and every model is good motivated. Besides the large learning capacity of CNN, the

final model also benefits from the low-light image enhancement pre-training. The effectiveness of the proposed method is evaluated on two classical datasets and it obtains outstanding performances.

Acknowledgments. This work was supported by the Guangdong Innovative and Entrepreneurial Research Team Program under Grant 2014ZT05G304.

References

1. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
2. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
3. Lin, M., Chen, Q., Yan, S.: Network in network. Computer Science (2013)
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
5. Szegedy, C., et al.: Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. arXiv preprint arXiv:1608.06993 (2016)
8. Jobson, D.J., Rahman, Z.-U., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997)
9. Petro, A.B., Sbert, C., Morel, J.M.: Multiscale retinex. Image Process. Line **4**, 71–88 (2014)
10. Provenzi, E., De Carli, L., Rizzi, A., Marini, D.: Mathematical definition and analysis of the retinex algorithm. J. Opt. Soc. Am. A Optics Image Sci. Vis. **22**(12), 2613–2621 (2005)
11. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**(6088), 533–536 (1986)
12. Pizer, S.M., et al.: Adaptive histogram equalization and its variations. Comput. Vis. Graphics Image Process. **39**(3), 355–368 (1987)
13. Celik, T., Tjahjadi, T.: Contextual and variational contrast enhancement. IEEE Trans. Image Process. **20**(12), 3431–3441 (2011)
14. Land, E.H.: The retinex theory of color vision. Sci. Am. **237**(6), 108–129 (1977)
15. Jobson, D.J., Rahman, Z.-U., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE Trans. Image Process. **6**(3), 451–462 (1997)
16. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
17. Wang, L., Ouyang, W., Wang, X., Lu, H.: Visual tracking with fully convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3119–3127 (2015)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)

19. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 295–307 (2016)
20. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1715–1723 (2017)
21. Cai, B., Xiangmin, X., Jia, K., Qing, C., Tao, D.: DehazeNet: an end-to-end system for single image haze removal. IEEE Trans. Image Process. **25**(11), 5187–5198 (2016)
22. Chevalier, M., Thome, N., Cord, M., Fournier, J., Henaff, G., Dusch, E.: LR-CNN for fine-grained classification with varying resolution. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 3101–3105. IEEE (2015)
23. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: AOD-Net: all-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4770–4778 (2017)
24. Liu, D., Wen, B., Liu, X., Huang, T.S.: When image denoising meets high-level vision tasks: A deep learning approach. arXiv preprint arXiv:1706.04284 (2017)
25. Vasiljevic, I., Chakrabarti, A., Shakhnarovich, G.: Examining the impact of blur on recognition by convolutional networks. arXiv preprint arXiv:1611.05760 (2016)
26. Diamond, S., Sitzmann, V., Boyd, S., Wetzstein, G., Heide, F.: Dirty pixels: optimizing image classification architectures for raw sensor data. arXiv preprint arXiv:1701.06487 (2017)
27. Nilsback, M.-E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, December 2008
28. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
29. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. Computer Science (2014)