Improving Generalization Ability of Deep Neural Networks for Visual Recognition Tasks

Takayuki Okatani[1,2(✉)], Xing Liu[1], and Masanori Suganuma[1,2]

[1] Graduate School of Information Sciences, Tohoku University, Sendai, Japan
okatani@vision.is.tohoku.ac.jp
[2] RIKEN Center for AIP, Tokyo, Japan
http://www.vision.is.tohoku.ac.jp

Abstract. This article discusses generalization ability of convolutional neural networks (CNNs) for visual recognition with special focus on robustness to image degradation. It has been long since CNNs were claimed to surpass human vision, for example, in an object recognition task. However, such claims simply report experimental results that CNNs perform better than humans on a closed set of testing inputs. In fact, CNNs can easily fail for images to which noises are added, when they have not learned the noisy images; this is the case even if humans are barely affected by the added noises. As a solution to this problem, we discuss an approach that first restores the clean image from an input distorted image and then uses it for the target recognition task, where a CNN trained only on clean images is used. For solutions to the first step, we show our recent studies of image restoration. There are multiple different types of image distortion, such as noise, defocus/motion blur, rain-streaks, raindrops, haze etc. We first introduce our recent study of architectural design of CNNs for image restoration targeting at a single, identified type of distortion. We then introduce another study, which proposes to use a single CNN to remove combination of multiple types of distortion with unknown mixture ratio. Although it achieves only lower accuracy than the first method in the case of a single, identified type of distortion, the method will be more useful in practical applications.

Keywords: Visual recognition · Convolutional neural networks · Generalization ability

1 Introduction

The emergence of convolutional neural networks has reshaped research in the field of computer vision in the past seven years. Their employment has brought about solutions to unsolved problems or contributed to (sometimes significant) improvements in performance (e.g., inference accuracy, computational speed etc.). It was claimed in the past years that CNNs can even surpass human vision in several visual recognition tasks, in particular, the task of object category classification [7].

S. Tominaga et al. (Eds.): CCIW 2019, LNCS 11418, pp. 3–13, 2019.
https://doi.org/10.1007/978-3-030-13940-7_1

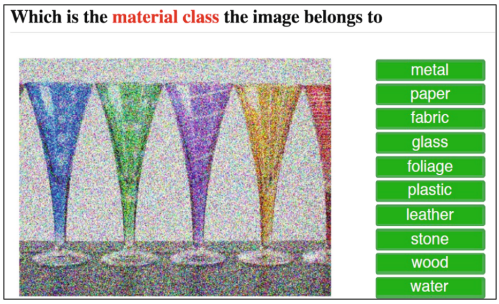

Fig. 1. Material classification from a noisy image. Humans and CNNs choose one of the ten material categories shown on the right.

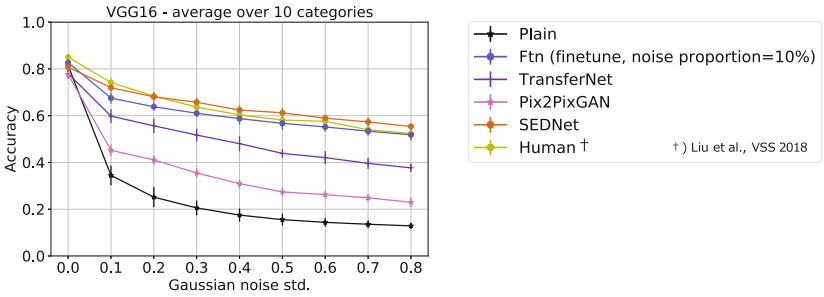

Fig. 2. Accuracy of material classification versus the strength of Gaussian noise added to the input images. Plain indicates a CNN trained only on clean images and Ftn is a CNN trained on both clean and noisy images.

However, we should be precise about the meaning (or underlying condition) of such claims. Each of them is made based on experiments that compare CNNs and humans on a recognition task using a *particular* dataset. The experimental results merely indicate that CNNs are better in terms of recognition accuracy than humans on a *closed* set of test inputs. In other words, CNNs may correctly classify inputs that are sampled from the same distribution as the training data they have learned, but will wrongly classify inputs sampled from a different distribution. The two distributions usually need to be very close; their difference is called *domain shift*, which is known as one of major causes that impede applications of CNNs to real-world problems.

One such example is shown in Fig. 1. The image shown on the left is a noisy version of a sample belonging to Flickr Material Databse (FMD) [18], which is a popular dataset for ten-class material classification task; the ten classes are shown on the right of the figure. We consider here the material classification task from noisy input images. The original images of FMDs are noise-free, and we add Gaussian random noises with a certain strength to them.

Figure 2 shows the results. It shows performance of various CNN models and humans for different strengths of additive Gaussian noises. An overall tendency is that humans and all the CNN models show similar accuracy in the noise-free case, and they all deteriorate as the noise strength increases. It is, however, seen that the performance decrease for humans is almost the smallest, while the CNN trained only on clean (i.e., noise-free) images performs the worst for noisy inputs. On the other hand, the CNN trained also on noisy images shows comparable performance to humans. This phenomenon demonstrates the aforementioned issue with neural networks; they work very well for trained data but can fail for inputs sampled from a slightly different distribution, even if the difference is mostly negligible for humans.

In this article, we consider how to cope with the issue with deep learning. We first discuss how to enable to perform visual recognition from degraded images such as noisy images considered above.

2 Image Restoration for Robust Visual Recognition

2.1 Visual Recognition Robust to Image Distortion

There are three approaches to visual recognition from distorted (or degraded) images, as shown in Fig. 3. The first approach, which is conceptually the simplest, is to train the CNN using not only clean images but noisy images. Then, the CNN will accurately recognize noisy inputs, as discussed above and shown in Fig. 2. However, this approach is often impossible to employ, since it requires to have training data of distorted images (i.e., noisy images in the aforementioned case), which need to be given labels (i.e., material or object categories), as well as to perform training on a larger dataset.

The second approach (Fig. 3(c)) is to make the CNN more robust to image distortion, so that it can correctly recognize distorted images even though it is trained only on clean images. This may be the most difficult one of the three approaches. In fact, there is only a few studies pursuing this approach. For example, Sun et al. [32] show that a type of activation functions mitigates decrease in recognition accuracy due to distortion of input images. However, the improvements are limited for real-world applications.

The third approach (Fig. 3(d)), which is the one we discussed in what follows, is to use another CNN model to restore quality of input images with distortions. We insert this CNN for image restoration before the CNN for classification (or other purposes); the first CNN estimate clean version of the input distorted image, which is fed to the second CNN for classification. We can use the CNN trained only on clean images for the second CNN. Instead, it is necessary to train the second CNN, which requires pairs of a distorted image and its clean version. On the other hand, it is not necessary to give the input distorted images labels for classification, which is advantageous. Moreover, the cost for creating the training data for image restoration (i.e., the second CNN) tends to be lower than the classification task.

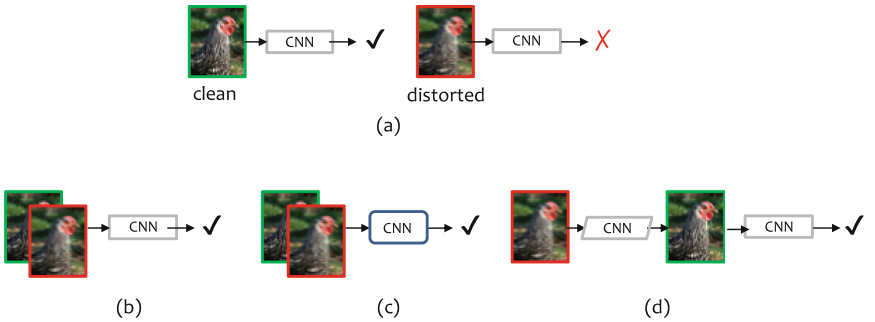

Fig. 3. Three ways to improve robustness of CNNs to image distortion. (a) A CNN trained only on clean images is vulnerable to distortion in input images. There are three methods to cope with this issue: (b) inclusion of distorted images in training data; (c) robustification of the CNN itself; and (d) cascading an image-restoration CNN to the CNN trained only on clean images.

Now, the problem is how to build a CNN that can perform image restoration with sufficient accuracy. We wish the first CNN to restore a clean image from an input distorted image and then the second CNN to correctly classify objects etc from the restored image. An example is the SEDNet shown in Fig. 2; it is a cascade of two CNNs (i.e., image restoration + classification) and it achieves slightly better performance than humans for inputs with large noises.

The problem of image restoration has been studied for a long time. In the past, researchers mainly tackle the problem by modeling natural images, where they consider their statistics based on edge statistics [4,16], sparse representation [1,25] etc. Recently, learning-based methods, particularly those using CNNs [8, 10] have shown better performance than those previous methods for the problems of denoising [21,24,28,29], deblurring [9,14,20], and super-resolution [2,11,30].

3 Image Restoration for Single Type of Distortions

We first consider the case where input images undergo a single type of distortion. This is a standard setting of image restoration, for which there have been a vast amount of studies conducted so far. Recent applications of CNNs have contributed to performance improvement. We have developed better architectural design of networks that can be shared across many tasks of image restoration. We briefly summarize the study here.

In the study, we pay attention to the effectiveness of paired operations on various image processing tasks. In [19], evolutionary computation is employed to search for optimal design of convolutional autoencoders for a few image restoration tasks; network structures repeatedly performing a pair of convolutions with a large- and small-size kernels perform well for image denoising. In [5], it is shown that a CNN iteratively performing a pair of up-sampling and down-sampling contributes to performance improvement for image-superresolution.

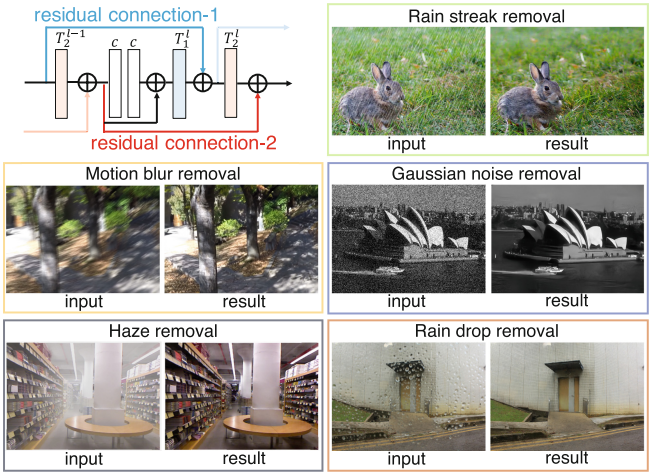

Fig. 4. Structure of the Dual Residual Block (DuRB) (upper left) and five popular image restoration tasks.

To accommodate such paired operation effectively, we propose a general architecture named Dual Residual Block (DuRN). A DuRN consists of an initial group of layers starting from the input layer, followed by an arbitrary number of blocks called Dual Residual Blocks (DuRBs), and the last group of layers ending at the output layer. Each DuRB has containers for the paired first and second operations. Normalization layers (such as batch normalization [6] or instance normalization [22]) and ReLU [15] layers can be incorporated when it is necessary.

In our experiments, we consider the five types of image distortions and restoration from them, as shown in Fig. 4. We design DuRBs for each of them; to be specific, we choose the two operations to be inserted into the containers T_1^l and T_2^l in the DuRBs. We have designed and used four different implementations, i.e., DuRB-P, DuRB-U, DuRB-S, and DuRB-US; DuRB-P are used for noise removal, rain-streak removal and raindrop removal, DuRB-U for motion blur removal, DuRB-S for raindrop removal, and DuRB-US for haze removal. For $[T_1^l, T_2^l]$, we specify [conv., conv.] for DuRB-P, [up-sampling + conv., down-sampling (by conv. with stride = 2)] for DuRP-U, [conv., channel-wise attention for DuRP-S, and [up-sampling + conv., channel-wise attention + down-sampling] for DuRB-US, respectively. We will show experimental results for noise removal, rain-streak removal, and motion-blur removal in what follows.

3.1 Noise Removal

We use DuRN-P for this task. Based on the findings in a study of neural architectural search [19], we choose convolution with large- and small-size receptive fields for T_1 and T_2, respectively. We also choose the kernel size and dilation rate for each DuRB so that the receptive field of convolution in each DuRB grows

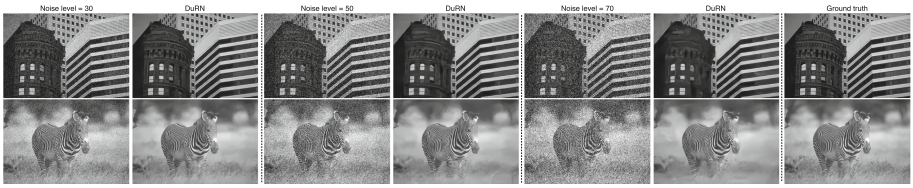

Fig. 5. Some examples of the results by DuRN-P for additive Gaussian noise removal. Sharp images can be restored from heavy noises ($\sigma = 50$).

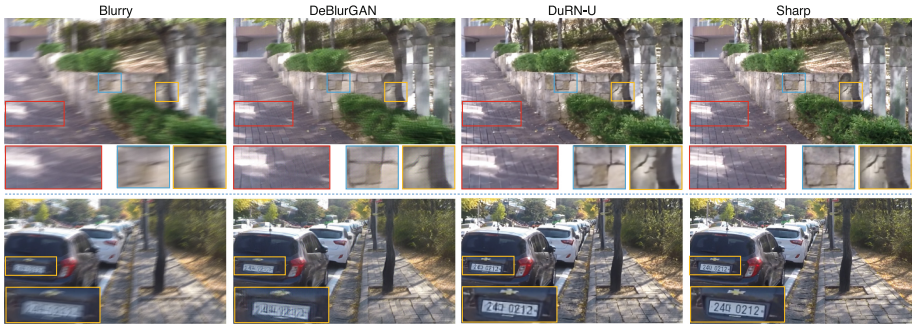

Fig. 6. Examples of motion blur removal on GoPro-test dataset.

its size with l. The entire network consisting of six DuRB-P's along with initial and final groups of layers is named DuRN-P. We train the network using l_2 loss (Fig. 5).

3.2 Motion Blur Removal

We use DuRN-U for this task. Following many previous works [9,23,27,31], we choose a symmetric encoder-decoder network for overall network structure. Then, following previous work [23] reporting the effectiveness of up- and down-sampling operations for this task, we employ the same operations for the paired operation. It achieves PSNR 29.9 dB and SSIM 0.91 on the GoPro-test dataset [14], while the state-of-the-art DeblurGAN (precisely, the "DeblurGAN-wild" introduced in the original paper [9]) achieved PSNR 27.2 dB and SSIM 0.95. Examples of deblurred images are shown in Fig. 6. It is seen that cracks on a stone-fence and numbers written on the car plate are restored well enough to be recognized.

Object Detection from Deblurred Images. We tested the two-step approach discussed earlier using the above CNN for motion blur removal. Given an image with motion blur, we first apply the above CNN, DuRN-U, to the input image and then use an object detector to the restored image. We follow the experimental procedure and dataset used in [9]. Note that DuRN-U is trained on the

GoPro-train dataset. For the object detector, we use YOLO v3 [17] trained on the Pascal VOC [3]. Table 1 shows quantitative results. The detection results for sharp images of the same YOLO v3 detector are used as the ground truths. It is seen that the proposed DuRN-U outperforms the state-of-the-art DeBlurGAN.

Table 1. Accuracy of object detection from deblurred images obtained by DeBlurGAN [9] and the proposed DuRN-U on Car Dataset.

	Blurred	DeBlurGAN[9]	DuRN-U (ours)
mAP (%)	16.54	26.17	31.15

4 Image Restoration for Combined Distortions

As explained above, there are many types of image distortion, such as various types of noises, defocus/motion blur, compression artifacts, haze, raindrops, etc. Thus, there are two cases for application of image restoration methods to real-world problems. One is the case where the user knows what image distortion need to be removed, e.g., a deblurring filter tool in a photo editing software. This is the case that we have considered above. The other is the case where the user wants to improve quality of an image but does *not* know what distortion(s) the image undergoes, e.g., applications to vision for autonomous cars or surveillance cameras.

We consider the second case here. Existing studies mostly consider the first case, which cannot be applied to the second case directly. However, real-world images usually suffer from a combination of different types of distortion, where we don't know mixture ratios and strengths of different distortion types in the input images. We need image restoration methods that work under such conditions.

There are only a few studies of this problem, such as Yu et al. [26]. The authors propose a method that adaptively selects and apply multiple light-weight CNNs; each CNN is trained for different image distortion. Their selection is done by an agent trained by reinforcemenet learning. However, the gain of accuracy obtained by their method is not so large, as compared with a dedicated method (CNN) for a single type of distortion.

We showed that a simple attention mechanism, named *operation-wise attention* layer, can better deals with the case of combined image distortions [13]. We propose a layer that performs many operations in parallel, such as convolution and pooling with different parameters. These operations are weighted by an attention mechanism built-in the layer, which is intended to work as a switcher of the operations. The attention weights are multiplied with the outputs of the operations, which are concatenated and transferred to the next layer. This layer can be stacked to form a multi-layer network that is fully differentiable. Thus, it can be trained in an end-to-end manner by gradient descent (Fig. 7).

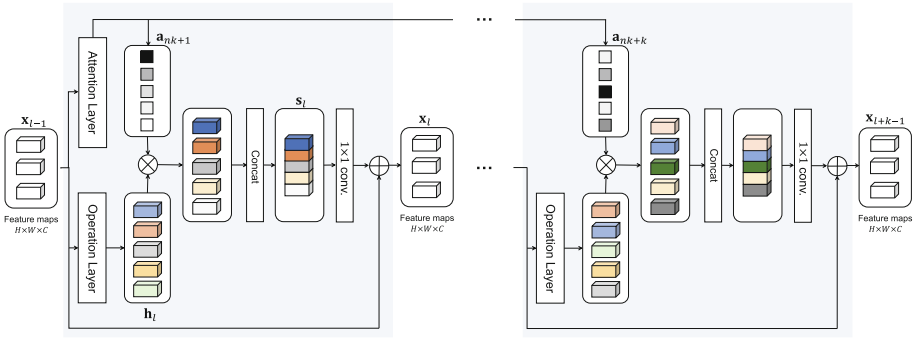

Fig. 7. Architecture of the operation-wise attention layer. It consists of an attention layer, an operation layer, a concatenation operation, and 1×1 convolution. Attention weights over operations of each layer are generated at the first layer in a group of consecutive k layers. Note that different attention weights are generated for each layer.

We evaluated the proposed approach by using the DIV2K dataset, which was created in [26] to evaluate their proposed method, RL-Restore. The dataset consists of about 0.3 million image patches of 63×63 pixels. They have multiple types of distortion, i.e., a sequence of Gaussian blur, Gaussian noise and JPEG compression with random levels of distortion. The proposed method achieves improvements of 0.3–0.7 dB (PSNR) and 0.015–0.02 (SSIM) over RL-Restore.

As in the aforementioned experiments on single type distortion, we evaluate the performance of the proposed method on the task of object detection. That is, we first restore an input image having combined distortion and then apply an object detector (SSD300 [12]) to the restored image. Employing the images from the PASCAL VOC detection dataset, we synthesize combined distortion of Gaussian blur, Gaussian Noise, and JPEG compression with random levels of distortion. The proposed method improves detection accuracy by a large margin (around 30% mAP) compared to the case of applying the same detector to the distorted images. It outperforms RL-Restore for almost all categories of objects. Figure 8 shows a selected examples of detection results. It is observed that the proposed method eliminate the combined distortion effectively and contributes to more accurate object detection.

5 Summary

We have discussed how to improve generalization ability of CNNs for visual recognition tasks, particularly in the case where input images undergo various types of image distortion. We have first pointed out that CNNs trained only on clean images are vulnerable to distortion in input images. This may be regarded as co-variate shift, the issue with CNNs or any other machine learning methods. We have further discussed the three possible approaches to the issue, i.e., (i) training the CNN also on distorted images, (ii) building more robust CNNs that can deal with distorted images, even if they are trained on clean images

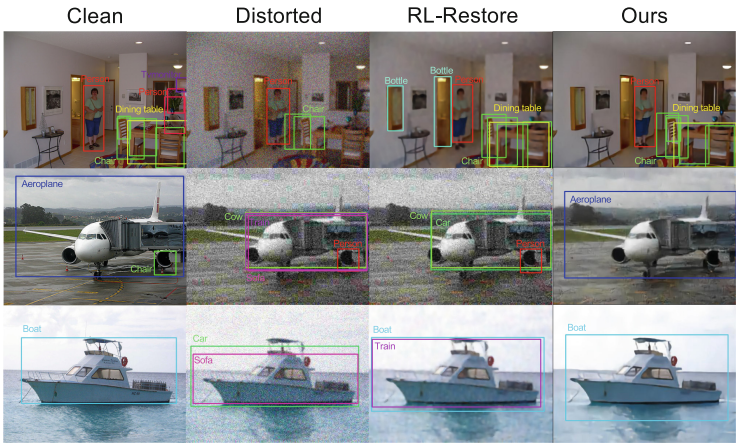

Fig. 8. Examples of results of object detection on PASCAL VOC. The box colors indicate class categories.

alone, and (iii) restoring a clean version from the input distorted image and then inputting it to a CNN trained only on clean images. We have then introduced two recent studies of ours that employ the third approach. The first study proposes an architectural design of CNNs for image restoration targeting at a single type of distortion. The second study proposes to use a single CNN that can restore clean image from an input image with combined types of distortion. The former provides better restoration performance but requires the type of distortion in input images to be identified beforehand. The latter is designed to be able to deal with unidentified type of image distortion, in particular, combined distortion of multiple types with unknown mixture ratios. We will be choosing between the two methods depending on conditions and requirements of applications. We believe that there will be room for further improvement in restoration accuracy, particularly for the second problem setting for unidentified distortion types. This will be studied in the future.

Acknowledgments. This work was partly supported by JSPS KAKENHI Grant Number JP15H05919, JST CREST Grant Number JPMJCR14D1, and the ImPACT Program "Tough Robotics Challenge" of the Council for Science, Technology, and Innovation (Cabinet Office, Government of Japan).

References

1. Aharon, M., Elad, M., Bruckstein, A.: k-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
2. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 184–199. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_13

3. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**, 98–136 (2015)
4. Fattal, R.: Image upsampling via imposed edge statistics. In: SIGGRAPH (2007)
5. Haris, M., Shakhnarovich, G., Ukita, N.: Deep back-projection networks for super-resolution. In: Proceedings of Conference on Computer Vision and Pattern Recognition (2018)
6. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of International Conference on Machine Learning (2015)
7. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of International Conference on Computer Vision (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Neural Information Processing Systems (2012)
9. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: DeblurGAN: blind motion deblurring using conditional adversarial networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2018)
10. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
11. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)
12. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part I. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
13. Suganuma, M., Lui, X., OKatani, T.: Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions. CoRR abs/1812.00733 (2018)
14. Nah, S., Kim, T.H., Lee, K.M.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of International Conference on Machine Learning (2015)
16. Perrone, D., Favaro, P.: Total variation blind deconvolution: the devil is in the details. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2014)
17. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv:1804.02767 (2018)
18. Sharan, L., Rosenholtz, R., Adelson, E.: Material perception: what can you see in a brief glance? J. Vis. **14**(9), 784 (2014)
19. Suganuma, M., Ozay, M., Okatani, T.: Exploiting the potential of standard convolutional autoencoders for image restoration by evolutionary search. In: Proceedings of International Conference on Machine Learning (2018)
20. Sun, J., Cao, W., Xu, Z., Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2015)
21. Tai, Y., Yang, J., Liu, X., Xu, C.: MemNet: A persistent memory network for image restoration. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)

22. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Instance normalization: the missing ingredient for fast stylization. arXiv:1607.08022 (2016)
23. Wieschollek, P., Hirsch, M., Schölkopf, B., Lensch, H.P.A.: Learning blind motion deblurring. In: Proceedings of International Conference on Computer Vision (2017)
24. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Proceedings of Neural Information Processing Systems (2012)
25. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Trans. Image Process. **19**(11), 2861–2873 (2010)
26. Yu, K., Dong, C., Lin, L., Change, L.C.: Crafting a toolchain for image restoration by deep reinforcement learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2018)
27. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: Proceedings of Conference on Computer Vision and Pattern Recognition (2018)
28. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. **26**(7), 3142–3155 (2017)
29. Zhang, K., Zuo, W., Zhang, L.: FFDNet: Toward a fast and flexible solution for cnn based image denoising. IEEE Trans. Image Process. **27**(9), 4608–4622 (2018)
30. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018, Part VII. LNCS, vol. 11211, pp. 294–310. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_18
31. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision (2017)
32. Sun, Z., Ozay, M., Zhang, Y., Liu, X., Okatani, T.: Feature quantization for defending against distortion of images. In: Proceedings of Computer Vision and Pattern Recognition, pp. 7957–7966 (2018)