Image Denoising Using Deep CGAN with bi-skip Connections

Peng Wang
SECE，Shenzhen Graduate School, Peking University
Shenzhen, China
145593758@qq.com

Abstract

With the rapid development of neural networks, many deep learning-based image processing tasks have shown outstanding performance. In this paper, we describe a unified deep learning-based approach for image image denoising. The proposed method is composed of deep convolutional neural and conditional generative adversarial networks. For the discriminator network, we present a new network architecture with bi-skip connections to address hard training and details losing issues. In the generative network, a objective optimization is derived to solve the problem of common conditions being non-identical. Through extensive experiments on image denoising task on both qualitative and quantitative criteria, we demonstrate that our proposed method performs favorably against current state-of-the-art approaches.

1. Introduction

Image processing is mainly concerned with extracting descriptions (that are usually represented as images themselves) from images. The analysis usually does not know anything about what objects are actually in the scene, nor where the scene is relative to the observer. There may be multiple, largely independent descriptions, such as edge fragments, spots, reflectances, line fragments, etc. As these descriptions are still linked to an image, these descriptions would apply everywhere in the image, not just to the mug.

As a fundamental task for medium-level or high-level processing, many Bayesian or regularization based conventional methods are utilized to fix many exiting processing issues. Most of problems can be solved satisfactorily. However, for the issue of images with a large amount of intricate noise still needs to be addressed. Recently, the superior performance based upon deep learning attracted attention from all fields. The outstanding performance has been demonstrated in many deep learning-based image processing applications. However, there still remains two areas that need to be explored: the lack of a robust and fast network architecture for more effective image processing, and a shortage of a unified network architecture for most of image denoising applications. Therefore, it is vital to seek a unified and effective machinery for image denoising.

To address the above difficulties, we propose a novel network architecture for image denoising using a deep conditional generative adversarial network (CGAN) with bi-skip connections and objective optimization. We observed the recent superior performance of deep learning-based image processing tasks, but we noticed that the hard training and details loss are in need of more attention. Therefore, taking into account the above problems, we propose a new symmetric bi-skip connections method to link convolutional and de-convolutional layers for the discriminator network, with which the training converges much faster and attains a higher-quality local optimum and image details can be recovered in deconvolution layers. Considering that the generative network has an excellent generating ability to recover image details of image denoising task and deriving that the existing of the non-identical optimization problem in CGAN.

In summary, our major contributions are briefly outlined as follows:

1) We propose a novel network architecture, which consists of a bi-skip connection based discriminator network and a objective optimization based discriminator network for image denoising is proposed in this paper.

2) We solve the problems of the hard training and details loss by proposing a new symmetric bi-skip connection, which helps to back-propagate the gradients to bottom layers and pass image details to the top layers, making training of the end-to-end mapping more easier and effective, and thus achieve performance improvement while the network going deeper.

3) We prove the existence of the non-identical optimization problem in CGAN, and we propose a multi-objective optimization to address the above issue. Relying on the new multi-objective optimization capacity and the generating ability of CGAN, most low-level image processing tasks have shown outstanding performance.

4) Experimental results demonstrate the advantages of the proposed network over other recent state-of-the-art methods on most of image denoising task, setting new records on these topics.

2. Related Work

2.1 Related Networks

The simple Encoder-Decoder model is mainly used to solve the problem of sequence to sequence conversion.

Image denoising is a challenging task that faces many problems in this area due to the lack of effective training data. The researchers proposed a "two-step" framework to train the noise distribution of the input image through GANs and generate rich training data using the generated noise samples. Based on which the denoising deep neural network was trained. The advantage of an auto encoder is that high-dimensional data can be converted to low-dimensional codes by training a multilayer network to reconstruct high-dimensional input vectors. That is, the features learned from networks are enough to describe the inputs. Based on this idea, some , such as Denoising Auto-encoder (DAE) [1] and Variational Auto-encoder (VAE) [2] have been used in image restoration. While it is head-scratching to extract proper features from blurred images.

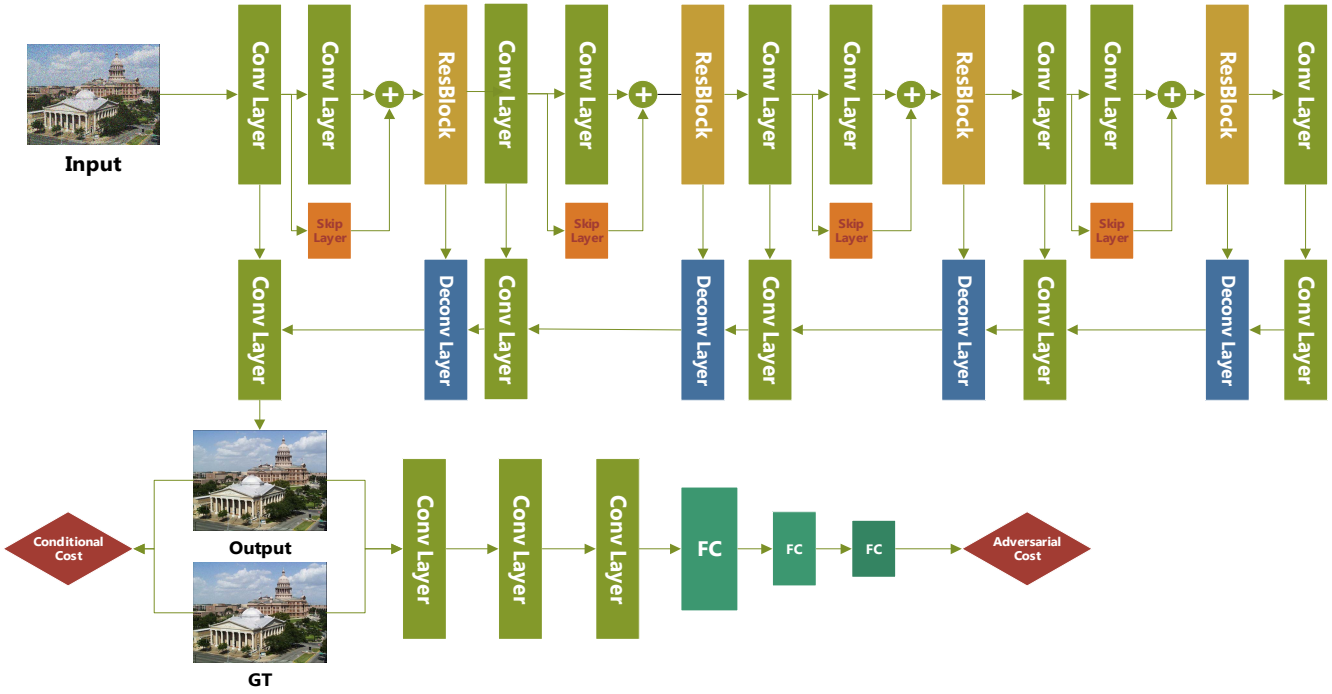

Figure 1. The framework of the proposed image processing adversarial networks, which gives an implementation of image denoising.

Image Priori-Based Method (BM3D) realizes the removal of unknown noise without training data by directly modeling noise images for image priori. At first, U-Net was proposed for image segmentation. It consists of a contracting path and an expansive path. The contracting path follows down sample steps and the expansive path follows up sample steps. Due to the symmetry of U-Net, it can localize the pixel, which makes features concatenating easy. Sip-Net proposed by Mao [3] mainly focuses on image denoising and super-resolution. Deriving from U-Net, Skip-Net links convolutional and deconvolutional layers with skip-layer connections. Furthermore, Ulyano [4] points out that untrained deep convolutional generators can be used to replace surrogate natural prior (TV norm) with dramatically improved results. All of the above shows that encoder-decoder has the ability to recover inputs with learned features, which is useful for low-level image processing tasks.

2.2 Conditional Adversarial Networks

Conditional Adversarial Networks are networks in which both the generator and discriminator are conditioned on some extra information. Isola[5], whose architecture is also known as pix2pix, proposed using cGAN in image-to-image translation. Unlike the vanilla version of GAN, cGAN learns a mapping from an observed image x and random noise z to the ground truth y with $G : x, z \rightarrow y$. In the cGAN architecture, the discriminator's job remains unchanged, but the generator not only fools the discriminator but also minimizes the divergence between the generated sample and the ground truth. As described above, the valuable insight of cGAN makes image-to-image translation more diverse and stable. Thus, some low-level image processing tasks, such as Kupyn [6] and Nah [7] use cGAN to deblur images in their models and output the very competitive results. Nah [7] proposes a multi-scale convolutional generator and presents the relative content loss to regularize the critic of cGAN. Kupyn et al. [6]

introduce perceptual loss which is obtained by some feature extractor in the improved WGAN. All of the above low-level image processing networks show that the cGAN that containing the specified generator has a unique strength to performing image denoising.

3. Proposed Approach

In this section, we propose a new deep CGAN architecture for the image denoising. The framework of the proposed method is illustrated in Figure 1.

3.1 Generator Network

The generator network is composed of multiple layers of convolution and de-convolution operators, and learns end-to-end mappings from images to tasks of the low-level image processing. The convolutional layers capture the abstraction of image contents. Deconvolutional layers have the capability to up-sample the feature maps and recover the image details. To deal with the problem of deeper networks tending to be more difficult to train,we propose symmetrically linking convolutional and deconvolutional layers with bi-skip-layer connections, with which the training converges much faster and attains better results, as shown in Figure 1. The linking of bi-skip connections from convolutional layers with their mirrored corresponding deconvolutional layers exhibit two main advantages. First, they allow the signal to be back-propagated to the bottom layers directly, and thus tackles the problem of gradient vanishing, making training deep networks easier and achieving processing performance gains a saconsequence. Second, these bi-skip connections pass image details from convolutional layers to deconvolutional layers, which is beneficial in recovering the image contents.

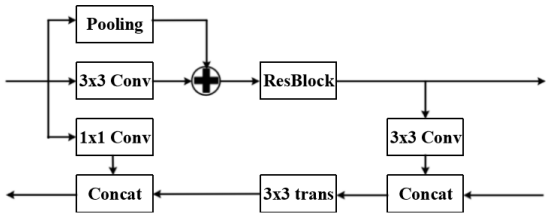

Figure 2. The illustration of the bi-skip-layer connections.

To better illustrate the bi-skip structure, we extract one scale (Figure 2) of the proposed generator network to illustrate the net. As shown in Figure 2, the contracting path consists of two subpaths in which one follows an average pooling and another follows 3×3 conv. The stride of the above two is 2 and follows an instance normalization layer and ReLU. Then the 3 × 3 conv path generates a residual and adds to the pooling path for downsampling followed by three ResBlocks. Each ResBlock consists of two repeated 3 × 3 conv, instance normalization layer and ReLU. The skipping path is composed of 1 × 1 conv which skipping to the same scale up-layer and 3×3 conv which skipping to the next scale up-layer. In the expansive path, the transposed layer is used to replace upsampling. An effective generator should digest the features as much as possible without increasing the depth of the network. The superiority of Bi-Skip-Net is that both shallow and deep priors can be concatenated for each scale. In our network, we explore 5 scales in Bi-Skip-Net and each scale is 2 times larger than the next. Besides, considering that residual is a very useful information both on the net optimization and the image details enhancement, we turn to use our Bi-Skip-Net to generate a residual adding to the inputs to output the image processing results and the framework is shown in Figure 1. The discriminator is designed by following the standard discriminator architecture. And the whole model is trained by CGAN with a specific content loss.

3.2 Discriminator Network

For the discriminator network, we use the content loss as the condition of the discriminator network. Given a target image x, our goal is to obtain a task-specific estimate of x based on the observed image y. We formulate the total loss as follows:

$$L_t = L_{adv} + uL_{cont}$$

Where L_{cont} is the content loss function which measures the divergence between x and our estimate x_e. A natural choice of the content loss function L_{cont} is L_2 loss $||x_e - x||_2^2$ which Nah [7] adopts it. However, L_2 loss will be much larger for outliers and will reduce the performance of the model. Another typical approach is to apply L_1 loss $||x_e - x||$,which is less sensitive to outliers and more robust when compared with L_2 loss. Unfortunately, these two-pixel losses will amplify noise in most cases. Kupyn [6] use perceptual loss $||f(x_e) - f(x)||_2^2$ to represent the content loss, which also has its pros and cons: it's capable of image processing well, but it cannot guarantee the structural consistency. What's more, x_e obtained from the perceptual loss may not be the ideal estimate of x even if x_e fulfills the condition $||f(x_e) - f(x)||_2^2 = 0$ perfectly. That is, although x_e satisfies $f(x_e) = f(x)$, it's not the case that $x_e = x$ as the feature extractor f is not a bijection.

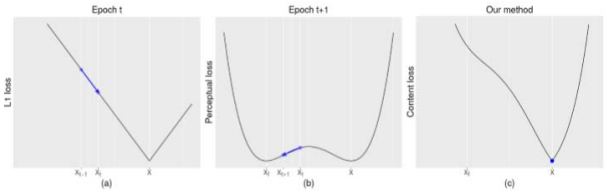

Figure 3. Problem caused by bi-level optimization. At epoch t, solution x_{t-1} tends to get close to x and we get x_t by the gradient of L1 loss function. Then at epoch t + 1, solution x_t tends to get close to x_f rather than x and we get x_{t+1} by the gradient of perceptual loss function. This fluctuation will plague neural network optimization and the solution may get stuck in the local minimum x_f or become divergent.

First, we apply a bi-level optimization to solve the above mentioned problem. It is known that the pixel loss can amply the image noise and that the perceptual loss cannot guarantee the structure consistency. It means that both of the two losses are non-identical with each other under the condition that the image must be identical with the image processing result. Thus, we propose a bi-level optimization to constrain the loss to satisfy the condition. Specifically, we make the two-level losses to interact with each other. The pixel loss can promote the perceptual loss tend to the condition and avoid the structure inconsistency. And the perceptual loss in turn to restrict the applied image noise. To balance the effects of the two losses, we normalize them to the same magnitude. The optimization consists of two steps that are base model training with MSE loss and the bi-level loss interaction. In the beginning, only the MSE is used as the content loss for the divergence between the input and the output is big in initial iterations. After some iterations, we turn to apply the bi-level loss interaction. L_1 loss is identical with MSE loss in value space and it has the ability to contract the noise space. Thus, we use L_1 loss and the perceptual loss to interact with each other. However, it may cause a serious problem for gradient-based optimization. As shown in Figure3(a) and (b)(here we illustrate with a one-dimensional signal), the interaction of L_1 loss and perceptual loss actually plagues the optimization and the solution may get stuck in the local minimum x_f or become divergent. Although x_f ' sperceptual loss is zero, it has high cost in fact considering the pixel level divergence between x_f and x, which poses a serious problem for gradient-based optimization algorithms.

Having considered the above issues, we innovatively propose another multi-objective optimization framework and formulate our content loss by combining L_1 loss and perceptual loss together:

$$L_{cont} = \alpha||x_e - x||+\beta||f(x_e) - f(x)||_2^2$$

where α and β are coefficients used to unify the magnitudes. To remain the similarity of content, pixel level loss and feature level loss both matters and sometimes they will conflict with each other: pixel level loss always amplifies image noise and damages the features of the image, while feature level loss could not guarantee the structure consistency in pixel level. Based on this motivation, our multi objective optimization is capable to obtain a trade-off between pixel level and feature level. What's more, as shown inFigure3(c),the local minimum x_f is eliminated and there exist only one global optimum x by combining L_1 loss and perceptual loss together.

For the sake of the significant part that the content loss plays in the training process where its gradient induces the direction of optimization, we have separated it from the total loss to discuss so far. The above-mentioned inducement to remain the similarity of the images' content is consistent with the direction of the adversarial loss which encourages the network to prefer solutions that reside on the manifold of images by trying to fool the discriminator network. In other words, the adversarial loss and the content loss supplement each other.

4. Experiments

For the experiments, we implement tasks for image denoising. The experiments have been performed on a desktop with i7-6700K CPU and NVIDIA GTX-1080 GPU. The experimental results are illustrated as follows.

4.1 Image Denosing

We evaluate the performance of our proposed network against several state-of-the-art mixed noise removal methods. We use BSD400 [8] for training and apply BSD68 and 12 commonly used images for testing. To generate mixed noise, two main types of mixed noise were considered: 1) AWGN + SPN and 2) AWGN + RVIN. We choose four state-of-the-art methods for comparison: MBM3D (BM3D coupled with median filter) [9], the $l_1 - l_0$ [10], and the WJSR [11] for comparison. The source codes of all compared methods are downloaded from the authors' websites. To evaluate the visual quality of the reconstructed images, Peak Signal to Noise Ratio (PSNR) and visual information fidelity (VIF)[12] are calculated to evaluate the visual quality. We apply l_2 commonly used images in our experiments. For AWGN + SPN, the AWGN with noise levels σ = 10,30 mixed with SPN with noise densities ρ = 20%,40% were considered. For AWGN + RVIN, the AWGN with noise levels σ = 15,25 mixed with RVIN with noise densities ρ = 25%,45% were considered. Tables 1 and 2 present the PSNR values of all these methods for all the tested images. These images are corrupted by the AWGN + SPN or AWGN + RVIN respectively. The best metrics calculated are marked in bold. As shown in tables 3 and 4, our proposed method achieves much higher PSNR values than other methods in most cases. The average values of PSNR of our method is also the highest.

Visual comparisons of denoising results are shown in Figure 4 and Figure 5. The proposed method can produce more clearly results.

Table 1. Comparison pf restoration results in PSNR for images corrupted byAWGN+SPN

		$\sigma = 10$				$\sigma = 30$			
Image	ρ	BM3D	$l_1 - l_0$	WJSR	Ours	M-BM3D	$l_1 - l_0$	WJSR	Ours
lena	20%	32.97	35.44	36.12	**36.45**	28.93	30.27	30.97	**31.61**
	40%	29.42	34.58	35.08	**35.44**	26.74	28.98	29.57	**29.78**
barbara	20%	24.07	31.37	32.28	**33.42**	23.28	26.66	27.13	**27.68**
	40%	23.28	29.41	31.19	**31.93**	22.29	25.65	25.89	**26.60**
boat	20%	28.00	31.23	31.94	**33.19**	25.70	27.04	27.54	**28.34**
	40%	25.19	30.20	30.81	**31.97**	23.98	26.19	26.56	**27.54**
bridge	20%	24.49	27.17	27.68	**29.50**	23.04	23.82	24.00	**24.81**
	40%	22.61	26.22	26.43	**27.70**	21.89	23.08	23.29	**24.17**
couple	20%	27.72	31.05	31.77	**33.16**	25.48	26.91	27.35	**28.06**
	40%	24.96	29.99	30.58	**31.85**	23.78	26.13	26.29	**27.36**
house	20%	35.91	37.38	37.95	**38.37**	31.17	32.28	33.42	**33.86**
	40%	32.52	36.83	37.26	**37.57**	28.35	30.54	31.65	**32.68**
F16	20%	30.23	32.94	33.94	**35.30**	27.25	28.52	29.15	**29.80**
	40%	26.70	32.15	32.91	**33.93**	25.12	27.14	27.85	**28.60**
peppers	20%	31.18	32.69	33.03	**34.56**	28.27	29.18	29.62	**30.64**
	40%	28.82	31.96	32.36	**33.28**	26.37	28.14	28.77	**29.73**
pentagon	20%	27.32	29.88	30.33	**30.40**	25.40	26.09	26.12	**26.20**
	40%	25.61	28.77	29.20	**29.31**	24.12	25.55	**25.59**	25.25
hill	20%	28.94	31.63	32.04	**32.95**	26.67	27.70	27.97	**28.55**
	40%	27.00	30.72	30.99	**31.90**	25.19	26.87	27.04	**27.76**
baboon	20%	25.74	28.92	31.00	**31.13**	23.54	25.20	25.43	**25.51**
	40%	22.89	27.38	29.24	**29.30**	22.13	24.26	24.46	**24.51**
man	20%	28.73	31.20	31.75	**33.33**	26.36	27.15	27.47	**28.31**
	40%	26.47	30.41	30.63	**32.04**	24.76	26.34	26.54	**27.63**
Average	20%	28.78	31.74	32.49	**33.48**	30.90	27.57	28.01	**28.61**
	40%	26.29	30.72	31.39	**32.19**	24.56	26.57	26.96	**27.63**

Table 2. Comparison pf restoration results in PSNR for images corrupted byAWGN+SPN

		$\sigma = 15$				$\sigma = 25$			
Image	ρ	BM3D	$l_1 - l_0$	WJSR	Ours	M-BM3D	$l_1 - l_0$	WJSR	Ours
lena	25%	29.79	29.95	31.91	**32.31**	26.59	26.08	28.46	**29.05**
	45%	22.62	26.25	28.47	**29.52**	20.15	23.26	25.61	**27.71**
barbara	25%	23.79	24.02	24.37	**24.89**	22.78	22.46	23.24	**24.13**
	45%	21.45	22.44	23.07	**23.97**	18.26	21.11	22.39	**24.00**
boat	25%	27.08	26.86	27.60	**28.87**	24.06	24.45	25.36	**26.12**
	45%	22.88	24.58	25.58	**25.87**	19.53	22.57	23.91	**25.19**
bridge	25%	23.83	24.19	24.21	**24.38**	21.80	22.64	22.65	**22.91**
	45%	20.72	23.37	22.60	**23.86**	18.45	20.90	21.40	**22.48**
couple	25%	26.72	26.78	27.43	**28.96**	23.91	24.37	25.18	**25.56**
	45%	23.00	24.46	25.38	**25.95**	19.45	22.50	23.65	**24.12**
house	25%	31.54	31.06	34.45	**34.98**	28.68	26.72	31.13	**32.53**
	45%	23.31	27.33	30.74	**31.77**	20.99	24.13	27.76	**29.07**
F16	25%	28.12	28.53	29.92	**30.45**	24.87	25.29	27.12	**28.36**
	45%	21.63	25.21	26.82	**29.37**	19.29	22.49	24.49	**26.10**
peppers	25%	28.95	28.90	30.30	**31.03**	26.60	25.71	28.04	**28.97**
	45%	23.07	25.98	27.71	**29.13**	20.64	23.22	25.53	**27.05**
pentagon	25%	26.14	26.36	**26.55**	26.41	24.65	24.21	24.82	**24.89**
	45%	24.69	25.04	25.45	**25.56**	21.97	23.05	23.90	**24.03**
hill	25%	27.75	27.74	28.52	**28.97**	25.59	25.08	26.30	**27.32**
	45%	23.24	25.53	26.71	**26.97**	20.73	22.98	24.84	**25.52**
baboon	25%	25.44	25.83	25.86	**26.37**	22.30	23.09	23.23	**23.98**
	45%	22.93	23.75	24.01	**24.71**	20.11	21.98	22.15	**22.74**
man	25%	27.51	27.44	28.22	**28.95**	25.02	24.83	25.95	**27.03**
	45%	23.22	25.26	26.27	**27.71**	20.62	22.96	24.42	**25.03**
Average	25%	27.22	26.47	28.28	**28.88**	24.74	24.58	25.96	**26.74**
	45%	22.73	24.93	26.07	**27.03**	20.02	22.60	24.17	**25.25**

Figure 4. Restoration results of Barbara image (GN + SPN, α = 20; ρ = 0.2). From left to right and top to bottom: original image, noise image, the reconstructed images by M −BM3D (PSNR = 23.64dB), l1−l0 (PSNR = 29.41dB), WJSR (PSNR = 30.01dB), and our proposed method (PSNR = 31.20dB).

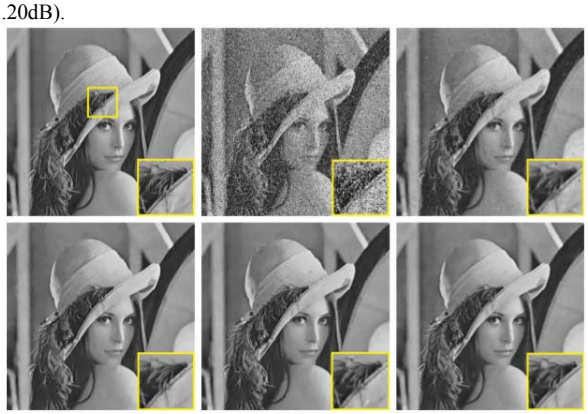

Figure 5. Restoraion results of Lena image (GN + SPN, α = 20; ρ = 0.2). From left toright and top to bottom: original image, noise image, the reconstructed images by M −BM3D (PSNR = 27.61dB), l1−l0 (PSNR = 31.31dB), WJSR (PSNR = 31.69dB), and our proposed method (PSNR = 32.61dB).

5 Conclusion

In this paper, we unify image denoising task with a novel deep conditional generative adversarial network(CGAN) model using bi-skip connections and multi objective optimization. We demonstrate that with symmetric bi-skip as connections to integrate both deep and shadow features in our generator network, we achieved more accurate image processing results. Furthermore, we design two new optimizations, which are proved that multi-objective optimization is better than bi-level optimization, to solve the problem of common conditions being non-identical in our discriminator network, which constrain the loss to obey the condition and achieve a better performance. Extensive experimental analysis on image denoising task, , and comparisons to state-of-the-art method show that our proposed network outperforms existing approaches with a wide margin.

References

[1] Vincent P , Larochelle H , Lajoie I , et al. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion[J]. Journal of Machine Learning Research, 2010, 11(12):3371-3408.

[2] Kingma, Diederik P., and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).

[3] Mao X, Shen C, Yang Y B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections[C]//Advances in neural information processing systems. 2016: 2802-2810.

[4] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep Image Prior. arXiv preprint arXiv:1711.10925 (2017).

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-toimage translation with conditional adversarial networks. arXiv preprint arXiv:1611.07004 (2016).

[6] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas (2017) DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. arXiv preprint arXiv:1711.07064.

[7] SeungjunNah,TaeHyunKim,and Kyoung MuLee (2017) Deep multi-scale convolutional neural network for dynamic scene deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 2.

[8] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[9] Rolf Khler, Michael Hirsch, Betty Mohler, Bernhard Schlkopf, and Stefan Harmeling (2014) Recording and playback of camera shake: Benchmarking blind deconvolution with a realworld database. In European Conference on Computer Vision. Springer, 2740.

[10] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce (2015)Learning a convolu-tional neural network for non-uniform motion blur removal. In Computer Vision and Pattern Recognition. 769777.

[11] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang (2016) Deep Video Deblurring. arXiv preprint arXiv:1611.08387.

[12] Zhang Y , Zhu C , Li G , et al. Bi-Skip: A Motion Deblurring Network Using Self-paced Learning[J]. 2019.