Dynamic Scene Deblurring Based on Semantic Information Supplement

Yiming Liu, Junhui Li, Wenzhuo Huang, Kang Tang and Dahong Xu*

College of Information Science and Engineering, Hunan Normal University

Email: *xudahong@hunnu.edu.cn

Abstract. To solve the problem of semantic information dilution in network propagation, a semantic information supplement mechanism (SIS) is proposed to improve the performance of dynamic scene deblurring algorithm. Based on GANs structure, our generator is to recycle the semantic information and features spanning across multiple receptive scales to restore a sharp image, when a blur image is given. What’s more, in order to better integrate semantic information with the latent-feature and solve the problem of training difficulty in very-deep network, we put forward a long and short skip-connection method. Extensive experiments show that our Semantic Information Supplement network (SIS-net) achieves both qualitative and quantitative improvements against state-of-the-art methods.

1. Introduction

Since spatially variant blur is the combined effect of multiple factors, such as camera shake and defocusing, dynamic scene deblurring is more realistic and challenging compared with uniform blind deblurring. There are a lot of effective algorithms. The multi-scale [1], scale-recurrent architecture [2], and a U-net like structure [3] exploit the deblurring cues at different processing levels. However, there still exists two major challenges in deep deblurring: One is that high-level semantic information is progressively propagated to shallower layers, and hence the semantic information captured by deeper layers maybe gradually diluted in the “coarse to fine” network process. Another one is as the increase of neural network depth, the difficulty of network training becomes more difficult, which leads to the problem of gradient explosion and vanishing.

In this paper, we propose not only a Semantic Information Supplement mechanism (SIS) aiming at the problem that the high-level semantic information is diluted in the propagation, but also a skip-connection method to solve the problem of network training and make a easy way of SIS aggregation.

2. Related Work

2.1 Dynamic Scene Blurring

In recent years, inspired by the traditional “coarse-to-fine” optimization framework, Nah, et al.[1] exploited a multi-scale network to restore sharp images from images whose blur is caused by various factors. A multi-scale loss function is employed to mimic the “coarse-to-fine” pipeline in conventional deblurring approaches. Tao, et al. [2] introduce the ConvLSTM module to share parameters from coarse-to-fine scales. Kupyn, et al. [4] proposed a Generative Adversarial Network (GAN) architecture to tackle this Motion Deblurring.
2.2 Semantic Information
In order to make use of more valuable image prior and have the ability to preserve details, the semantic information has gradually emerged into image deblurring[3,5,6]. To take advantage of multi-scale feature information, Orest, et al. [3] also introduces FPN [5] network into the generator, so that low-level content information can be integrated with high-level semantic information integration, providing more reference information for image deblurring without losing details. Liu, et al. [6] uses sophisticated pooling techniques that fuse high-level semantic information with low-level feature layers and successfully solves the problem of semantic loss in salient object detection. However, it is not suitable for image deblurring.

2.3 Skip Connections
With the deepening of the neural network, the effectiveness of training is seriously affected by the disappearance of the gradient. Therefore, a large number of architectures have been proposed to solve this problem. After the theory of U-net [7] was proposed, the skip connection mode between the corresponding encoder and decoder stages in the problem of pixel regression in image restoration was used as an effective framework [8, 9]. Jump connections, which are a key factor in solving the vanishing gradient problem, are connections that add input to output after two or more convolutional layers. So it's easy to build and optimize very deep networks. Skip connections also link from the bottom to the top to implement functional mapping. This scheme [1,2] achieves better detail reconstruction, which allows more flexible back-propagation of information and allows image details to be passed from the bottom to the top.

3. Proposed Method
The model architecture of this paper is based on DeblurGAN-v2 architecture [3]. The generator uses U-net [7] and integrates FPN mode [5]. Compared with the original model, this paper adds two parts: semantic box and skip connections. The architecture is illustrated in Figure 1. The generator restores a sharp image $I_i$ from a blurred image $B_i$.

![Figure 1. network architecture](image)

3.1 Model Architecture
Based on a generative adversarial networks structure, Our architecture consists of a skip connected generator and two discriminators. In our generator, the contracting path and the expanding path follow the FPN, who is a lightweight alternative to multi-scale features. The FPN backbone reserve latent-feature maps at five different scales. Those features which aggregate different levels of semantic information are later up-sampled to the same $\frac{1}{4}$ size of input and concatenated into one tensor.
Similar to [10], we designed two types of skip connections, which can make better use of semantic. In addition to the multi-scale feature aggregation capability, choosing FPN as generator also achieves a balance between accuracy and speed: please see our experiment parts. Two discriminators were set respectively responsible for the discrimination of the global ambiguous degree and the local patch ambiguous degree.

3.2 Semantic Information Supplement mechanism (SIS)
To solve the problem that the semantic information of FPN is gradually diluted with the increase of fusion times, we design a supplementary semantic information storage structure. This structure is located at the top of generator, and plays an significant role in feature propagation. The feature map of the final layer in the contracting path is considered as semantic information. In order to integrate the semantic information into the layers of different scales in the expanding path, we do four sets of up sampling of different scales for the semantic information blocks. In Figure 1, the “infos” section is the semantic information. This mechanism is named as semantic information supplement (SIS). At the same time, we design two types of skip connections, which use semantic information in the expanding path. See Section 3.3 for details.

3.3 Skip-connection
In this paper, we design two types of skip connections: long connection and short connection. In Figure 1, the green line represents the long connection which can solve the problem of global training difficulty, while the blue line means a short connection which can solve the problem of difficult local training. In the expanding path of generator, the feature map is up-sampled, and the sharp image is gradually restored. In this process, the short connections integrates local SIS into each feature map. In the final layer in the expanding path, the global multi-scale SIS is integrated into the feature map through the long connection. The detailed principle of aggregation can refer to the algorithm pseudo code in Table 1.

| Table 1. Algorithm Solving info aggregation |
|------------------------------------------------|
| Input: semantic information (info) and latent-feature (latent) |
| Output: merge[k] |
| 1: merge[0] = info[0] + latent[1] |
| 2: for k in range(1,3):
| merge[k] = merge[k-1] + info[k] + latent[k] |
| 3: for i in range(1,3):
| merge_smooth +=merge[i]
| merge[4] = info[4] + latent[4] + merge_smooth |

3.4 Loss function
GANs for training image recovery, it is difficult to optimize the objective function, due to these problems such as model collapse and exploding/vanishing gradient ,we choose the WGAN Loss[11] proposed by Arjovsky et al. What's more, one needs to compare the image of the reconstructed and the ground-truth ones according to some metric on the training stage. Generally, one uses the loss function of per-pixel to measure the diversity of the reconstructed and the ground-truth images. However, the loss of per-pixel don't capture different perception between the reconstructed and the ground-truth images. For example, consider two identical images of offsetting by one pixel from each other. Although they are similar in perception, they are quite different in terms of loss per pixel. In parallel,
the latest work shows that using perceptual loss functions can generate high-quality images based on the diversity of high-level image feature representations extracted from pre-trained convolutional neural networks.

In this paper we combine two approaches to replace the loss function in DeblurGAN-v2. When training networks for image restoration tasks, we not only use per-pixel loss functions, but also use perceptual loss functions [12] that compare relative images on high-level features from a pre-trained network. During the training process, the loss of perception can cover the shortage of per-pixel losses, and during the test-time, the restoration network with light parameters runs in real-time.

\[
L_{\text{MSE}} = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} ((I_i)_{x,y} - G_{\theta G}(B_i)_{x,y})^2
\]  

(1)

\[
L_{\text{percep}} = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} (\phi(I_i)_{x,y} - \phi(G_{\theta G}(B_i))_{x,y})^2
\]  

(2)

Here, w,h are the width and height of the feature map which obtained by the ReLU 3_3 layer of VGG-16 network with the parameter of \( \phi \).

\[
L_{\text{WGAN}} = \sum_{i=1}^{I} -D_{\theta d}(G_{\theta G}(B_i))
\]  

(3)

In total, our network loss function combined as follows:

\[
L_g = 0.5 \times L_{\text{MSE}} + 0.006 \times L_{\text{percep}} + 0.01 \times L_{\text{WGAN}}
\]  

(4)

4. Experimental evaluation

4.1 Training detail

We use GOPRO dataset to train our network according to the same strategy as [1]. The network contains 3214 blurry and clear image pairs, of which 2103 are trained and 1111 are used for evaluation. The experiment performed 2500 epochs with a 0.0001 learning rate which decays 0.000001 each iteration after 500 epochs. Our model, using Adam as an optimizer, is trained on a single GTX2080Ti and takes for a total of 14 days for convergence. Our model can receive images of any size because of the completely convolutional structure.

4.2 Effectiveness of SIS and Skip Connections

In this paper, we first design long/short connection in the expanding path of U-net to supplement semantic information repeatedly. The purpose is to enhance the restoration effect and solve the training problem. In this section, we divide the model into four different structures to demonstrate the effectiveness of semantic information supplement mechanism and long/short connection. For the reliability of experimental results, we use the same training method and the same data set. Model Plain only have GAN structure without SIS and skip connection. Model NoSIS-net keeps long connections, but does not use semantic boxes. Model NoSkip-net removed skip connection mechanism from the original network. Quantitative comparison results can be seen in Table 2, which indicates that the complete model (SIS-net) has better performance, and also indicates that the effectiveness of SIS and long/short connection.
Table 2. PSNR and SSIM comparison on the GoPro test dataset.

| Model          | Plain | NoSIS-net | Noskip-net | SIS-net (complete) |
|----------------|-------|-----------|------------|--------------------|
| PSNR           | 27.44 | 28.9      | 28.5       | 30.24              |
| SSIM           | 0.823 | 0.851     | 0.842      | 0.902              |

4.3 Quantitative Evaluation on GoPro Dataset

We compare our SIS-net with recent state-of-art methods on the GoPro dataset: two of are traditional method [13,14], while the rest are deep learning-based: [1] by Nah et al., [15] by Sun et al., [16] by Gong et al., SRN [2], DeblurGAN [4], and DeblurGAN-v2 [3]. We compare on standard performance metrics (PSNR, SSIM). Results are summarized in Table 3.

In terms of PSNR/SSIM, SIS-net (ours) and SRN are ranked top-2: SIS-net (ours) has slightly lower PSNR, but it outperforms DeblurGAN-v2 (Inception- ResNet-v2) in SSIM. However, we are very encouraged to observe that SIS-net (ours) takes 80% less inference time than SRN. The traditional methods [14] based on the prior assumption does not get good results. Nah et al. [1] and Tao et al. [2] use multi-scale scheme to remove motion blur, however, due to the lack of sharing mechanism, it [1] results heavy network parameters and long running time. Orest Kupyn et al. [4] builds a generative adversarial model to remove ambiguity, which demonstrated very good results and shorter execution time. Our network presents a good accuracy and can do real-time deblurring. For pictures with a resolution of 720P, our average run-time reaches 0.303 seconds on our platform, which means it supports real-time image deblurring.

Table 3. Quantitative comparison on the GoPro test dataset

| Methods        | Xu et al. [14] | Sun et al. [15] | Nah et al. [1] | Tao et al. [2] | Gong et al. [16] | DeblurGAN [4] | Li, et al. DeblurGANv2 [3] | SIS-net (ours) |
|----------------|----------------|-----------------|---------------|---------------|------------------|----------------|--------------------------|----------------|
| PSNR           | 25.1           | 24.64           | 29.08         | 30.10         | 26.06            | 26.435         | 27.08                    | 29.55          | 30.24                      |
| SSIM           | 0.894          | 0.842           | 0.913         | 0.932         | 0.863            | 0.892          | 0.857                    | 0.934          | 0.902                      |
| Time           | 13.41s         | 20min           | 4.33s         | 1.31s         | 20min            | 0.92s          | 1.7min                   | 0.35s          | 0.303s                     |

4.4 Subjective Evaluation based on perception

In order to evaluate several models from the perspective of perception, canny edge detection is performed on the sharp image restored by each model: SRN, DeblurGAN, DeblurGAN-v2, and SIS-net (ours). Figure 2 displays visual examples on the GoPro dataset.

DeblurGAN, DeblurGAN-v2, and SIS-net (ours) effectively restore the edges and textures, without noticeable artifacts. But, SRN for this specific example shows some color artifacts when zoomed in, and SRN also lost more edge pixels. DeblurGAN and DeblurGAN-v2 have better edge pixels preserving ability than SRN, but it gets wider edge pixels. This indicates that the edge part of the object may have a higher degree of blur. From the perspective of edge detection results, SIS-net (ours) is outperforms SRN, DeblurGAN and DeblurGAN-v2 in terms of edge pixel preserving ability and edge width refinement ability.
4.5 Quantitative Evaluation on PSNRs-FLOPs trade-off

In Figure 3, the PSNRs-FLOPs trade-off plot on the GoPRO dataset. Compared to four state-of-the-art competitors (in blue): DeblurGAN[4], Scale-Recurrence Network(SRN)[2] and Deepblur[1], Our models (in red) is shown to achieve superior or comparable quality, and is much efficient.
5. Conclusion

This paper introduces SIS-net, a powerful and efficient dynamic scene deblurring framework, with promising quantitative and qualitative results. Dynamic scene deblurring is realized by this algorithm. Because of its real-time and light-weight computing, it not only can be transplanted to the intelligent terminal to realize the anti-shake of the camera of the intelligent terminal, but also can be used to overcome the defocusing blur caused by the long exposure time of the camera and improve the results of the camera of the intelligent terminal. At the same time, the algorithm can also be used as a preprocessing in other visual tasks to improve the performance of visual tasks, such as object detection, segmentation, classification, action recognition and so on. We plan to improve the supplementary mechanism of semantic information in SIS-net to achieve better detail preserving ability of the model.

References

[1] Nah, S., Hyun Kim, T., and Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. 2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 3883-91.
[2] Tao, X., Gao, H., Shen, X., Wang, J., and Jia, J. Scale-recurrent network for deep image deblurring. 2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 8174-82.
[3] Kupyn, O., Martyniuk, T., Wu, J., and Wang, Z. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. 2018 Proceedings of the IEEE International Conference on Computer Vision pp 8878-87.
[4] Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., and Matas, J. Deblurgan: Blind motion deblurring using conditional adversarial networks. 2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 8183-92.
[5] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. 2017 Proceedings of the IEEE conference on computer vision and pattern recognition pp 2117-25.
[6] Liu, J. J., Hou, Q., Cheng, M. M., Feng, J., and Jiang, J. 2019 A Simple Pooling-Based Design
for Real-Time Salient Object Detection. arXiv preprint arXiv:1904.09569.

[7] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention pp 234-41.

[8] Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. Enhanced deep residual networks for single image super-resolution. 2017 Proceedings of the IEEE conference on computer vision and pattern recognition workshops pp 136-44.

[9] Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. 2018 Proceedings of the European Conference on Computer Vision (ECCV) pp 286-301.

[10] H Gao, X Tao, X Shen, J Jia. Dynamic Scene Deblurring with Parameter Selective Sharing and Nested Skip Connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[11] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. ArXiv e-prints, Jan. 2017. 1, 3, 4, 5, 7

[12] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, pages 694–711. Springer, 2016

[13] Li, L., Pan, J., Lai, W. S., Gao, C., Sang, N., and Yang, M. H. Blind image deblurring via deep discriminative priors. 2019 International Journal of Computer Vision, pp 1-19.

[14] Xu, L., Zheng, S., and Jia, J. Unnatural l0 sparse representation for natural image deblurring. 2017 Proceedings of the IEEE conference on computer vision and pattern recognition pp 1107-14.

[15] Sun, J., Cao, W., Xu, Z., and Ponce, J. Learning a convolutional neural network for non-uniform motion blur removal. 2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 769-77.

[16] Gong, D., Yang, J., Liu, L., Zhang, Y., Reid, I., Shen, C., ... and Shi, Q. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. 2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp 2319-28.