Dynamic Scene Deblurring of Multi-Scale Progressive Attention Network

Wenzhuo Huang\textsuperscript{1,2,b}, Yueming Deng\textsuperscript{1,2,r,a}, Yiming Liu\textsuperscript{1,c}, Jun Wu\textsuperscript{1,2,d} and Xiaojun Li\textsuperscript{1,e}

\textsuperscript{1}College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
\textsuperscript{2}Hunan Xiangjiang Artificial Intelligence Academy, Changsha 410081, China

E-mail: \textsuperscript{a}dengyueming@hunnu.edu.cn; \textsuperscript{b}walterzhuo@163.com; \textsuperscript{c}516148514@qq.com; \textsuperscript{d}964299463@qq.com; \textsuperscript{e}1332863830@qq.com

Abstract. In order to enhance the attention to the image foreground targets in dynamic scenes and better suppress the generation of image edge artifacts, a multi-scale progressive attention network (MSPA) deblurring algorithm is proposed. MSPA is based on the GAN structure. In the feature extraction stage, MSPA designs multi-scale width perception and pyramid perception residual blocks to help our skeleton network better extract multi-scale local features and reduce the difficulty of network training, and In order to better obtain the semantic information of the image, the channel attention mechanism is used to correct the deep features of the blurred image; in the feature fusion stage, the progressive channel and spatial attention network (PCSA) is designed to enhance the receptive field, and selectively integrate multiple levels semantic information and enhance the non-local connection between features in progressive method. A large number of experiments show that the MSPA algorithm is superior to many advanced algorithms on the GoPro and Kohler datasets.

1. Introduction

With the advent of the mobile intelligence era, people have more requirements for high-definition technology for deblurring images in dynamic scenes. Recently, a method based on deep learning has been proposed that using external data and using an end-to-end trainable CNN to facilitate the deblurring process, and it has shown encouraging results. In the network design of deep learning, it is not difficult to find that the design from coarse to fine can achieve good results. For example, Deep multi-scale\textsuperscript{[1]}, SRN\textsuperscript{[2]}, DeblurGAN\textsuperscript{[3]} and other network structures all import a large number of parameters, so it runs slowly. This paper chooses DeblurGAN-V2 \textsuperscript{[4]} as the framework network, but we use an improved FPN to replace the generator part of the framework network. Because FPN has an intermediate layer, it can not only use multi-scale local features, but also reduce the weight of the network and increase the speed of the network.

In order to enhance the attention to the foreground target and better suppress the generation of image edge artifacts, this paper uses the semantic information provided by the channel attention correction module at the top of the feature extraction layer, the features of the upper layer of the feature fusion layer and the channel attention correction module and the spatial attention correction module to form an attention correction network (CSA) and output $F_{CSA}$. Since we need to use up-sampling 4 times in the feature fusion stage, the attention correction
network is used many times to form a progressive channel and spatial attention network (PCSA), as shown in Figure 1.

Deep multi-scale shows that the deeper the network, the stronger the ability to extract features, due to deep network training, the gradient disappears and explodes. According to the enlightenment of literature [4], in the feature extraction stage of the generated network, we use the residual network as our backbone network. There are two reasons: First, because the blurred and sharpened image pairs are similar in value, the blurred features extracted by letting the parameters understand the differences between them are effective. Second, the residual network can improve the effect of blurred feature extraction. However, simply stacking residual modules at will to achieve the effect of deepening the number of network layers is invalid, and the prototype of the residual network is used to solve advanced vision tasks. Since image restoration such as deblurring or super-resolution is a low-level vision task, the residual module needs to be improved. As shown in Figure 2 and Figure 3, this paper designs the width perception residual block and the pyramid perception residual block, and achieves a better deblurring effect through aggregation.

FPN has the phenomenon of semantic information degradation, and the response of different feature channels in the feature map of network deepening is different, so it is necessary to filter the multi-scale features obtained from top-down coding. Refer to the previously published article (SIS algorithm) [5], this article improves the channel attention correction mechanism, which not only chooses the powerful features of semantic information, but also supplements the attention correction mechanism as semantic information (as shown in Figure 4).

In this article, we explore a more effective structure to blur in the dynamic scene. The following is the main content of this article:

- In order to reduce network parameters and increase the speed of the network, this paper improves the residual block, and designs the width perception and pyramid perception residual blocks.

- To enhance the attention to the foreground targets in the dynamic scene and better suppress the generation of image edge artifacts, this paper designs a progressive channel and spatial attention network (PCSA).

- In order to alleviate the phenomenon of semantic information degradation in FPN, inspired by the article we have published before (SIS algorithm), we use the channel attention mechanism to correct the deep features of the blurred image, and choose strong semantic features to supplement the progressive channel and spatial attention network (PCSA).

2. Model of Multi-Scale Progressive Attention Correction Network

In Figure 1, the left side is the generator, and the right side is the two identical discriminators. The generator module adopts the U-Net structure and is mainly composed of 3 parts (the feature extraction layer, the middle layer and the feature fusion layer). The discriminator adopts a dual-scale structure, which can make full use of global and local features.

2.1. Improved Residual Network

Regarding the feature extraction of residual deep network, the skeleton network has 5 feature extraction sub-modules of different scales, and each single-scale feature extraction sub-module is composed of 3 modules (a convolution block for downsampling, a width perception residual block, a pyramid perception residual block).
Figure 1. MSPA network structure.

2.1.1. The Width Perception Residual Block In this paper, inspired by the literature [4], the width perception residual block is designed (as shown in Figure 2). The difference with its traditional residual module is that in the feature extraction process, we double the number of convolution kernels so that the channel number of the feature map in the propagation process also doubles. In this way, the width perception residual module is constructed into a vase-like structure, and the number of channels of the convolutional layer involved in the transmission process is twice the number of input and output channels. This construction method can make the skeleton network better retain the transmitted blurred information and reduce the difficulty of network training.

2.1.2. The Pyramid Perception Residual Block The the pyramid perception residual block only add a multi-scale mechanism on the basis of the width perception residual block, as shown in Figure 3. Because the multi-scale convolution can better ensure the detailed information and expand its receptive field to a greater extent. It uses 3 convolutional layers of different scales (7×7, 5×5, 3×3) to convolve the output features capture the contextual perception information of the multi-receptive field, and then concatenate the three feature maps obtained by convolution kernels of different sizes to obtain multi-scale perception features.

Figure 2. The width perception residual block.  
Figure 3. The pyramid perception residual block.

2.2. Progressive Channel and Spatial Attention Network (PCSA) In order for the deep features to acquire with strong semantic information, and for the shallow features to retain rich spatial details, it is necessary to combine the multi-level features. However,
due to redundant details and background interference, it is flawed to use FPN-net to perform feature fusion without distinction. So we design a progressive channel and spatial attention network (PCSA), which gradually decodes multi-level context information in order to better focus on the foreground area and suppress the generation of image edge artifacts. As shown in Figure 4, the progressive CSA network is the PCSA network. Since the generator in this article is based on FPN, which has the phenomenon of semantic information degradation, the feature uses Channel attention mechanism (CA) to select channel features with strong semantic information to provide semantic supplements to the CSA network structure in the feature fusion stage. F is the output of the previous layer feature.

![Figure 4. CSA network structure.](image)

### 2.3. Algorithm Flow

As shown in Figure 1, the blue area in the generator network is the feature extraction layer. The blurred image in the dynamic scene sequentially extracts the shallow and deep features of the image through 5 residual blocks of different scales (extract1 ~ extract5). The output of the previous residual block (fuse) is the input of the next residual block, and the image of each residual block The size decreases $\frac{1}{2}$ sequentially. And the deep features at the top of feature extraction use CA to extract channels with strong semantic information (FCA). The yellow area in the generator network is the middle layer, the middle layer also corresponds to 4 feature blocks of different scales (extract1 ~ extract5), each of which is reduced to half of the former in an orderly manner, and these multi-scale details features will be added to the fusion of features. The olive region of the generator network is the feature fusion layer, CA multipath feedbacks the features with strong semantic information (FCA) in the top convolutional layer to CSA, and then gradually feedbacks FCSA to generate powerful detailed features. The specific algorithm is shown in the Algorithm 1.

### 3. Loss Function

In this article, we combine two methods to replace the loss function in DeblurGAN-v2 [4]. We not only use a pixel-by-pixel loss function that depends on low-level pixel information $L_{MSE}$, but also a perceptual loss function $L_{percep}$ that relies on the high-level features of the pre-trained loss network. In the training process, pixel loss does not measure the similarity of images more reliably than perceptual loss, and the pixel loss is recoverable during testing and the network cannot run in real time. Among them, $B_i$ is the input blurred image, $G_{θG}$ is the generator, $I_i$ is the output clear image, and $D_{θG}$ is the discriminator. $w$ is the width of the feature map and height $s$ the width of the feature map, which are obtained by using the Relu 3x3 layer of the VGG-16 network with parameters $φ$.

$$L_{MSE} = \frac{1}{wh} \sum_{x=1}^{w} \sum_{y=1}^{h} (I_{i,x,y} - G_{θG}(B_i,x,y))^2. \quad (1)$$
Algorithm 1 Algorithm of MSPA-net

**Input:** blurring image  
**Output:** sharp image

1. Feature extraction layer and middle layer:
2. for i in range(2,3):
3. \( \text{extract}_{i-1} = \text{Feature extraction}(\text{extract}_{i-1}) \) (extract\(_1\)=blurring image)
4. for i in range(1,4):
5. \( \text{detail}_i = \text{inception}_i(\text{extract}_i) \)
6. \( F_{CA} = \text{CA} \cdot \text{extract}_5 \)
7. Feature fusion layer:
8. \( \text{fuse}_1 = \text{extract}_5 \)
9. for i in range(1,3):
10. \( \text{merge}_i = \text{detail}_{5-i} + \text{fuse}_i \)
11. \( F_{CSAi} = \text{CSA} \cdot (F_{CA} \cdot \text{merge}_i) \)
12. \( \text{fuse}_{i+1} = \text{upsampling}(F_{CSAi}) \)
13. sharp image = upsampling(\( \text{fuse}_4 + \text{detail}_1 \))

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

In order for our model to solve the problem of gradient explosion and disappearance, the objective function is difficult to optimize while training more stable. We choose the perceptual WGAN loss proposed in [6].

\[ L_{\text{WGAN}} = \sum_{i=1}^{I} -D_{\theta G}(G_{\theta G}(B_i)). \] (3)

Therefore, our total loss function is as follows:

\[ L_G = 0.5L_{\text{MSE}} + 0.006L_{\text{percep}} + 0.01L_{\text{WGAN}}. \] (4)

4. Experimental Evaluation

4.1. Ablation Experiment
The quantitative comparison results are shown in Table 1 and the qualitative results are shown in Figure 5. It shows that MSPA-net has better performance. It also shows that the progressive attention network, width perception residual block and pyramid perception residual block are all valid.

4.2. Quantitative and Qualitative Evaluation
4.2.1. Quantitative and Qualitative Evaluation on GOPRO Dataset
We choose the remaining 1111 blur/clear images in the GOPRO dataset as the test set for model evaluation, and compare them with the current ones tested on the GOPRO dataset. We choose PSNR, SSIM and time spent as the three major indicators for our model evaluation. As shown in Table 2, the traditional
methods based on prior information [7] do not perform well, and the multi-scale literature [1] takes a long time because there is no parameter sharing. Looking at structural similarity only, DeblurGANv2 [4] performed best. In terms of PSNR and SSIM as a whole, our model (MSPA) and SRN [2] perform best, but the time we spend testing a single image is much less than SRN [2], and our network model costs The time is only 0.323s, which means that our network can deblur in real time.

In order to evaluate several models subjectively, this article uses DMPNH, SRN, DeblurGanv2 and this paper model (MSPA) to deblur a blurred image in the GoPro dataset as shown in Figure 6. It can be seen that the MSPA model can effectively restore edges and textures without obvious artifacts.

Table 1. The ablation model in this article is quantitatively evaluated on the GoPro test set

| Model          | Plain | No-Width residual block | No-Pyramid residual block | No-PCSANet | MSPA(complete) |
|----------------|-------|-------------------------|---------------------------|------------|----------------|
| PSNR           | 27.44 | 29.76                   | 29.46                     | 29.16      | 30.24          |
| SSIM           | 0.823 | 0.861                   | 0.851                     | 0.832      | 0.907          |

Table 2. Performance and efficiency comparison on the GoPro test dataset

| Methods        | Nah et al. [1] | SRN [2] | DeblurGAN [3] | DeblurGAN-v2 [4] | Li et al. [7] | Xu et al. [8] | Sun et al. [9] | DMPHN | MSPA (ours) |
|----------------|----------------|--------|---------------|------------------|---------------|---------------|---------------|-------|-------------|
| PSNR           | 29.08          | 30.10  | 26.435        | 29.55            | 27.08         | 25.1          | 24.64         | 28.70 | 30.24       |
| SSIM           | 0.913          | 0.932  | 0.892         | 0.934            | 0.857         | 0.894         | 0.842         | 0.913 | 0.915       |
| Time           | 4.33s          | 1.31s  | 0.92s         | 0.35s            | 1.7min        | 13.41s        | 20min         | 0.005s | 0.323s      |

Table 3. The model is quantitatively evaluated on the Kohler standard dataset

| Methods        | Nah et al. [1] | SRN [2] | DeblurGAN [3] | DeblurGAN-v2 [4] | Li et al. [7] | Xu et al. [8] | Sun et al. [9] | MSPA (ours) |
|----------------|----------------|--------|---------------|------------------|---------------|---------------|---------------|-------------|
| PSNR           | 26.48          | 26.75  | 26.10         | 26.36            | 25.88         | 25.1          | 25.245        | 26.81       |
| SSIM           | 0.807          | 0.837  | 0.816         | 0.820            | 0.801         | 0.894         | 0.784         | 0.822       |
4.2.2. The Model is Quantitatively Evaluated on the Kohler Standard Dataset The Kohler dataset is a standard data set for deblurring. The camera records the motion trajectory of 6 degrees of freedom and analyzes the real camera motion trajectory. The Kohler data set applies 12 sets of different blur kernels to 4 blur images. As shown in Table 3, this method compares peak PSNR and SSIM with other advanced methods in the Kohler data set. Moreover, as shown in Figure 7, it can be seen from a subjective qualitative that our algorithm (MSPA) is superior to the SRN and DeblurGanv2 algorithms.

5. Conclusion
In the feature extraction layer, this paper designs the width perception residual and the pyramid perception residual block which allow the multi-scale skeleton network to better retain the transmitted blur information and reduce the difficulty of network training. In the feature fusion layer, we design a progressive channel and spatial attention network (PCSA), which not only solves the problem of semantic sparseness, but also enhances the attention to the foreground target of the image in the dynamic scene and better suppresses the generation of image edge artifacts. Compared with the algorithm proposed recently, the algorithm in this paper has a better blur effect in dynamic scenes. The algorithm in this paper also has some flaws. In
the image restoration process, the color fidelity is not good, which means that our algorithm will improve the standard performance indicators of SSIM in the next work. In addition, the algorithm is only limited to blurring a single image, and the integration of binocular vision technology into image deblurring is also our next research trend.

Figure 7. Visual comparison of model adaptation to Kohler dataset.

References

[1] Nah S, Hyun Kim T, 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. 3883-3891.

[2] Tao X, Gao H, Shen X, et al. Scale-recurrent network for deep image deblurring. 2018 In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8183-8192.

[3] Kupyn O, Budzan V, Mykhailych M, et al. Deblurgan: Blind motion deblurring using conditional adversarial networks. 2018 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8183-8192.

[4] Kupyn O, Martyniuk T, Wu J, et al. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. 2019 Proceedings of the IEEE International Conference on Computer Vision, pp. 8878-8887.

[5] Liu Y, Li, J., Huang W, et al. Dynamic Scene Deblurring Based on Semantic Information Supplement. 2020 Journal of Physics: Conference Series. IOP Publishing, Vol. 1518, No. 1, p. 012069.

[6] Arjovsky M, Chintala S, Bottou L. Wasserstein gan. 2017 arXiv preprint arXiv:1701.07875.

[7] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. 2017 Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117-2125.

[8] Xu L, Zheng S, Jia J. Unnatural l0 sparse representation for natural image deblurring. 2013 Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1107-1114.

[9] Sun J, Cao W, Xu Z, et al. Learning a convolutional neural network for non-uniform motion blur removal. 2015 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 769-777.

[10] Gao H, Tao X, Shen X, et al. Dynamic scene deblurring with parameter selective sharing and nested skip connections. 2019 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3848-3856.