Single Image Deblurring Using Bi-Attention Network

Yaowei Li[1], Ye Luo[1] , Jianwei Lu[1]

[1] 1 School of Software Engineering, Tongji University

ABSRACT

Recently, deep convolutional neural networks have been extensively applied into image deblurring and have achieved remarkable performance. However, most CNN-based image deblurring methods focus on simply increasing network depth, neglecting the contextual information of the blurred image and the reconstructed image. Meanwhile, most encoder-decoder based methods rarely exploit encoder' s multi-layer features. To address these issues, we propose a bi-attention neural network for single image deblurring, which mainly consists of a bi-attention network and a feature fusion network. Specifically, two criss-cross attention modules are plugged before and after the encoder-decoder to capture long-range spatial contextual information in the blurred image and the reconstructed image simultaneously, and the feature fusion network combines multi-layer features from encoder to enable the decoder reconstruct the image with multi-scale features. The whole network is an end-to-end trainable. Quantitative and qualitative experiment results validate that the proposed network outperforms state-of-the-art methods in terms of PSNR and SSIM on benchmark datasets.

INTRODUCTION

Single image Deblurring has recently received much attention in our daily life. It is to generate a high-level deblurred image from its blurred image, which is caused by camera shaking or object moving. However, the inverse problem is ill-posed since it is hard to restore a deblurred image from its blur one. Therefore, numerous deblurring methods have been proposed ranged from prior-based deblurring and deep learning-based deblurring. Although considerable progress has been achieved, those methods still have some limitations: (1) underutilized the information of the original image and the reconstructed image: most CNN-based deblurred methods do not make full use of the structure information from either the blurred image or the reconstructed image; (2) rarely exploiting multiple layers of features in encoder: although a large number of methods have achieved good results, few consider the information of multiple layers in the encoder as the input of decoder.

In order to address above issues, we propose a single image deblurring method via a bi-attention network. The bi-attention network is designed by plugging two attention modules(Denoting criss-cross attention module) before and after the encoder and the decoder separately to capture long-range spatial contextual information from the blurred image and the reconstructed image, respectively. Between the encoder and the decoder, the feature fusion module is proposed to concatenate multi-layers of features from encoder such that the concatenated feature, which is the input of decoder, is rich of scale information.

THE PROPOSED METHOD

☐ Framework

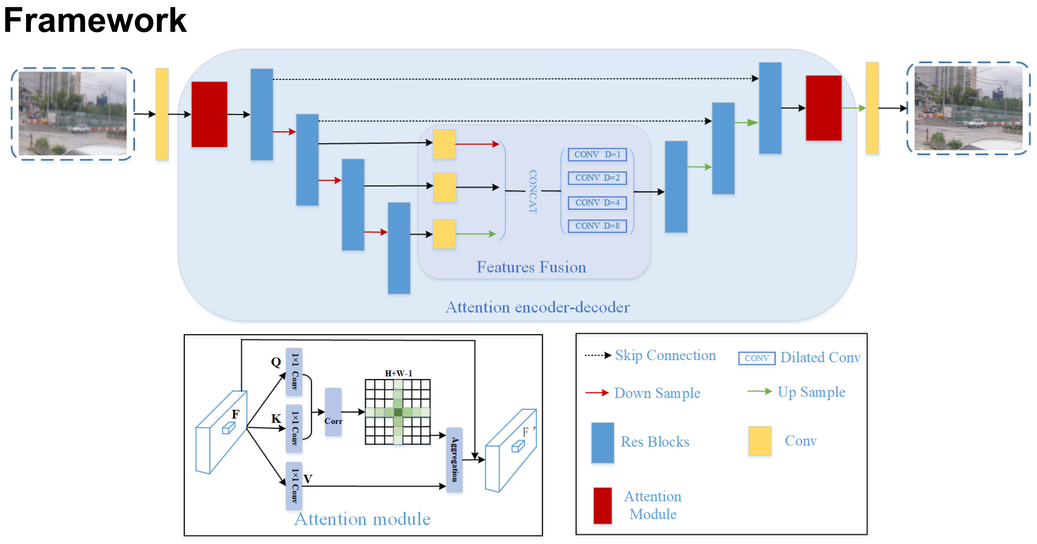


Fig.1. Framework of our proposed method, and its submodules including the attention module and the feature fusion module.

☐ The Attention Module

Given a local feature map $F \in \mathbb{R}^{C\times W\times H}$, the attention module firstly applies two convolutional with the 1×1 filters on F to generate two feature maps Q and K, where $Q,K \in \mathbb{R}^{C'\times W\times H}$ and C' is less than the original channel C. Given any spatial position at s in Q. We want to calculate the correlations of the feature at this position s on Q to features to some points on K by:

$$Corr_{i,s} = Q_s K_{i,s}^T \quad (1)$$

Where $Q_s \in \mathbb{R}^{C'}$ represents the feature vector at spatial position s in Q. T is the transportation operation. $K_s \in \mathbb{R}^{(H+W-1)\times C'}$ denotes the feature matrix with each row a feature vector collected from the pixel point on a cross centered at the spatial position s on K. C_i stands for the correlation between the features Q_s and the features $K_{i,s}$, $i = \{1,\cdots,H+W-1\}$. Overall, $Corr \in \mathbb{R}^{(H+W-1)\times H\times W}$ represents the correlations of all the points on Q to their counterparts on K.

After generating the attention map $Corr$, another 1×1 convolutional is used on F to generate $V \in \mathbb{R}^{C\times H\times W}$. For each spatial position s at V, we can get a feature vector $v_{i,s} \in \mathbb{R}^C$.

Thus, the contextual information of the point at position s can be aggregated into its feature as:

$$F_s' = \sum_{v_{i,s}\in\Omega_s} v_{i,s} C_{i,s} + F_s \quad (2)$$

Where $F_s' \in \mathbb{R}^C$ represents the attention weighted feature vector at spatial position s of $F' \in \mathbb{R}^{C\times H\times W}$ and $\Omega_s \in \{v_{i,s}\}_{i=1,\cdots,H+W-1}$ stands for the points around s on V.

☐ Bi-Attention Network

To exploit abundant structure information in the blurred image, we put the criss-cross attention module before the encoder-decoder network. Moreover, to utilize the self-similarities in reconstructed feature map, another cris-cross attention module is plugged after the encoder-decoder part. Specifically, given the shallow feature F_0, which is sent into the attention module:

$$F_{FCCAM} = H_{CCAM}(F_0) \quad (3)$$

Where H_{CCAM}, F_{CCAM} stands for the criss-cross attention module and the features that is obtained by the first criss-cross attention module. The same operation is applied to the second criss-cross attention module for the reconstructed feature F_{rec} as following:

$$F_{LCCAM} = H_{CCAM}(F_{rec}) \quad (4)$$

☐ Feature Fusion Module

Though the encoder-decoder network is an exceedingly effective method in image deblurring, during the upscaling stage of the decoder, usually only the feature map from the last residual block of the encoder is used to upscale. And this process under-utilizes the rich features from the encoder.

We apply three down-scales convolutional stages at encoder, where we downscale the last 3th feature map and upscale the last 1th feature map in encoder, respectively, and concatenate the features from these three layers:

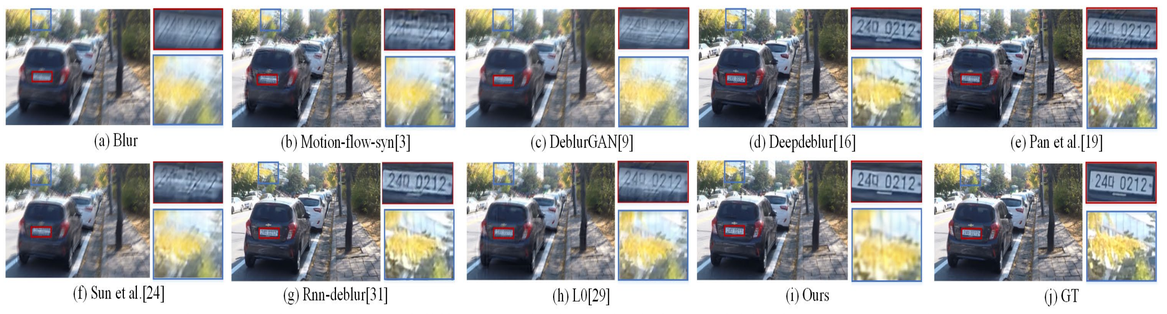
$$F_{con} = H_{\downarrow}(F_{3th}) + H(F_{2th}) + H_{\uparrow}(F_{1th}) \quad (5)$$

EXPERIMENTAL RESULTS

We quantitatively and qualitatively evaluate the proposed network method against state-of-the-art single image deblurring methods on GOPRO and DVD datasets as shown in Figure 3 and Table 1.

(a) Blurred Image (b) Ours (c) GT (a) Blurred Image (b) Ours (c) GT

Fig.2. Sampled deblurring results of our method from GOPRO and DVD dataset



(a) Blur (b) Motion-flow-syn[3] (c) DeblurGAN[9] (d) Deepdeblur[16] (e) Pan et al.[19]

(f) Sun et al.[24] (g) Rnn-deblur[31] (h) L0[29] (i) Ours (j) GT

Fig.3. Visual comparison of different methods on single image deblurring on GOPRO dataset

Table 1 Average PSNR and SSIM of different methods on GOPRO Dataset

Method	Motion-flow-syn[3]	deepdeblur[16]	deblurGAN[9]	Sun's[23]	Rnn-deblur[31]	Pan's[18]	L0[29]	**Ours**
PSNR	26.10	26.98	26.01	25.30	29.22	25.30	27.80	**30.20**
SSIM	0.80	0.84	0.77	0.78	0.87	0.72	0.85	**0.90**

CONCLUSION

We propose a bi-attention network for single image deblurring. This network mainly consists of bi-attention network and feature fusion network. Specifically, the bi-attention network captures long range feature information before and after encoder-decoder part and the feature fusion network is to combine more feature information from encoder part. Extensive experiments show the effectiveness of our proposed network.