Attention Network for Non-Uniform Deblurring

QING QI\(^{1,2}\), JICHANG GUO\(^{1}\), (Member, IEEE), AND WEIPEI JIN\(^{1}\)

\(^{1}\)Department of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
\(^{2}\)Department of Physics and Electronics Information Engineering, Qinghai University for Nationalities, Xining 810007, China

Corresponding author: Jichang Guo (jcguo@tju.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 61771334, and in part by the Chunhui Project of the Ministry of Education under Grant Z2016105.

ABSTRACT Recently, image deblurring task is valuable and challenging in computer vision. However, existing learning-based methods can not produce satisfactory results, such as lacking of salient structures and fine details. In this paper, we propose a solution to transform spatially variant blurry images into the photo-realistic sharp manifold. In this paper, we investigate an attention network for image deblurring. Instead of relying on local receptive fields to construct features by previous state-of-the-art methods, the non-local features for capturing long-range dependencies and the local features rely on receptive fields should be jointly considered. Therefore, we propose a novel dense feature fusion block that consists of a channel attention module and a pixel attention module. In addition, we further densely connected multiple dense feature fusion blocks to acquire high-order feature representation. Moreover, a scale attention module is further introduced for removing unfavorable features while retaining features that facilitate image recovery. Comprehensive experimental results show that our method is able to generate photo-realistic sharp images from real-world blurring images and outperforms state-of-the-art methods.

INDEX TERMS Generative adversarial networks, non-uniform image deblurring, attention network.

I. INTRODUCTION

Image deblurring is a challenging problem in the field of image restoration. Blind image deblurring is to estimate the underlying image from the single degraded observation where the kernels are absent. Thus, the image recovery needs to be restricted by additional hand-crafted priors [1]–[7]. Let \( B \) denotes the convolution of the sharp image \( S \) with a kernel \( K \). The mathematical expression is represented as \( B = S \ast K + N \), where \( N \) implies noise. The minimization of the cost function for solving image deblurring task is formulated by

\[
S = \arg \min_S \left( \|B - S \ast K\|_2^2 + f(B, K) \right)
\]

where \( f(B, K) \) implies priors information based on image statistics or hand-crafted priors. These early methods are proposed based on the assumption of linear space-invariant blurring process. Due to these methods involve solving optimization equations, thus sometimes may converge to local solutions. Consequently, it leads the inaccurate kernel estimation and the incomplete clean image. In reality, real-world blurring observations are often tend to be a complicated variant condition. Non-uniform image deblurring caused by the camera shake, object motion or even more complicated factors cannot be well approximated by linear solutions. Previous deblurring methods [8], [11], [12] highly rely on moving object segmentation of non-uniformly degraded scenes for estimating multi-kernel and restoring the related regions. However, the patch-level deblurring strategy can not reveal the intrinsic essence of the non-uniform degraded scenes.

Recently, learning-based methods have proven useful for tackling a variety of image processing tasks [10], [13] which also include image deblurring problem. Instead of solving complicated optimization equations, methods [14]–[19] estimate kernels which are significant components in the traditional image deblurring framework. However, once the kernel estimation is incorrect, consequently the subsequent deconvolution operator will be effected. Lately, methods [20]–[27] adopt the end-to-end trainable networks which input the degraded observations and directly obtain the underlying clean images. This learning manner directly avoids the potential effect of the kernel estimation. Recently, VGGNets [28] has been demonstrated that increasing the depth of the network could significantly enhance the quality
of the features representations. Therefore, image deblurring methods [23], [24] explicitly or [21], [25], [27] implicitly increase receptive fields for achieving promising performance. However, this strategy inevitably leads to increase model size and computational complexity.

In this paper, we investigate an attention network rather than the local features for blind image deblurring. The main contributions of the proposed method are listed as follows:

Firstly, the specially designed GAN-based model that effectively transforms the images from the blurry manifold to the sharp manifold. Our method is able to generate high-quality clean images on both synthetic and real-world images, which are usually better than the state-of-the-art methods.

Secondly, we propose an attention network for blind image deblurring, the attention network mainly includes a dense feature fusion group and a scale attention module. Image deblurring task can be well investigated by multiple perspectives rather than the local receptive fields.

Thirdly, we propose a dense feature fusion group consists of a channel attention module and a pixel attention module. The proposed dense feature fusion block aims at modeling rich contextual dependencies through jointly considering the non-local features and the local features. The dense feature fusion group effectively learns uncertainty maps to guide the pixel loss for more robust optimization.

Fourthly, instead of connecting the feature representations of the encoder and the decoder directly, we propose a novel scale attention module to extract essential information for features reconstruction. In this way, replacing shortcut connections with the scale attention module significantly facilitates the performance of the proposed method.

Finally, we propose an edge loss function and a content loss function for the generator. The proposed edge loss aims at generating latent images having the quality of sharp edges and the content loss function is utilized for preserving the semantic features of images.

II. RELATED WORK

In recent years, the single image deblurring issue has made significant progress. In this section, we focus on the traditional optimization-based methods and the learning-based methods of the non-uniform image deblurring.

A. TRADITIONAL OPTIMIZATION-BASED METHODS

Unlike the generic image deblurring task, non-uniformly degraded scenes are varying everywhere. To handle this problem, Kim et al. [8] assume that the non-uniformly degraded scenes caused by objects motion and camera shake. They adopt the segmentation methods to cope with the distinct motion regions. Segmentation-based image deblurring methods [29] are developed. However, these methods require a known model of the blur type. Lately, Kim and Lee [11] assume that the locally linear motion can approximate the complicated non-uniform blur. Then, they propose a segmentation-free approach to tackle this task. However, in the case that the object motion and camera shake simultaneously happen, the assumption of the locally linear model cannot reveal the intrinsic essence of the non-uniform blur. Lately, Pan et al. [12] propose a method to cope with the segmentation confidence map which is used for reducing the segmentation ambiguity between different motion regions.

B. LEARNING-BASED METHODS

Recently, with the advanced development of deep learning in computer vision tasks, such as underwater image processing [30]–[32], object salient detection [33], and image dehazing [34]. Learning-based methods have been employed in the image deblurring. According to whether the image deblurring process involves kernel estimation, the learning-based image deblurring methods can be roughly divided into two categories.

The first type method based on “kernel-unfree” which exploits CNNs to obtain kernels by replacing principle components of kernel estimation in the traditional framework. Sun et al. [18] propose a CNN network to estimate the motion blur by image patch-level. Then, the Markov random field (MRF) is applied to obtain a dense motion field. However, due to the network is trained according to the patch-level, it fails to take the image global information into account. Gong et al. [19] utilize a CNN to estimate the motion flow, and then the non-blind deconvolution operation is still adopted to estimate the latent image. The “kernel-unfree” based methods are unavoidable to produce incorrect results caused by kernel estimation. Consequently, the second category method [23]–[27] named as “kernel-free” is emerged. They employ the end-to-end learning fashion which inputs the degraded observations to the network and directly obtains the deblurred results. Nah et al. [23] propose a multi-scale network with 40 convolution layers in each scale and 120 convolution layers in total. The network learns in a pyramidal manner and finally obtains clear images. However, parameter dependencies across scales are not well utilized. In other words, this pyramidal approach does not share weights across scales. Therefore, the method confronts with the problem of large model size and high computation budget. Tao et al. [24] consider the method [23] does not take the dependencies across scales into consideration, and this architecture may not restrict solution space. Thus, Tao et al. propose a recurrent multi-scale network that maximum facilitates feature sharing.

Recently, as the Generative Adversarial Networks (GAN) [35] proposed, many image style transfer methods are developed. CycleGAN [36] brings us the inspiration and view to reconsider the task of image deblurring. However, directly adopting the CycleGAN to image deblurring task is infeasible. First, due to the highly unstable of simultaneously train a pair of network, thus the network is difficult to achieve convergence. Second, weakly supervise learning pattern barely captures lost high-frequency information, and thus fails to produce promising results with fine details for image deblurring. Reference [25] propose a conditional GAN, the method requires pairs of images to constrain the
high-frequency detail information. However, it is difficult to understand the nature of non-uniform image deblurring by limiting the restoration process of the image by a single semantic loss. Nimisha et al. [26] utilize the blurry image to self-guide the network training which recovers the content and the color to the correct manifold. Furthermore, the method is trained by the weakly supervised learning manner based on unpaired datasets. Lately, in order to meet the real-time processing requirements, [27] proposes a feature pyramid network as an alternative scheme. In addition, they propose three backbones to pursue high-quality image and real-time processing respectively. As a comparison, the proposed method is based on an attention network. Our method is able to generate high-quality images with clean edges and fine details.

III. PROPOSED METHOD

In this section, we first introduce the proposed dense feature fusion group and the scale attention module in Section III-A and Section III-B, respectively. In Section III-C, we detail the network architecture. Finally, we describe the loss functions utilized in the proposed model in Section III-D.

A. DENSE FEATURE FUSION GROUP

The proposed dense feature fusion group contains multiple attention blocks. Each attention block contains cascaded connected spatial attention and channel attention modules. More importantly, each blocks are connected to each other in a dense manner. Specifically, we design a channel attention module to draw reweigh non-local features that access the significance of each channel. Then, based on this, we further design a pixel attention module to obtain better feature representations for pixel-level restoration. Next, we introduce these two modules in detail.

1) CHANNEL ATTENTION MODULE

Generally, in the network, a set of filters of a convolutional layer express neighborhood spatial connectivity patterns within local receptive fields. However, we observed that under the complicated non-uniform condition, when the heavy camera shake and object motion simultaneously happen, the global spatial dependencies also need to be considered. For example, in image neighborhoods containing a particularly strong blurring component, precise discrimination using any possible method may require contextual information from clean regions. Thus, besides the local features, we expect a richer non-local and overall structure by explicitly modeling channel interdependencies. The underlying essence is to introduce non-local contextual information across channels, meanwhile dramatically decrease huge parameters. Recently, the channel attention module [37] is originally developed based on the non-local feature and explicitly modeled the channel correlations in the field of high-level image processing. Here, we propose a channel attention module to achieve the proposed non-local context encoding which recalibrates feature maps according to the weight of the feature itself.

The architecture of the channel attention module is depicted in Figure 1. Let \( U = [u_1, u_2, \ldots, u_C] \) denotes the original activation features which is transformed from a set of filters \( V = [v_1, v_2, \ldots, v_C] \) convolved with input images \( X = [x_1, x_2, \ldots, x_C] \). Here, \( C \) denotes the feature maps of the size of \( H \times W \). Since features from each channel are only closely related to the current receptive fields, transforming output \( u_C \) of each channel can represent non-local features. In order to acquire global spatial information, we apply a global average pooling operator \( f_{sq} \) as channel-specific descriptor \( z_c = f_{sq}(u_C) \), which squeezes global spatial information of each feature map \( u_C \in H \times W \). Besides the squeezed global statistics, adaptive access to the weights of each channel is also important. Thus, we further capture the channel-wise dependencies for emphasizing the relative importance of each channel and benefit in predicting the clean features. Consequently, the channel attention module is obtained by \( s = \sigma(W_2 \delta(W_1 z)) \) where \( z = [z_1, \ldots, z_C] \in R^C \) expresses statistics for the whole image, \( W_1 \) is a weight that downsamples \( z \), \( \delta \) is the ReLU activation function [38], \( W_2 \) represents the upsampling weights that return the dimension to the original activations features \( U \), and \( \sigma \) implicate Sigmoid activation function. The output is served as the reweighed feature of each channel and the final output vector is obtained by dot product \( u_C = s_C \cdot u_C \).

2) PIXEL ATTENTION MODULE

The channel attention module adaptively accesses to the weights of each channel and emphasizes the relative importance of each channel. Considering that the blurry image distribution is variant on the different image pixels, we further learn the spatially variant properties of the blurry images in an adaptive way by the pixel attention module. Specifically, the pixel attention module consists of two convolution layers, each layer activated by a ReLU function and a sigmoid function respectively.

\[
U_p = \sigma(\text{Conv}\delta(\text{Conv}(u_C)))
\]  

(2)

3) DENSE FEATURE FUSION

Every attention block arranged in a cascade form is not competitive enough to cope with the image deblurring task.
Due to the severe ill-posed of the degraded observations, the most high-frequency details are already lost. Meanwhile, the recalibrated features of the attention block merely flow in sequential discouraging feature reuse. Therefore, to further improve the features flow between attention modules and alleviate the gradient-vanishing problem, we densely connect these attention blocks.

Let an image $x_0$ pass through a CNN. The network comprises $L$ layers, each of which implements a non-linear transformation $F_n(\cdot)$, where $n$ indicates the layer. $F_n(\cdot)$ can be a composite function of the channel attention module and the pixel attention module. Figure 2a illustrates the layout of the resulting dense feature fusion group schematically. Recalling the statement of residual function as

$$u_n = u_{n-1} + F_n(u_{n-1})$$  \hspace{1cm} (3)

where $u_{n-1}$, $u_n$, and $F_n$ are the input, out and the residual function. Empirically, fitting the residual of residual is easier than the original residual map. Consequently, we transform the proposed dense feature fusion group as the form of the residual maps.

$$u_n = F_1(u_0) + F_2(F_1(u_0)) + F_3(F_1(u_0) + F_2(F_1(u_0)))$$  \hspace{1cm} (4)

Figure 2c shows the third-order residual function. According to Figure 2c, we can deduce the eleventh-order residual function. Higher-order residual functions have the ability to express more complicated representations than multiple stacked attention blocks. As shown in Figure 2, the layout of the connections is a nested network which is visually similar to DenseNet [39]. However, it should be noted that the proposed dense feature fusion group is different from DenseNet. DenseNet emphasizes the concatenation operations of feature maps, while our method utilizes the fusion features by the summation operations. The proposed dense feature fusion group not only encourages features reuse but fits the residual mapping easier to optimize.

A pipeline of the proposed image deblurring network is shown in Figure 4. In what follows, we explain the important parts of it.

The proposed network consists of six distinct blocks that are denoted by a different color. The first type of block is expressed as a group of “Conv-Ins.Norm-ReLU” layers. The parameters of the “Conv” layer are represented as “kernel size × output feature numbers × stride”. The second type of block is indicated as a downsample block which includes a “Conv-Ins.Norm-ReLU” layer and three residually connected groups of “Conv-Ins.Norm-ReLU”. The third type of block is a dense feature fusion group which contains eleven attention blocks with identical layout for easing the notorious vanishing-gradient problem and promoting feature reuse. Specifically, each attention block contains spatial attention module and channel attention module. The fourth type of block is represented by an upsample block which includes three residually connected groups of “Conv-Ins.Norm-ReLU” and a “Conv-Ins.Norm-ReLU” layer. The fifth type of block is expressed by a scale attention module which suppresses responses of irrelevant features.

Specifically, the output of the scale attention module is the element-wise multiplication of the decoded feature maps and the attention gate coefficients. The output of the scale attention module range from 0 to 1. Therefore, each pixel vector in the image is filtered by a single scalar attention value. Then, the attention gate features further enhanced by two “Conv-Ins.Norm-ReLU” blocks. As shown in Figure 3, the proposed scale attention module is incorporated into the generator architecture to highlight clean features that are passed through the skip connections.

**B. SCALE ATTENTION MODULE**

In U-net [40], feature maps encode at multiple scales are merged through skip connections to form the coarse-to-fine structure for reconstructing content and sharp manifold features. Features at the coarse spatial level model detailed information. However, it remains difficult to differentiate the desired clean features from those that can be discarded as blurry features. Therefore, we propose a scale attention module which suppresses responses of irrelevant features.
increase from $64 \times 64$ to $256 \times 256$. At the stage of the encoder, useful detailed features are extracted for transforming downstream. To the best of our knowledge, this is the first time that we have introduced the concept of attentional mechanisms (from high-level image processing) into image deblurring task (low-level image processing). Correspondingly, the decoder begins with three upsample blocks for reconstructing sharp manifold features. The feature numbers progressively decrease from $256 \times 256$ to $64 \times 64$. Furthermore, to facilitate the performance of the network, we establish the scale attention module between the encoder and the corresponding scales for reserving the essential features. Finally, the output photo-realistic image is obtained by a Tanh activation function.

The discriminator $D$ is utilized to distinguish whether the input image is a generated image or a real one. Different from the high-level image processing task (object classification), image sharpness discrimination relies on local features of the image. Instead of a general full-image discriminator, we adopt PatchGAN [41] for the discriminator $D$. PatchGAN starts with a flat convolutional layer, after which the network adopts three stride convolutional layers to reduce the resolution and encode essential local features for classification. Afterward, a convolutional layer and a Sigmoid activation function are used to obtain the classification response. Instance normalization and Leaky ReLU are appended after each convolutional layer.

**D. LOSS FUNCTIONS**

The loss function $\mathcal{L}(G, D)$ of the proposed network in Eq.5 is composed of three parts: (1) the semantic content loss $\mathcal{L}_{\text{content}}$ (Section III-D1), which preserves the image content coherence during the process of discriminative learning. (2) the edge loss $\mathcal{L}_{\text{edge}}$ (Section III-D2), which measures the edge features dissimilarity between the generated image and the related clean image. (3) the adversarial loss $\mathcal{L}_{\text{adv}}(G, D)$ (Section III-D3), which propels the generator to reach the desired manifold. The loss function is expressed by a simple
additive form:
\[ \mathcal{L}(G, D) = \beta \mathcal{L}_{content} + \lambda \mathcal{L}_{edge} + \sigma \mathcal{L}_{adv} \]  
(5)

where \( \beta, \lambda, \) and \( \sigma \) are weights of semantic content loss, edge loss, and adversarial loss respectively. A Larger value indicates the corresponding component is more significant. Generally, the task of transforming the images of the blurry manifold to the sharp one expects to recover the faithful characteristics of photo-realistic images. In all our experiments, we empirically set \( \beta = 10, \lambda = 12 \) and \( \sigma = 1 \) respectively.

1) CONTENT LOSS

Instead of utilizing pixel-wise Mean Squared Error (MSE) loss as loss function, inspired by [42], we introduce a semantic content loss that captures high-level semantic and perceptual differences between the deblurred image and the corresponding label for the proposed network optimization. One significant advantage of the semantic content loss is that reduces the effects of the artifact induced by pixel-wise restriction and improves the subjectively visual perception of the deblurred results. We employ the high-level features of the layer “Conv4-3” of the pre-trained VGG19 layer network. Accordingly, the semantic content loss is defined as follows:

\[ L_{content}(s, G(b)) = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left\| \phi_{ij}(s_k)_{x,y} - \phi_{ij}(G(b_k))_{x,y} \right\|_2 \]  
(6)

where \( \phi_{ij} \) is the feature map obtained by the j-th convolutional after activation and before the i-th pooling layer within the pre-trained VGG19 network, and H and W are the height and width of inputs respectively. \( \phi_{ij}(s_k) \) denotes the semantic correspondence of the ground truth, and \( \phi_{ij}(G(b_k)) \) indicates the semantic correspondence of the synthesized image.

2) EDGE LOSS

Images having sharp edges is an intuitive judgment to prove its clarity. Thus, we introduce an edge loss that measures the edge features dissimilarity between the generated image and the corresponding ground truth for the proposed blurry removal network optimization. We calculate the edge loss with the designed Sobel convolution layer, whose kernels are Sobel operators [43]. This loss is mathematically expressed as follows:

\[ L_{edge}(s, G(b)) = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \left( \left\| \nabla h(s_k)_{x,y} - \nabla h(G(b_k))_{x,y} \right\|_1 + \left\| \nabla v(s_k)_{x,y} - \nabla v(G(b_k))_{x,y} \right\|_1 \right) \]  
(7)

where \( \nabla h \) and \( \nabla v \) are the gradient operations along the horizontal and the vertical directions respectively, \( H \) and \( W \) are the height and width of inputs respectively.

3) ADVERSARIAL LOSS

The adversarial loss is applied to both networks \( G \) and \( D \), which affects the image deblurring process of the generator network \( G \). Its probability value represents to what extent the deblurred image generated by \( G \) looks like a photo-realistic image. The competitive relationship between \( G \) and \( D \), correctly driving \( G \) to transform the degraded observation into a high-quality image. In the proposed model, we use WGAN-GP [44] as the discriminator loss function, which can stabilize the training process. The mathematical expression of the WGAN-GP objective function is as

\[ L_{GAN} = E[D(s_k)] - E[D(G(b_k))] - \lambda E_{\tilde{x}\sim p_{\mathcal{b}}}[(\|\nabla \tilde{x}D(\tilde{x})\|_2 - 1)^2] \]  
(8)

where \( G(b_k) \) is the generated image, \( s_k \) is the real image, \( E[D(G(b_k))] \) and \( E[D(s_k)] \) are the expectation of the discriminator assigns the correct label to the generated image \( G(b_k) \) and the clean image \( s_k \) respectively. \( \lambda E_{\tilde{x}\sim p_{\mathcal{b}}}[(\|\nabla \tilde{x}D(\tilde{x})\|_2 - 1)^2] \) represents samples as a linear combination of the real image and the generated image.

IV. EXPERIMENTS

In this section, we first depict the dataset and implementation details in Section IV-A. Then, in Section IV-B, we make quantitative evaluations of the proposed method and representative non-uniform image deblurring methods, such as CNN based method (Nah et al. [23] and Mustaniemi et al. [9]), methods Sun et al. [18] and Gong et al. [19] which bridge the gap between the traditional and learning-based methods, and the recently GANs-based methods (Kupyn et al. [25], Kupyn et al. [27]) (InceptionResNet-v2) on the synthetic and real-world non-uniform blurry images. To be fair, we recurrent these methods by conducting their official implementations with default settings and parameters.

A. DATASET AND IMPLEMENTATION DETAILS

1) DATASET

In this paper, the datasets utilized in the proposed method includes [23], [45], [46] and [47]. Nah et al. [23] propose a dataset named GOPRO that simulates the camera imaging process rather than designing specific kernels. GOPRO dataset includes 3214 pairs of aligned clean/blurry images with 1280 \times 720 resolutions. Specifically, 2103 pairs of pixel-aligned blurry and clean images as the training dataset, the remain 1111 pairs of images for testing. The Köhler dataset includes 4 clear images and each clear image has 12 different blur kernels, thus there are 48 blurry images. These kernels are recorded by sampling the 6D camera motion trajectory generated by recording real camera motion. These datasets are widely used for quantitative evaluating the performance of non-uniform image deblurring methods. The dataset of Lai et al. [47] is composed of real-world images and dataset of Su [46] consists of a variety of videos that take in real-time. In the proposed method, we employ the training dataset of GOPRO for training. Furthermore, we randomly crop the degraded observations and the corresponding ground truth for training.
Q. Qi et al.: Attention Network for Non-Uniform Deblurring

FIGURE 5. Qualitative evaluations on the non-uniform dataset GOPRO [23], zoom in for the best view. The proposed method has a photo-realistic effect and generates much clearer details.

FIGURE 6. Qualitative evaluations on the non-uniform dataset GOPRO [23], zoom in for the best view. The proposed method has a photo-realistic effect and generates much clearer details.

truth to $256 \times 256$ pixels during training. The test dataset of GOPRO, and the datasets of Köhler [45], Lai and Su are employed for testing.

2) IMPLEMENTATION DETAILS
The proposed model is built by utilizing PyTorch framework with Python interface. All experiments are implemented on a computer configured to include an Intel(R) Core(TM)i7 CPU (16GB RAM) 3.60 GHz CPU and an NVIDIA GeForce GTX 1080Ti GPU. For the first 150 epochs, the learning rate is 0.0001 for training both $G$ and the $D$, for the next 150 epoch we linearly decay the learning rate to zero. The networks are optimized using Adam optimizer [48], and the Momentum and weight decay coefficients are set to 0.5 and 0.999 respectively. The slope of the Leaky ReLU is 0.2. All models are trained for 300 epochs with a batch size of 4.

B. QUANTITATIVE EVALUATIONS
In order to verify the effectiveness and generalization of the proposed model, we quantitatively evaluate the proposed method with the state-of-the-art methods on the synthesized datasets of [23] and [45]. We employ evaluation metrics of Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [49] for quantitative evaluations. A higher SSIM value implies a deblurred image that is closer to the label image in terms of structural similarity. A higher PSNR value indicates the similarity in terms of pixel-wise values. Several deblurred results on the synthesize testing dataset [23] and [45] are shown in Figure 5, Figure 6, Figure 7, and Figure 8 respectively.
1) QUANTITATIVE EVALUATIONS ON DATASET OF GOPRO

Figure 5 shows some deblurred results of heavy camera shake condition. Obviously, the deblurred results obtained from the proposed method, the clean underlying scenes and finer details can be visually observed. Sun et al. [18] and Gong et al. [19] fail to deblur the extremely blurry images. Compared to the previous methods, the method of Mustaniemi et al. [9] has some deblurring performance. However, the results produce ringing-artifact effects. The multi-scale CNNs-based method [23] can not handle large
blurry caused by heavy camera shake. The GANs-based methods [25] and [27] also fail to deal with non-uniform blurry scenes properly.

Figure 6 and Figure 7 display some deblurred results of complicated blurring scenes which include the heavy camera shake and object motion. Learning-based methods [9], [18], [19], [23], [25], [27] fail to have convincing deblurring performance and their results produce significant ringing-artifacts and unclear details. Both multi-scale CNNs method [23] and GANs-based methods [25], [27] are not able to remove blur and generate results with salient edges. In contrast, our method has the ability to cope with the complicated conditions which include large camera shake and object motion. As shown in Figure 6h and Figure 7h, the recovered images with finer details and more salient structures.

2) QUANTITATIVE EVALUATIONS ON THE DATASET KÖHLER
In order to validate the generalization of the method, we test the proposed algorithms on the dataset of Köhler. In Figure 8, methods [9], [18], [19], [23], [25], [27] fail to have the deblurring performance. By contrast, the proposed method can generate clean images with good visual effects.

C. COMPARISONS ON REAL-WORLD IMAGES
We conduct several qualitative evaluations on the real-world blurry images to verify the generalization and effectiveness of the proposed method. We first select several real-world datasets [46] and [47] which are always used to qualitatively evaluate for non-uniform image deblurring. We compare the proposed method with the above-mentioned state-of-the-art methods in Figure 9, Figure 10, Figure 11 and Figure 12 respectively.

1) QUALITATIVE EVALUATION ON THE DATASET OF LAI
In Figure 9 and Figure 10, the results of [9] have obvious ringing-artifacts. Methods [18], [19] involve kernel estimation which inevitably generate the result deviated from the clean underlying. For the kernel-free methods [23], [25], [27], the deblur performance have significant improvement. However, the details cannot be recovered well, especially in the regions of the figure of the tower. By contrast, observing Figure 10h, the proposed method produces good deblur performance in the challenging low-light condition image deblurring task, which benefits from the proposed attention network.
In Figure 11 and Figure 12, the kernel-unfree method \[18\], \[19\] and the kernel-free methods \[9\], \[23\], \[25\], \[27\] have less deblur performance on the real-world image. The entire outline of the person cannot be recovered well and the visual effect of the deblurred image is unnatural. As shown...
FIGURE 12. Qualitative comparison on the real-world blurry images [46], zoom in for the best view. The proposed method has a photo-realistic effect and generates much clearer details.

in Figure 11h and Figure 12h, due to the attention network, the proposed method has good performance in coping with the challenging face outline recovery and our results have a visually pleasant effect.

D. EXECUTION TIME

We demonstrate the efficiency of the proposed method by conducting execution time with state-of-the-art methods. We calculate floating point operations per second (FLOPs) and the average execution time on images with the size of 1280 × 720, the results are shown in Table 2. Due to the deblurring methods [18], [19] involve with non-blind deconvolution process, they inevitably encounter the problems of high computational. Nah et al. [23] employ a multi-scale CNN to expand the receptive fields for recovering images. However, the multi-scale of the model is independent of each other, thus method [23] takes more testing time than the methods of Kupyn et al. [25] and Kupyn et al. [27]. Kupyn et al. [25] construct a Resblock based CNN to restrict image recovery process. Mustaniemi et al. [9] build a U-Net based CNN to restrict image recovery process. Kupyn et al. [27] enlarge the receptive fields by developing the feature pyramid network. Unlike image deblurring methods mentioned above, due to the proposed method involves the design of dense feature fusion group, therefore the proposed method spares a bit more time and parameters than other learning-based methods [25], [27].

V. ANALYSIS AND DISCUSSIONS

A. RELATION WITH IMAGE DEBLURRING METHODS

To solve dynamic scene deblurring, Gong et al. [19] and Sun et al. [18] utilize CNNs to achieve blur kernels estimation, and then the non-deconvolution is adopted to obtain underlying results. However, due to the inaccurate kernel estimation, consequently leading to the deblurred result which is indistinguishable from the degraded input. Method [9] introduces gyroscope measurements into a CNN. Kupyn et al. [25] using a semantic content loss function for capturing the high-level semantic correspondence. However, merely adopting the semantic content restriction may not tackle the non-uniform image deblurring task, especially caused by object motion and heavy camera shake.

TABLE 1. Performance comparison with the state-of-the-art methods by two quantitative evaluation metrics. We compute the average PSNR(dB) and SSIM values on the test dataset of GOPRO and Köhler.

| Methods      | GOPRO PSNR | GOPRO SSIM | Köhler PSNR | Köhler SSIM |
|--------------|------------|------------|-------------|-------------|
| Gong [19]    | 27.2778    | 0.8187     | 21.2371     | 0.6490      |
| Sun [18]     | 26.6260    | 0.8018     | 21.2335     | 0.6525      |
| Kupyn [25]   | 25.2363    | 0.7773     | 19.0843     | 0.5838      |
| Kupyn [27]   | 27.8086    | 0.8564     | 21.2987     | 0.6544      |
| Nah [23]     | 28.3225    | 0.8588     | 20.8507     | 0.6340      |
| Mustaniemi [9] | 25.9563    | 0.8285     | 20.4833     | 0.6442      |
| Ours         | 28.8148    | 0.8664     | 21.5722     | 0.6654      |

TABLE 2. Execution time comparison with state-of-the-art. We calculate the average execution time on images with the size of 1280 × 720 of the test dataset of GOPRO. "−" indicates not available.

| Methods      | FLOPs | Seconds |
|--------------|-------|---------|
| Gong [19]    | 4.12G | 1300    |
| Sun [18]     | 5.41G | 1500    |
| Kupyn [25]   | 678.29G | 1.1 |
| Kupyn [27]   | 411.34G | 0.65 |
| Nah [23]     | 1760.04G | 8.1 |
| Mustaniemi [9] | 128.13G | 4.5 |
| Ours         | 28.03G | 1.3    |
Furthermore, Kupyn et al. [27] enlarge the receptive fields by developing the feature pyramid network. The method [27] merely relies on the local features and ignores long-range dependencies from the non-local features. Nah et al. [23] propose a multi-scale deep network by stacking convolutional layers to cope with the dynamic scene deblurring. However, Nah et al. fail to utilize the dependencies between scales, thus method [23] involves the problem of convergence difficulties.

In contrast, we employ a diverse perspective to accomplish image deblurring task by the dense feature fusion group and the scale attention module. Therefore, the proposed method achieves the best PSNR and SSIM values on the dataset of GOPRO and Köhler, as shown in Table 1. In addition, the results of the proposed method on real-world images are in accordance with those of the synthetic blurry images. Even though the attention network is trained on synthesized blurry images, the experimental results demonstrate that our method has good generalization for real-world blurry images.

B. ABLATION STUDY

We conduct the ablation experiments to study the role of each part of the proposed method. We remove the dense feature fusion group and the scale attention module and refer to it as a baseline. In order to verify the importance of our network, we compare five architectures as depicted in Section III-C, including (1) Our baseline image deblurring network BaseNet; (2) BaseNet with the channel attention module BaseNet+CA; (3) BaseNet with the densely connected channel attention modules BaseNet+FCA; (4) The proposed BaseNet with dense feature fusion group that is the densely connected channel attention modules and pixel attention modules BaseNet+FAB (Fusion Attention Block); (5) The BaseNet + FAB and a scale attention module (BaseNet + FAB + SA). We train these networks using the identical training strategy and parameters as in Section IV-A.

As shown in Table 3, the performance of our BaseNet has little effect on extremely blurry images. Therefore, we first propose a channel attention module BaseNet+CA to recalibrate feature maps according to the weight of the feature itself for extracting the non-local features. However, from the visually results, we can find out that isolated adopt eleven channel attention modules arranged in a cascaded form can not generate clear enough results. In addition, the overly deep network BaseNet + CA module causes the notorious gradient-vanishing problem. Therefore, the ability to represent features and utilized for image deblurring is hindered. Recall that for alleviating the notorious problem, He et al. [50] fit the residual map is easier to optimize than the original map. Therefore, we densely connect the proposed channel attention modules. Consequently, the network BaseNet + FCA is proposed, the performance of image deblurring is significantly promoted. Furthermore, BaseNet + FCA can achieve strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. Furthermore, based on this, we further introduce the pixel attention module and embed it in the BaseNet + FCA network, finally from a new module named as BaseNet + FAB. Figure 13 visually shows the results of BaseNet and our BaseNet + FAB. It can be observed that BaseNet struggles to deblur the non-uniform image regions and yields some obvious trajectory of the car, while the BaseNet + FAB removes the blurring of the non-uniform regions which caused by the object motion. Moreover, we introduce the module BaseNet + FAB + SA to discard those blurry features, while retaining the features that are beneficial to image recovery. We analyze that the proposed attention network is important for recovering faithful characteristics of sharp images with photo-realistic.

![Visual results on the non-uniform dataset GOPRO [23], zoom in for the best view. The proposed method has a photo-realistic effect and generates much clearer details.](image_url)

**TABLE 3.** An ablation study of removing/changing components in the proposed method. We compute the average PSNR(dB) and SSIM values on the non-uniform dataset GOPRO [23].

| Methods             | GOPRO PSNR | GOPRO SSIM |
|---------------------|------------|------------|
| BaseNet             | 27.5495    | 0.8293     |
| BaseNet + CA        | 27.7276    | 0.8312     |
| BaseNet + FCA       | 28.1302    | 0.8404     |
| BaseNet + FAB       | 28.4873    | 0.8576     |
| BaseNet + FAB + SA  | 28.8148    | 0.8664     |
VI. CONCLUSIONS

In this paper, we have presented an attention model for image deblurring. For blur removal, the introduced a dense feature fusion group aims at investigating the relationship between non-local features and local features. Furthermore, a scale attention module is introduced for retaining features between non-local features and local features. Furthermore, feature fusion group aims at investigating the relationship for large-scale image recognition. In this paper, we have presented an attention model for this trade-off: http://arxiv.org/abs/1410.05495. [Online]. Available: http://arxiv.org/abs/1410.05495

[21] M. Jin, M. Hirsch, and P. Favaro, “Learning face deblurring fast and wide,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 745–753.

[22] X. Mao, C. Shen, and Y. B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Proc. NIPS, 2016, pp. 2802–2810.

[23] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 257–265.

[24] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, “Scale-recurrent network for deep image deblurring,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8174–8182.

[25] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, “DeblurGAN: Blind motion deblurring using conditional adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8183–8192.

[26] T. M. Nimisha, K. Sunil, and A. Rajagopalan, “Unsupervised class-specific deblurring,” in Proc. ECCV, Sep. 2018, pp. 353–369.

[27] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, “DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better,” 2019, arXiv:1908.03826. [Online]. Available: http://arxiv.org/abs/1908.03826

[28] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

[29] L. Graham and Y. Tizhakhy, “Blind restoration of space-variant Gaussian-like blurred images using regional PSFs,” Signal, Image Video Process., vol. 13, no. 4, pp. 711–717, Jun. 2019.

[30] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, “An underwater image enhancement benchmark dataset and beyond,” 2019, arXiv:1901.05495. [Online]. Available: http://arxiv.org/abs/1901.05495

[31] C. Li, J. Guo, and C. Guo, “Emerging from water: Underwater image color correction based on weakly supervised color transfer,” IEEE Signal Process. Lett., vol. 25, no. 3, pp. 323–327, Mar. 2018.

[32] C. Li, S. Anwar, and F. Porikli, “Underwater scene prior inspired deep underwater image and video enhancement,” Pattern Recognit., vol. 98, Feb. 2020, Art no. 107038.

[33] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, “Nested network with two-stream pyramid for salient object detection in optical remote sensing images,” 2019, arXiv:1906.08462. [Online]. Available: http://arxiv.org/abs/1906.08462

[34] C. Li, C. Guo, J. Guo, P. Han, H. Fu, and R. Cong, “PDR-Net: Perception-inspired single image dehazing network with refinement,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 704–716, Mar. 2019.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, “Generative adversarial nets,” in Proc. NIPS, 2014, pp. 2672–2680.

[36] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image translation using cycle-consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2223–2232.

[37] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[38] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in Proc. ICLR, Jun. 2013, p. 3.

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4700–4708.

[40] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, 2015, pp. 234–241.

[41] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation with conditional adversarial networks,” in Proc. ICLR, Jun. 2016, pp. 5766–5775.

[42] Q. Qi et al., “Improved training of wasserstein GANs,” in Proc. NIPS, 2016, pp. 694–711.

[43] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image translation using conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1125–1134.

[44] I. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. ECCV, 2016, pp. 694–711.

[45] M. K.-T. Chu and K. K. Ma, “A neural approach to blind motion deblurring,” in Proc. ECV, 2016, pp. 221–235.

[46] L. J. S. Xu Ren, C. Liu, and J. Jia, “Deep convolutional neural network for image deconvolution,” in Proc. Neural Inf. Process. Syst., 2014, pp. 1790–1798.

[47] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 769–777.

[48] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. V. D. Hengel, and Q. Shi, “From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur,” in Proc. CVPR, 2017, pp. 2319–2328.

[49] M. Hradil, J. Koterka, P. Zemcik, and F. Šroubek, “Convolutional neural networks for direct text deblurring,” in Proc. Brit. Mach. Vis. Conf., 2015, p. 2.

[50] M. Jin, M. Hirsch, and P. Favaro, “Learning face deblurring fast and wide,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2018, pp. 745–753.
QIN QI is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, China. Her research interests include image deblurring, and image restoration and its related vision problems.

JICHANG GUO (Member, IEEE) received the B.S. degree from the Nanjing University of Science and Technology, China, in 1988, and the M.S. and Ph.D. degrees in signal and information processing from the School of Electronics and Information Engineering, Tianjin University, Tianjin, China, in 1993 and 2003, respectively. From 2005 to 2006, he was a Senior Visiting Scholar with the Tokyo University of Science, Tokyo, Japan. Since 1993, he has been with the School of Electronics and Information Engineering, Tianjin University, where he is currently a Full Professor. His current research interests include digital video/image intelligent analysis, recognition and processing, and filter theory and design.

WEIPEI JIN is currently pursuing the master’s degree with the School of Electrical and Information Engineering, Tianjin University, China. Her research interests include image deblurring and underwater image restoration.

* * *