BANet: A Blur-Aware Attention Network for Dynamic Scene Deblurring

Fu-Jen Tsai, Yan-Tsung Peng, *Member, IEEE*, Chung-Chi Tsai, Yen-Yu Lin, *Senior Member, IEEE*, and Chia-Wen Lin, *Fellow, IEEE*

***Abstract*— Image motion blur results from a combination of object motions and camera shakes, and such blurring effect is generally directional and non-uniform. Previous research attempted to solve non-uniform blurs using self-recurrent multi-scale, multi-patch, or multi-temporal architectures with self-attention to obtain decent results. However, using self-recurrent frameworks typically leads to a longer inference time, while inter-pixel or inter-channel self-attention may cause excessive memory usage. This paper proposes a Blur-aware Attention Network (BANet), that accomplishes accurate and efficient deblurring via a single forward pass. Our BANet utilizes region-based self-attention with multi-kernel strip pooling to disentangle blur patterns of different magnitudes and orientations and cascaded parallel dilated convolution to aggregate multi-scale content features. Extensive experimental results on the GoPro and RealBlur benchmarks demonstrate that the proposed BANet performs favorably against the state-of-the-arts in blurred image restoration and can provide deblurred results in real-time.**

***Index Terms*— Image deblurring, blur-aware attention module, region-wise pooling attention.**

I. Introduction

DYNAMIC scene deblurring or blind motion deblurring aims to restore a blurred image with little knowledge about the blur kernel. Scene blur caused by camera shakes, object motions, low shutter speeds, or low frame rates not only degrades the quality of taken images/videos but also results in information loss. Therefore, removing such blurring artifacts to recover image details becomes essential to many downstream vision applications, such as facial detection [1], [2], text recognition [3], moving object segmentation [4], etc., where clean and sharp images are appreciated. Although significant progress has been made in conventional and deep-learning-based approaches [5], [6], [7], [8], we observe a compromise between accuracy and speed. Owing to this observation, we target to develop an efficient and effective algorithm in this paper for blurred image restoration with its current performance in accuracy and speed shown in Fig. 1.

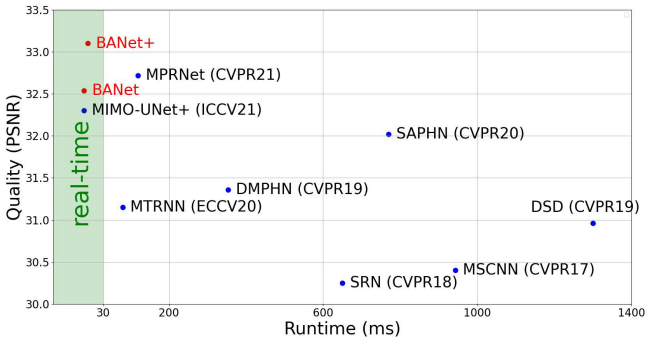


Fig. 1. Performance comparison on the GoPro test dataset in terms of deblurring quality and runtime complexity. The proposed BANet performs favorably against the state-of-the-art methods in both accuracy and efficiency.

Deep-learning-based approaches usually reach superior results, given their better feature representation capability toward dynamic scenes. Among the state-of-the-art architectures for deblurring, self-recurrent models have been widely adopted to leverage blurred image repeatability in either *multiple scales* (MS) [6], [9], [10], [11], *multiple patch levels* (MP) [7], [12], [13], or *multiple temporal behaviors* (MT) [8], as shown in Fig. 2(a)–(c). Specifically, the MS models distill multi-scale blur information in a self-recurrent manner and restore blurred images based on the extracted coarse-to-fine features [6], [9], [10]. However, scaling a blurred image to a lower resolution often results in losing edge information [8]. In contrast, the MP models split a blurred input image into multiple patches to estimate and then remove motion blurs of different scales [7], [13]. However, splitting the blurred input and features into equal-sized non-overlapping patches may cause contextual information

Manuscript received 16 August 2021; revised 12 July 2022 and 9 September 2022; accepted 13 October 2022. Date of publication 26 October 2022; date of current version 31 October 2022. This work was supported in part by the National Science and Technology Council (NSTC) under Grant 110-2634-F-002-050, Grant 111-2628-E-A49-025-MY3, and Grant 111-2221-E-004-010; in part by Qualcomm Technologies, Inc., through a Taiwan University Research Collaboration Project, under Grant NAT-487844; and in part by the National Center for High-performance Computing (NCHC), National Applied Research Laboratories (NARLabs), Hsinchu, Taiwan, for computational and storage resources for this project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wangmeng Zuo. *(Fu-Jen Tsai and Yan-Tsung Peng contributed equally to this work.) (Corresponding author: Chia-Wen Lin.)*

Fu-Jen Tsai is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan (e-mail: fjtsai@gapp.nthu.edu.tw).

Yan-Tsung Peng is with the Department of Computer Science, National Chengchi University, Taipei 116011, Taiwan (e-mail: ytpeng@cs.nccu.edu.tw).

Chung-Chi Tsai is with Qualcomm Technologies, Inc., San Diego, CA 92121 USA (e-mail: chuntsai@qti.qualcomm.com).

Yen-Yu Lin is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan (e-mail: lin@cs.nctu.edu.tw).

Chia-Wen Lin is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300044, Taiwan, and also with the Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan (e-mail: cwlin@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TIP.2022.3216216

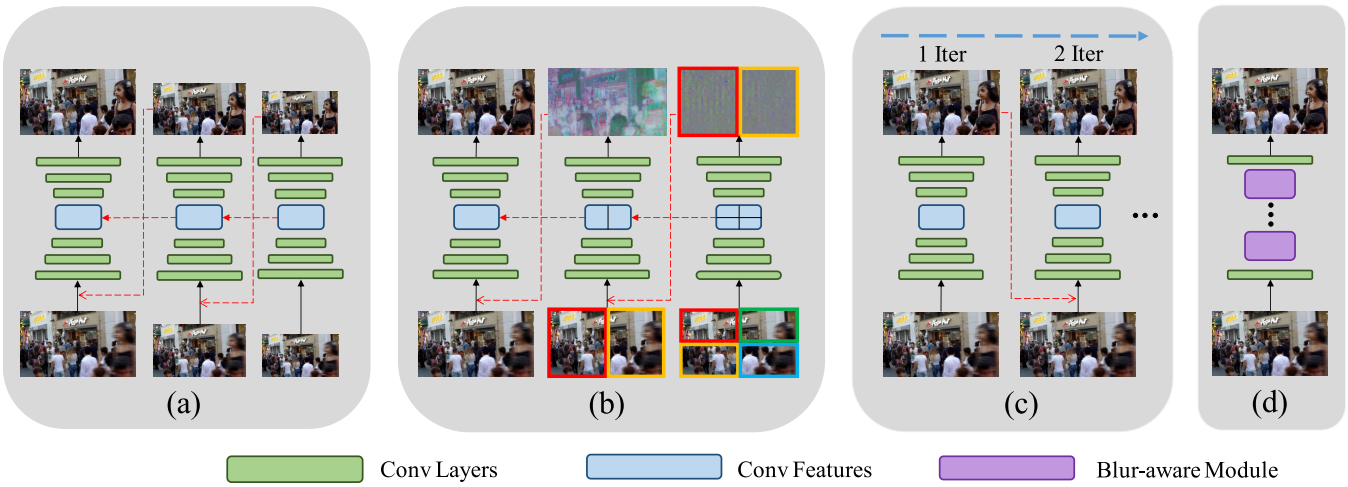


Fig. 2. Network architecture comparisons among (a) MS, (b) MP, (c) MT, and (d) our BANet. Recurrent models are typically less efficient. BANet completes deblurring via a single forward pass.

discontinuity, sub-optimal for handling non-uniform blur in dynamic scenes. In [8], a self-recurrent MT structure was proposed to progressively eliminate non-uniform blurs over multiple iterations. Each iteration would gradually deblur the image until it becomes sharp. However, its inflexible progressive training and inference process may not generalize well for images of varying region-wise blurring degrees. Besides, these existing self-recurrent models, including MS, MP, and MT, cannot achieve high-quality deblurring in real-time (say, 30 HD frames per second).

In addition to model architectures, recent research studies [13], [14] further exploit self-attention to address blur non-uniformity. Suin et al. [13] utilize MP-based processing with self-attention to extract features for areas with global and local motions. However, using a self-recurrent mechanism to generate multi-scale features often leads to a significantly longer inference time. To shorten the latency, Purohit and Rajagopalan [14] selectively aggregate features through learnable pixel-wise attention [15] enabled by deformable convolutions for modeling local blurs in a single forward pass. Despite its effectiveness, self-attention exploring pixel-wise or channel-wise correlations via trainable filters often causes high memory usage, thus only applicable to small-scale features [14]. Furthermore, motion blurs coming from object motions manifest smeared effects and produce directional and local averaging artifacts, which cannot be handled well by inter-pixel/channel correlations.

This paper proposes a *Blur-aware Attention Network* (BANet) to overcome the above-mentioned issues. BANet is an efficient yet effective single-forward-pass model, as illustrated in Fig. 2(d), which achieves state-of-the-art deblurring performance while working in real-time, as shown in Fig. 1. Specifically, our model stacks multiple layers of the *Blur-Aware Module* (BAM) for removing motion blurs. BAM separates the deblurring process into two branches, Blur-aware Attention (BA) and Cascaded Parallel Dilated Convolution (CPDC), where BA locates region-wise blur orientations and magnitudes while CPDC adaptively removes blurs based on the attended blurred features. Based on an observation of directional and regional averaging artifacts caused by dynamic blurs, the proposed BA derives region-wise attention by using computationally inexpensive regional averaging to capture blurred patterns of different orientations and magnitudes globally and locally. To derive the orientations and magnitudes of different blurred regions in an image, we reassemble horizontal and vertical blurred responses to catch irregular blur orientations and utilize multi-scale kernels to learn the magnitudes. CDPC leverages two cascaded multi-scale dilated convolutions to deblur image features. As a result, BANet possesses the superior deblurring capability and can support subsequent real-time applications superbly.

In short, our contributions are two-fold. First, BANet is featured with a novel BAM module that exploits region-wise attention to capture blur orientations and magnitudes, making BANet capable of disentangling blur contents of different degrees in dynamic scenes. With the disentangled region-wise blurred patterns, it then utilizes cascaded multi-scale dilated convolution to restore blurred features. Second, our efficient single-forward-pass deep networks perform favorably against state-of-the-art methods with fast inference time.

II. Related Work

A. Conventional Methods

Dynamic Scene image deblurring is a highly ill-posed problem since blurs stem from various factors in the real world. Conventional image deblurring studies often make different assumptions, such as uniform [16], [17], [18], [19] or non-uniform [20], [21], [22], [23], [24] blurs, and image priors [25], [26], [27], [28], [29], to model blur characteristics. Namely, these methods impose different constraints on the estimated blur kernels, latent images, or both with handcrafted regularization terms for blur removal. Nevertheless, these methods often attempt to solve a non-convex optimization problem and involve heuristic parameter tuning that is entangled with the camera pipeline; thus, they cannot generalize well to complex real-world examples.

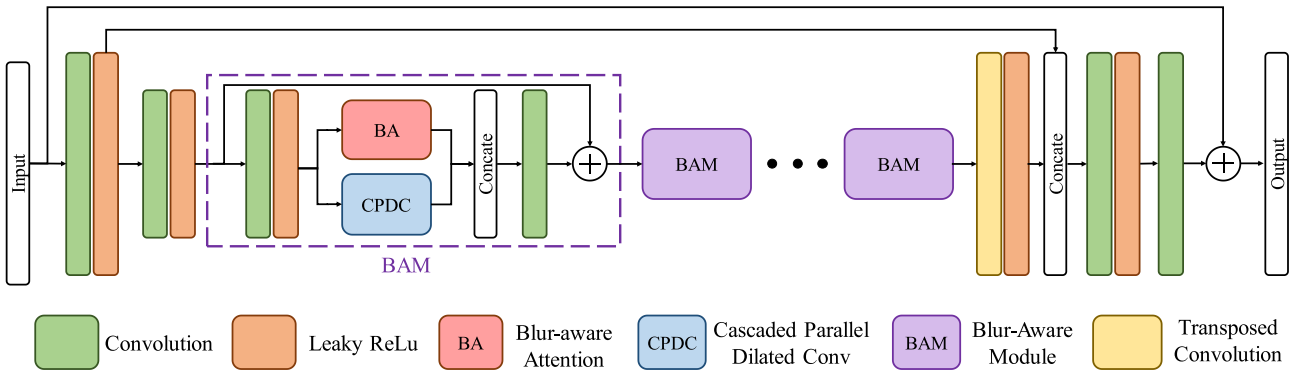


Fig. 3. Architecture of the proposed blur-aware attention networks (BANet). The blur-aware modules (BAM) serve as the building blocks of BANet. The first BAM is detailed in the purple dotted box while the rest are represented by solid purple boxes.

B. Deblurring via Learning

Learning-based approaches with self-recurrent modules gain great success in single-image deblurring. Particularly, the *coarse-to-fine* schemes can gradually restore a sharp image on different resolutions (MS) [6], [9], [10], [11], fields of view (MP) [7], [13], or temporal characteristics (MT) [8]. Despite the success, self-recurrent models usually lead to longer inference runtime. Recently, non-recurrent methods [14], [30], [31], [32], [33], [34] were proposed for efficient deblurring. For instance, Kupyn et al. [30], [31] suggested using conditional generative adversarial networks to restore blurred images. However, these methods do not well address non-uniform blurs in dynamic scenes, often causing blur artifacts in the deblurred images. To address this issue, Yuan et al. [32] proposed a spatially variant deconvolution network with optical flow estimation to guide deformable convolutions and capture moving objects during model training. Li et al. [33] proposed a depth-guided model for deblurring. However, the optical flow and depth information may not always correlate with blur, which may cause less effective deblurring. Cho et al. [34] proposed an efficient multi-scale deblurring structure with a multi-input multi-output. With multi-scale input, the process adopts a shallow convolution to turn the images into attention masks and multiply them by the same scales' features. However, its simple feature attention mechanism may not be able to extract blur information comprehensively from an input image, hence limiting its deblurring performance.

C. Self-Attention

Self-attention (SA) [35] has been widely adopted to advance the fields of image processing [15], [36] and computer vision [37], [38]. Recent advances [13], [14] revealed that attention is beneficial for learning inter-pixel correlations to emphasize different local features for removing non-uniform blur. Specifically, Purohit et al. [14] proposed to deblur using SA to explore pixel-wise correlation for non-local feature adaptation. However, since SA requires much memory in $\mathcal{O}(H^2W^2)$ space, where H and W are the height and width of the input to SA, the method can only apply SA to the smallest-scale features (from a 1280×720 blurred input to 160×90 SA's input), limiting the efficacy of SA. Also, motion blurs cause directional and local averaging artifacts, which merely pixel-wise SA may not address well. Suin et al. [13] proposed an MP architecture with less memory-intensive SA by using global average pooling with space complexity $\mathcal{O}(d_a d_c HW)$, where d_a is the channel dimension of the components *query* and *key* in SA, d_c is the dimension of the component *value*, and $d_a d_c < HW$. Despite the method's less space complexity, compressing pixel information into the channel domain may lose spatial information, thus degrading deblurring performance. In contrast, we propose an efficient and low memory-cost regional averaging SA to capture non-uniform blur information more accurately. It is with space complexity $\mathcal{O}(CHW)$, where C is the number of output channels. It can deblur high-resolution input images and achieve superior performance in real-time.

III. Proposed Approach

We present the blur-aware attention network (BANet) to address the potential issues in two commonly used techniques for deblurring: self-recurrence and self-attention. Self-recurrent algorithms result in longer inference time due to repeatedly accessing input blurred images. Self-attention based on inter-pixel or inter-channel correlations is memory intensive and cannot explicitly capture regional blurring information. Instead, the proposed BANet is a one-pass residual network consisting of a series of stacked blur-aware modules (BAMs), which serve as the building blocks, to disentangle different patterns of blurriness and remove blurs based on the attended blurred features.

As illustrated in Fig. 3, BANet starts with two convolutional layers, which contain a stride of 2 to downsample the input image to half resolution. BANet employs one transposed convolutional layer to upsample features to the original size. In between, we stack a set of BAMs to correlate regions with similar blur and extract multi-scale content features. A BAM consists of two components, BA and CPDC, where BA distills global and local blur orientations and magnitudes, and CPDC captures multi-scale blurred patterns to eliminate blurs adaptively. Combining BA and CPDC, BAM is a residual-like architecture that derives both global and local multi-scale blurring features in a learnable manner. We detail the two key components, BA and CPDC, in the following.

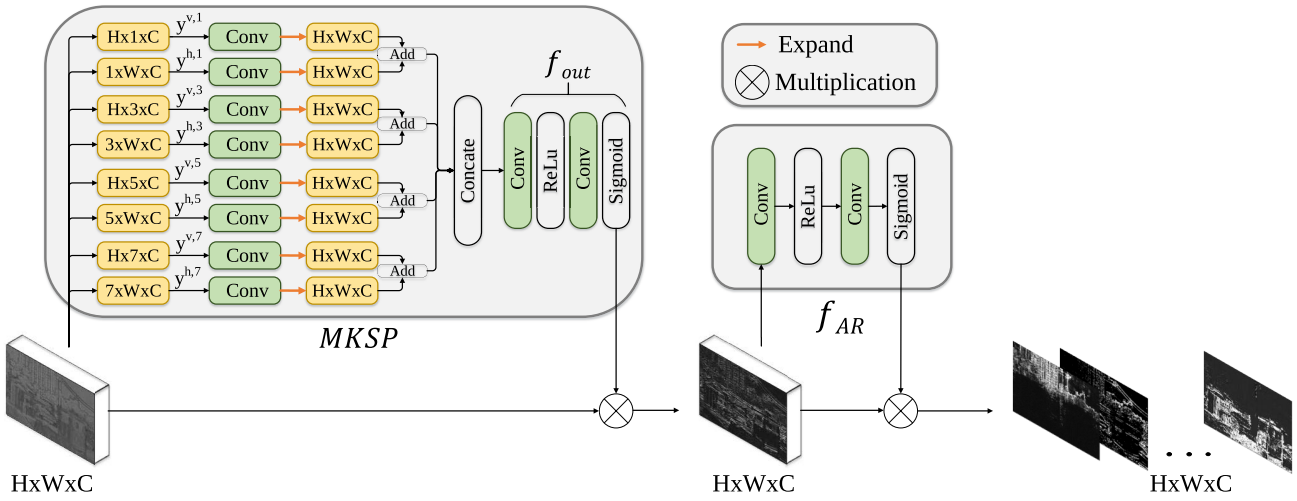


Fig. 4. Architecture of blur-aware attention (BA). It cascades two parts, including multi-kernel strip pooling (MKSP) and attention refinement (AR). It is developed to disentangle blurred contents in an efficient way. See the text for details.

A. Blur-Aware Attention (BA)

To accurately restore the motion area displaying directional and averaging artifacts caused by object motions and camera shakes, we propose a region-based self-attention module, called BA, to capture such effects in the global (image) and local (patch) scales. As shown in Fig. 4, BA contains two cascaded parts: multi-kernel strip pooling (MKSP) and attention refinement (AR). MKSP catches multi-scale blurred patterns of different magnitudes and orientations, followed by AR to refine them locally.

1) Multi-Kernel Strip Pooling (MKSP): Hou et al. [39] presented an SP (strip pooling) method that uses horizontal and vertical one-pixel long kernels to extract long-range band-shape context information for scene parsing. SP averages the input features within a row or a column individually and then fuses the two thin-strip features to discover global cross-region dependencies. Let the input feature maps $\mathbf{x} = [x_{i,j,c}] \in R^{H\times W\times C}$, where C denotes the number of channels. Applying SP to $\mathbf{x}$ generates a vertical and a horizontal tensor followed by a 1D convolutional layer with a kernel size of 3. This produces a vertical tensor $\mathbf{y}^v = [y^v_{i,c}] \in R^{H\times C}$ and a horizontal tensor $\mathbf{y}^h = [y^h_{j,c}] \in R^{W\times C}$, where $y^v_{i,c} = \frac{1}{W}\sum_{j=0}^{W-1} x_{i,j,c}$ and $y^h_{j,c} = \frac{1}{H}\sum_{i=0}^{H-1} x_{i,j,c}$. The SP operation, after a convolution layer, fuses the two tensors into $\mathbf{y} = [y_{i,j,c}] \in R^{H\times W\times C}$, where $y_{i,j,c} = y^v_{i,c} + y^h_{j,c}$, and then turns the fused tensor into an attention mask $\mathbf{M}_{sp}$ as

$$\mathbf{M}_{sp} = \sigma_{sig}(f_1(\mathbf{y})), \tag{1}$$

where f_1 is a 1×1 convolutional layer and $\sigma_{sig}(\cdot)$ is the sigmoid function. Although SP has shown its effects on segmenting band-shape objects for scene parsing, it is unsuitable to directly apply SP to an image deblurring task, aiming to locate blurred patterns that tend to involve different orientations and magnitudes, and restore a sharp image.

Motivated by SP, we propose MKSP that adopts strip pooling with different kernel sizes to discover regional and directional averaging artifacts caused by dynamic blurs.

MKSP combines and compares multiple sizes/scales of averaging results followed by concatenation and convolution to catch blurred patterns of different magnitudes and orientations. The idea behind our design is to reassemble different orientations by horizontal and vertical operations on multi-scale results, e.g., the difference between consecutive kernel sizes, and reveal the scales of blurred patterns. We apply convolutional layers to automatically discover these blur-aware operations on the feature level to learn irregular attended features rather than a fixed cropping method on the image level used in MP methods [7], [13]. MKSP averages the input tensors within rows and columns by adaptive average pooling to generate $H\times n\times C$ and $n\times W\times C$ long features, where $n\in\{1,3,5,7\}$ represents different scales. Thus, MKSP generates four pairs of tensors, each of which has a vertical and a horizontal tensor followed by a 1D (for $n=1$) or 2D (for the rest) convolutional layer with the kernel size of 3 or 3×3, respectively. This produces the vertical tensor $\mathbf{y}^{v,n} \in R^{H\times n\times C}$ and the horizontal tensor $\mathbf{y}^{h,n} \in R^{n\times W\times C}$, where the vertical tensor is

$$y^{v,n}_{i,j,c} = \frac{1}{K_h}\sum_{k=0}^{K_h-1} x_{i,(j\cdot S_h+k),c}, \tag{2}$$

where the horizontal stride $S_h = \lfloor \frac{W}{n} \rfloor$ and the horizontal-strip kernel size $K_h = W-(n-1)S_h$. Symmetrically, the horizontal tensor is defined by

$$y^{h,n}_{i,j,c} = \frac{1}{K_v}\sum_{k=0}^{K_v-1} x_{(i\cdot S_v+k),j,c}, \tag{3}$$

where the vertical stride $S_v = \lfloor \frac{H}{n} \rfloor$ and the vertical-strip kernel size $K_v = H-(n-1)S_v$.

After determining the horizontal and vertical magnitudes, the orientations of blur patterns are estimated jointly considering the two orthogonal magnitudes. More specifically, MKSP, after a 1D (for $n=1$) or 2D (for the rest) convolutional layer, fuses each pair of tensors ($\mathbf{y}^{v,n}$, $\mathbf{y}^{h,n}$) into a tensor

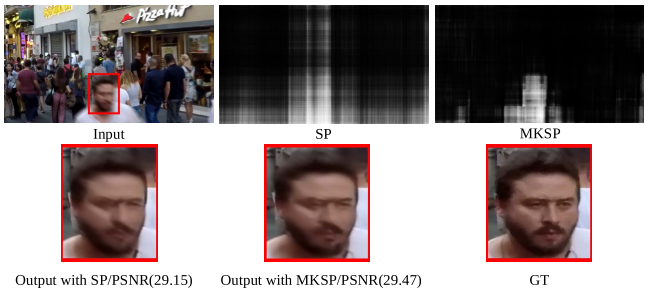


Fig. 5. Visualization of the attention masks of SP and MKSP, and the corresponding effects on the final results on GoPro test set.

$\mathbf{y}^n \in R^{H\times W\times C}$ by

$$y^n_{i,j,c} = y^{v,n}_{i,\lfloor\frac{n\times j}{W}\rfloor,c} + y^{h,n}_{\lfloor\frac{n\times i}{H}\rfloor,j,c}. \tag{4}$$

Similar to SP, we concatenate all the fused tensors to yield an attention mask as $\mathbf{M}_{mksp} = f_{out}(\mathbf{y}^1 \oplus \mathbf{y}^3 \oplus \mathbf{y}^5 \oplus \mathbf{y}^7)$, where $\oplus$ stands for the concatenation operation, and $f_{out}(\cdot) = \sigma_{sig}(Conv(\sigma_{ReLU}(Conv(\cdot))))$ represents a non-linear mapping function consisting of two 3×3 convolutional layers. The first layer uses the ReLU activation function, and the second uses a sigmoid function. As shown in Fig. 5, the proposed MKSP can generate attention masks that better fit objects or local scenes than those by using SP with only $H \times 1$ and $1 \times W$ kernels used, which yields rough band-shape masks.

2) Attention Refinement (AR): After obtaining the globally attended features by the element-wise multiplication of attention masks $\mathbf{M}_{mksp}$ and input tensor $\mathbf{x}$, we further refine these features locally via a simple attention mechanism using $f_{AR}(\cdot)$. The final output of our BA block through the MKSP and AR stages is computed as

$$f_{AR}(\tilde{\mathbf{x}}) \otimes \tilde{\mathbf{x}}, \tag{5}$$

where $\otimes$ represents element-wise multiplication, and $\tilde{\mathbf{x}} = \mathbf{M}_{mksp} \otimes \mathbf{x}$ denotes the global features extracted using MKSP. Figs. 6(c) and (d) demonstrate that cascading MKSP with AR can refine the attended feature maps.

The proposed BA facilitates the attention mechanism applied to deblurring since it requires less memory, i.e.$\mathcal{O}(HWC)$, where C represents the channel dimension, than those adopted in [14] and [13]. It disentangles blurred contents with different magnitudes and orientations. Fig. 7 showcases three examples of blur content disentanglement using BA, where we witness that background scenes are differentiated from the foreground scenes because those objects closer to the camera move faster, thus more blurred. Fig. 8 shows more examples of attention maps yielded by BA, which implicitly acts as a gate for propagating relevant blur contents.

B. Cascaded Parallel Dilated Convolution (CPDC)

Atrous convolution, also called *dilated convolution*, has been widely applied to computer-vision tasks [40], [41] for enlarging receptive fields and extracting features from objects with different scales without increasing the kernel size. Inspired by this, we design a *cascaded parallel dilated convolution* (CPDC) block with multiple dilation rates to capture multi-scale blurred objects. Instead of stacking dilated convolutional layers with different rates in parallel, which we call *parallel dilated convolution* (PDC), our CPDC block cascades two sets of PDC with a single convolutional layer working as a fusion bridge. It can distill patterns more beneficial to deblurring before passing through the second PDC. As an example, Fig. 9(a) shows a PDC block consisting of three 3×3 dilated convolutional layers with a dilation rate D ($D = 1$, 3, and 5), each of which outputs features with half the number of input channels. After concatenation, the number of the output channels of the PDC block increases by 1.5 times. As shown in Fig. 9(b), our CPDC block consists of two PDC blocks bridged by a 3×3 convolutional layer, which would be more effective in aggregating multi-scale content information for deblurring.

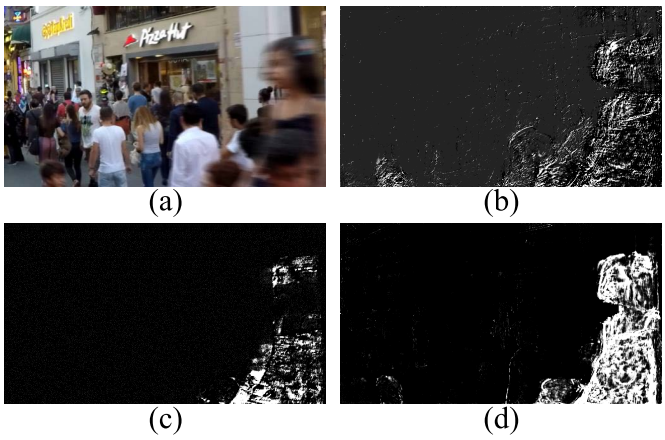


Fig. 6. (a) Input blurred image in GoPro testing set. (b)–(d) Comparisons among the attended feature maps by using different components of the proposed BA including (b) AR, (c) MKSP, and (d) MKSP + AR.

Fig. 7. Three disentanglement examples of blurred patterns of different degrees using our BA on GoPro test set. (a) Input blurred images and (b) attended feature maps on different regions.

C. Loss Function

In BANet, we utilize the Charbonnier loss as suggested in [42] and [12]:

$$L_{char} = \sqrt{||\mathbf{R} - \mathbf{Y}||^2 + \varepsilon^2}, \tag{6}$$

where $\mathbf{R}$ and $\mathbf{Y}$ respectively denote the restored image and the ground-truth image, and $\varepsilon = 10^{-3}$ as in [42] and [12].

Fig. 8. Visualization of the blur-aware attended features on GoPro test set, where moving objects in the blurred images are highlighted while background is mostly excluded. These blur-aware masks are crucial for handling blurred images with diverse blur patterns.

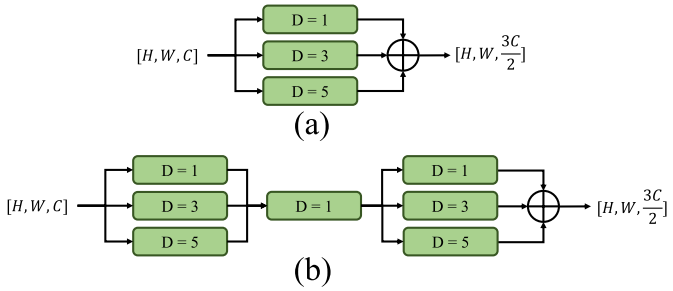


Fig. 9. Architectures of (a) parallel dilated convolution (PDC) and (b) cascaded parallel dilated convolution (CPDC).

In addition, to enhance the restoration performance, we add an FFT loss to supervise the results in the frequency domain, as adopted in MIMO-UNet+ [34]:

$$L_{FFT} = ||\mathcal{F}(\mathbf{R}) - \mathcal{F}(\mathbf{Y})||_1, \tag{7}$$

where $\mathcal{F}$ represents the fast Fourier transform function. At last, we optimize BANet using the total loss L as

$$L = L_{char} + \lambda L_{FFT}, \tag{8}$$

where λ is set to 0.01 empirically.

IV. Experiments

This section evaluates the proposed method. In the following, we first describe the experimental setup, then compare our method with the state-of-the-arts, and finally conduct ablation studies to analyze the effectiveness of individual components.

A. Experimental Setup

We evaluate the BANet on three image deblurring benchmark datasets: 1) GoPro [6] that consists of 3,214 pairs of blurred and sharp images of resolution 720 × 1280, where 2,103 pairs are used for training, and the rest for testing, 2) HIDE [43] that contains 2,025 pairs of HD images, all for testing, and RealBlur [44] that consists of 3,758 pairs for training and 980 pairs for testing. The RealBlur dataset is further split into two datasets: RealBlur-R collected from raw images and RealBlur-J from JPEG images. We train our model using Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the initial learning rate to 10^{-4}, which then decays to 10^{-7} based on the cosine annealing strategy. Following [8], [32], we utilize random cropping, flipping, and rotation for data augmentation. Lastly, we implement our model with PyTorch library on a computer equipped with Intel Xeon Silver 4210 CPU and NVIDIA 2080ti GPU.

TABLE I

Evaluation Results on GoPro Test Set. The Best Score in Its Column Is Highlighted in Bold and the Second Best Is Underlined. Symbol ∗ Indicates Those Methods Without Released Code; Thus We Cite the Results From the Original Papers or Evaluate on the Released Deblurred Images. All Methods Are Trained on GoPro Training Set. Time and Params Are Measured in Millisecond (ms) and Million (M)

Model	PSNR ↑	SSIM ↑	Time ↓	Params ↓	GFLOPs ↓
MSCNN [6]	30.40	0.936	943	12	336
SRN [10]	30.25	0.934	650	7	167
DSD [9]	30.96	0.942	1300	3	471
DeblurGAN-v2 [31]	29.55	0.934	42	68	**42**
DMPHN [7]	31.36	0.947	354	22	235
EDSD* [32]	29.81	0.934	**10**	**1**	–
MTRNN [8]	31.13	0.944	53	3	164
RADN* [14]	31.85	0.953	38	–	–
SAPHN* [13]	32.02	0.953	770	–	–
MIMO-UNet+ [34]	32.45	0.957	23	16	154
MPRNet [12]	32.66	0.959	138	20	760
BANet	32.54	0.957	23	18	264
BANet+	**33.03**	**0.961**	25	40	588

B. Experimental Results

1) Quantitative Analysis: We compare our method with 11 latest approaches, including MSCNN [6], SRN [10], DSD [9], DeblurGAN-v2 [31], DMPHN [7], EDSD [32], MTRNN [8], RADN [14], SAPHN [13], MIMO-UNet+ [34], and MPRNet [12], which also handle dynamic deblurring on the GoPro [6] test set. For HIDE [43], we choose nine recent deblurring methods, including DeblurGAN-v2 [31], SRN [10], HAdeblur [43], DSD [9], DMPHN [7], MTRNN [8], SAPHN [13], MIMO-UNet+ [34], and MPRNet [12], according to their availability in released pre-trained weights. For RealBlur [44], we choose four methods that trained on the RealBlur training set, including DeblurGAN-v2 [31], SRN [10], MPRNet [12], and MIMO-UNet+ [34].

Fig. 10. Qualitative comparisons on GoPro [6] test set. The deblurred results listed from left to right are from MTRNN [8], DSD [9], DMPHN [7], MIMO-UNet+[34], MPRNet [12], and Ours.

Fig. 11. Qualitative comparisons on HIDE [43] dataset. The deblurred results listed from left to right are from MTRNN [8], DSD [9], DMPHN [7], MIMO-UNet+[34], MPRNet [12], and Ours.

TABLE II

EVALUATION RESULTS ON HIDE DATASET. THE BEST SCORE IN ITS COLUMN IS HIGHLIGHTED IN BOLD AND THE SECOND BEST IS UNDERLINED. SYMBOL ∗ INDICATES THOSE METHODS WITHOUT RELEASED CODE; THUS WE CITE THE RESULTS FROM THE ORIGINAL PAPERS OR EVALUATE ON THE RELEASED DEBLURRED IMAGES. ALL METHODS ARE TRAINED ON GOPRO TRAINING SET. TIME IS MEASURED IN MILLISECOND (MS)

Model	PSNR ↑	SSIM ↑	Time ↓
DeblurGAN-v2 [31]	27.40	0.882	42
SRN [10]	28.36	0.904	424
HAdeblur* [43]	28.87	0.903	–
DSD [9]	29.10	0.913	1200
DMPHN [7]	29.10	0.918	341
MTRNN [8]	29.15	0.918	53
SAPHN* [13]	29.98	0.930	–
MIMO-UNet+ [34]	30.00	0.930	28
MPRNet [12]	**30.93**	**0.939**	138
BANet	30.16	0.930	**23**
BANet+	<u>30.58</u>	<u>0.935</u>	<u>25</u>

TABLE III

EVALUATION RESULTS ON REALBLUR TEST SET. ALL METHODS ARE TRAINED ON REALBLUR TRAINING SET. TIME IS MEASURED IN MILLISECOND (MS)

Model	RealBlur-J		RealBlur-R		
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	Time
DeblurGAN-v2 [31]	29.69	0.870	36.44	0.935	44
SRN [10]	31.38	0.901	38.65	0.965	420
MPRNet [12]	31.76	0.922	39.31	**0.972**	81
MIMO-UNet+ [34]	31.92	0.919	–	–	<u>23</u>
BANet	<u>32.00</u>	<u>0.923</u>	<u>39.55</u>	<u>0.971</u>	**22**
BANet+	**32.42**	**0.929**	**39.90**	**0.972**	24

To better compare with recent approaches, we devise two versions of our model, BANet and BANet+. The only difference between them is the number of channels used in a BAM, and BANet with 128 channels involves 18 million parameters while BANet+ has 40 million parameters with 192 channels. Table I lists the objective scores (PSNR and SSIM), runtime, parameters, and GFLOPs on the GoPro test set for all the compared methods. We observe that the self-recurrent models, MSCNN [6], SRN [10], DSD [9], MTRNN [8], SAPHN [13], and MPRNet [12], consume longer runtime than the non-recurrent ones, i.e., DeblurGAN-v2 [31], RADN [14], MIMO-UNet+ [34], and ours. As reported in Table I, BANet runs faster with fewer parameters and GFLOPs as well as achieves better performance than recurrent-based methods, MSCNN [6], SRN [10], DSD [9], DMPHN [7], MTRNN [8], and SAPHN [13] and non-recurrent methods, such as DeblurGAN-v2 [31] and RADN [14] on the GoPro test set. BANet also performs favorably against an efficient multi-scale model, MIMO-UNet+ [34], with the same runtime and a comparable model size. BANet+ outperforms the best competitor, MPRNet [12], by 0.37 dB in PSNR with faster runtime (−113ms) and lower GFLOPs (−172). Table II shows the quantitative results on HIDE [43]. As can be seen, BANet outperforms all the compared methods except for MPRNet [12] with a faster inference time. BANet+ only works comparably to MPRNet [12] since MPRNet seems to perform favorably on HIDE [43] particularly, but our model runs much faster. Table III lists the quantitative comparisons on the RealBlur test set, demonstrating that both BANet and BANet+ outperform the compared methods on the RealBlur-J and RealBlur-R test sets.

Fig. 12. Examples of deblurred results obtained using DeblurGAN-v2 [31], SRN [10], MPRNet [12], MIMO-UNet+ [34], and Ours on RealBlur [44] test set.

2) Qualitative Analysis: Fig. 10 and Fig. 11 show qualitative comparisons on the GoPro test set and HIDE dataset with previous state-of-the-arts MTRNN [8], DSD [9], DMPHN [7], MIMO-UNet+ [34], and MPRNet [12]. As observed in Fig. 10, MTRNN [8], DSD [9], DMPHN [7], MIMO-UNet+ [34], and MPRNet [12] do not well recover regions with texts or severe blurs whereas BANet can restore those regions better. In Fig. 11, MTRNN [8], DSD [9], DMPHN [7], and MIMO-UNet+ [34] do not deblur the striped t-shirt and texts well, while BANet recovers those parts better. Fig. 12 and Fig. 13 demonstrate some deblurred results using DeblurGAN-v2 [31], SRN [10], MPRNet [12], MIMO-UNet+ [34], and ours, on the RealBlur [44] test set. As can be seen, although all these models can remove blurs, BANet performs favorably on delicate image details.

3) User Study: We further conduct a user study to evaluate the subjective quality of deblurred results on real blurred images chosen from the RealBlur-J test set. We compare our method (BANet+) against four methods, including MIMO-UNet+ [34], MPRNet [12], SRN [10], and DeblurGAN-v2 [31]. Note that all the methods are trained on the RealBlur-J training set.

In the study, 34 subjects aged from 21 to 40 years participated in the study without any prior knowledge of the experiment. Their vision is either normal or corrected to be normal. We picked 16 blurred images with varying scenes for the experiment and obtained the deblurred results using all the compared approaches. Since each method is compared against BANet with all the chosen blurred images in the experiment, we have $16 \times 4 = 64$ image pairs in total. Each subject is shown all the image pairs, one at a time, and asked which one he/she prefers in terms of visual quality. Each image pair is displayed randomly and placed side by side. Subjects are asked to check images carefully before choosing without a time limit.

TABLE IV

RESULTS OF USER STUDY. THE VALUES REPRESENT THE PERCENTAGE THAT OUR METHOD WAS CHOSEN OVER THE OTHER COMPARED METHODS

	MIMO-UNet+ [34]	MPRNet [12]	SRN [10]	DeblurGAN-v2 [31]
Ours	95.6%	95.6%	98.5%	99.6%

Table IV shows the subjective evaluation results, where the values represent the percentage that the deblurring results with our method are preferred to the counterparts with the other compared methods for all the votes collected. It indicates that our method obtains over 95% preference votes compared to all the compared methods, which again demonstrates that our approach achieves better subjective visual quality.

C. Ablation Study

In the ablation studies, we do all the experiments on the BANet (18M) version.

Fig. 13. Examples of deblurred results obtained using DeblurGAN-v2 [31], SRN [10], MPRNet [12], MIMO-UNet+ [34], and Ours on RealBlur [44] test set.

TABLE V

ABLATION STUDY ON GOPRO TEST SET USING DIFFERENT COMPONENT COMBINATIONS IN BAM

Model	PDC	AR	SP	MKSP	CPDC	PSNR
Net1	√					31.39
Net2	√	√				31.80
Net3	√		√			31.81
Net4	√			√		32.11
Net5	√	√		√		32.24
Net6		√		√	√	32.54

TABLE VI

PERFORMANCE COMPARISONS OF THE STACKING NUMBER OF BAMS IN BANET ON GOPRO TEST SET

BANet	stack-4	stack-8	stack-10	stack-12
PSNR	31.36	32.37	32.54	**32.55**

TABLE VII

PERFORMANCE COMPARISONS OF STRIP POOLING (SP) AND MKSP ON GOPRO TEST SET WITH PDC

BANet	SP	$MKSP_{135}$	$MKSP_{1357}$	$MKSP_{13579}$
PSNR	31.81	32.03	**32.11**	32.04

1) BAM With Different Components: Table V shows an ablation on different component combinations in our Blur-Aware Module (BAM) tested on the GoPro test set. As can be seen, adding a simple attention refinement (AR) mechanism to PDC (Net1 vs. Net2) can improve PSNR by 0.41 dB, which shows the effectiveness of spatial attention for deblurring. Using MKSP in PDC (Net1 vs. Net4) improves PSNR by 0.72 dB, which has much more performance gain compared to using strip pooling (SP) [39] in Net3 or AR in Net2. Substituting PDC in Net5 with CPDC (Net6), our proposed version of BAM leads to a further performance gain. Thanks to its mechanism for locating blur regions based on both global attention and local convolutions, our BAM attains the best performance while achieving fast inference time.

2) Numbers of Stacked BAMs: Using more layers to enlarge the receptive field may improve performance for computer vision or image processing tasks. Nevertheless, stacking more layers for deblurring does not guarantee better performance [13] and might consume extra inference time. However, using our residual learning-based BAM design, we can stack multiple layers to expand the effective receptive field for better deblurring. In Table VI, we show performance comparisons with various numbers of BAMs stacked in our model on the GoPro test set. We list four versions: stack-4, stack-8, stack-10, and stack-12, corresponding to 4, 8, 10, and 12 BAMs stacked in BANet. Although the quantitative performance improves with the number of BAMs, the improvement became saturated after 12. Therefore, we choose 10 for its excellent balance between efficiency and visual quality.

3) Effectiveness of MKSP and CPDC: In Table VII, we investigate the effects of kernel combination of MKSP on the GoPro test set. MKSP with five kernel sizes of 1, 3, 5, 7, and 9 performs a little worse than the first four sizes (1, 3, 5, and 7), indicating that adding the kernel size of 9 would not catch blur features more accurately, thus not helping with

TABLE VIII

ABLATION STUDY OF CPDC (W/O BA) COMPARED TO PDC (W/O BA) ON GOPRO TEST SET

	PDC_{135}	PDC^2_{135}	CDPC
PSNR	31.39	31.78	**32.13**
Parms (M)	**6**	10	10

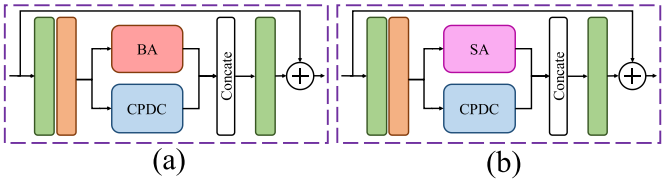


Fig. 14. Architecture comparisons between (a) our original BAM and (b) BA replaced by SA [15] in BAM.

TABLE IX

PERFORMANCE COMPARISON BETWEEN BA AND SA [15] USING BANET (STACK-4) ON GOPRO TEST SET. $*$ REPRESENTS DEBLURRING ON EIGHT SUB-IMAGES INSTEAD OF AN ENTIRE IMAGE

	SA*	BA*	BA
PSNR	31.11	31.09	**31.36**
Time (ms)	770	16	**12**

the performance. In Table VIII, we verify that CPDC, which uses a single convolution as a fusion bridge, outperforms PDC. For a fair comparison, we also compare CPDC against a PDC variant that stacks two PDCs in a series, called PDC^2, with a similar parameter size, and CPDC still performs better.

D. Blur-Aware Attention vs. Self-Attention

RADN [14] utilizes a similar self-attention (SA) mechanism proposed in [15] for deblurring. It helps connect regions with similar blurs to facilitate global access to relevant features across the entire input feature maps. However, its high memory usage makes applying it to high-resolution images infeasible. Thus, SA is usually employed in network layers on a smaller scale like in RADN [14], where important blur information would be lost due to down-sampling. In contrast, our proposed region-based attention is more suitable for correlating regions with similar blur characteristics. Moreover, it can process high-resolution images thanks to its low memory consumption. To further demonstrate our BA's efficacy, we compare the SA [15] with BA using our BANet (stack-4) as a backbone network, as shown in Fig. 14(b). Due to the high memory demand for SA ($\mathcal{O}(H^2W^2)$) to process 720×1280 images, we adopt our stack-4 model for training. When testing the networks, we separate the input image into eight sub-images for both SA and BA to deblur, each equipped with a single 2080ti GPU. Since our BA requiring lower memory usage ($\mathcal{O}(CHW)$, where $C \ll H \times W$) can process the image with the full resolution, we also show its result. In Table IX, SA* and BA* represent deblurring an image with its eight sub-images separately, whereas BA for processing the entire image at once. As can be observed, the proposed BA* works much more efficiently than SA* with a comparable result. When deblurring the entire image at once, BA undoubtedly performs the best.

V. CONCLUSION

This paper proposes a novel blur-aware attention network (BANet) for single image deblurring. BANet consists of stacked blur-aware modules (BAMs) to disentangle region-wise blur contents of different magnitudes and orientations and aggregate multi-scale content features for more accurate and efficient dynamic scene deblurring. We have investigated and examined our design through demonstrations of attention masks and attended feature maps, as well as extensive ablation studies and performance comparisons. Our extensive experiments demonstrate that the proposed BANet achieves real-time deblurring and performs favorably against state-of-the-art deblurring methods on the GoPro and RealBlur benchmark datasets.

REFERENCES

[1] D. Tian and D. Tao, "Coupled learning for facial deblur," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 961–972, Feb. 2016.

[2] R. Yasarla, F. Perazzi, and V. M. Patel, "Deblurring face images using uncertainty guided multi-stream semantic networks," *IEEE Trans. Image Process.*, vol. 29, pp. 6251–6263, 2020.

[3] H. Lee, C. Jung, and C. Kim, "Blind deblurring of text images using a text-specific hybrid dictionary," *IEEE Trans. Image Process.*, vol. 29, pp. 710–723, 2020.

[4] L. Pan, Y. Dai, M. Liu, F. Porikli, and Q. Pan, "Joint stereo video deblurring, scene flow estimation and moving object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 1748–1761, 2020.

[5] X. Xu, J. Pan, Y.-J. Zhang, and M.-H. Yang, "Motion blur kernel estimation via deep learning," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 194–205, Jan. 2018.

[6] S. Nah, K. 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. 3883–3891.

[7] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5978–5986.

[8] D. Park, D. U. Kang, J. Kim, and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020, pp. 327–343.

[9] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3848–3856.

[10] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8174–8182.

[11] J. Cai, W. Zuo, and L. Zhang, "Dark and bright channel prior embedded network for dynamic scene deblurring," *IEEE Trans. Image Process.*, vol. 29, pp. 6885–6897, 2020.

[12] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 14821–14831.

[13] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3606–3615.

[14] K. Purohit and A. Rajagopalan, "Region-adaptive dense network for efficient motion deblurring," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 11882–11889.

[15] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2019, pp. 7354–7363.

[16] S. Cho and S. Lee, "Fast motion deblurring," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–8, Dec. 2009.

[17] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," in *Proc. ACM SIGGRAPH Papers*, Jul. 2006, pp. 787–794.

[18] L. Xu and J. Jia, "Two-phase kernel estimation for robust motion deblurring," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2010, pp. 157–170.

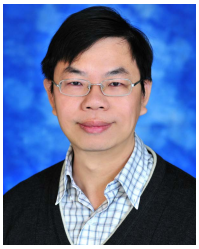
[19] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–10, Aug. 2008.
[20] A. Gupta, N. Joshi, C. L. Zitnick, M. Cohen, and B. Curless, "Single image deblurring using motion density functions," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2010, pp. 171–184.
[21] S. Harmeling, H. Michael, and B. Schölkopf, "Space-variant single-image blind deconvolution for removing camera shake," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2010, pp. 829–837.
[22] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Schölkopf, "Fast removal of non-uniform camera shake," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 463–470.
[23] T. H. Kim and K. M. Lee, "Segmentation-free dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2766–2773.
[24] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 491–498.
[25] L. Chen, F. Fang, T. Wang, and G. Zhang, "Blind image deblurring with local maximum gradient prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1742–1750.
[26] N. Joshi, C. L. Zitnick, R. Szeliski, and D. J. Kriegman, "Image deblurring and denoising using color priors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1550–1557.
[27] J. Pan, Z. Hu, Z. Su, and M.-H. Yang, "Deblurring text images via L_0 -regularized intensity and gradient prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2901–2908.
[28] J. Pan, D. Sun, H. Pfister, and M.-H. Yang, "Deblurring images via dark channel prior," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2315–2328.
[29] Y. Yan, W. Ren, Y. Guo, R. Wang, and X. Cao, "Image deblurring via extreme channels prior," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 6978–6986.
[30] 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.
[31] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-V2: Deblurring (orders-of-magnitude) faster and better," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8877–8886.
[32] Y. Yuan, W. Su, and D. Ma, "Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3552–3561.
[33] L. Li, J. Pan, W.-S. Lai, C. Gao, N. Sang, and M.-H. Yang, "Dynamic scene deblurring by depth guided model," *IEEE Trans. Image Process.*, vol. 29, pp. 5273–5288, 2020.
[34] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4641–4650.
[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 6000–6010.
[36] N. Parmar et al., "Image transformer," 2018, *arXiv:1802.05751*.
[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2018, pp. 7132–7141.
[38] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[39] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. CVPR*, Jun. 2020, pp. 4002–4011.
[40] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[41] S. Liu, D. Huang, and Y. Wang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 404–419.
[42] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Mar. 2017, pp. 624–632.
[43] Z. Shen, W. Wang, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5571–5580.
[44] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Proc. Euro. Conf. Comput. Vis.*, Aug. 2020, pp. 184–201.

Fu-Jen Tsai received the B.S. degree in electrical engineering from the National Taiwan Ocean University, Keelung, Taiwan, in 2019, and the M.S. degree in communications engineering from the National Tsing Hua University, Hsinchu, Taiwan, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. His research interests include image/video restoration and computer vision.

Yan-Tsung Peng (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of California, San Diego, in 2017. In February 2019, he joined the National Chengchi University, where he is currently an Assistant Professor with the Department of Computer Science. Before that, he was a Senior Engineer with Qualcomm Technologies, Inc., San Diego, CA, USA. His research interests include image processing, video compression, and machine-learning applications.

Chung-Chi (Charles) Tsai received the B.S. degree in electrical engineering from the National Tsing-Hua University, Hsinchu, Taiwan, in 2009, the M.S. degree in electrical engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2012, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2018. He received full scholarship from the Ministry of Education, Taiwan, to attend a one-year exchange program, at the University of New Mexico, Albuquerque, NM, USA, in 2007. He also participated in the summer internship with MediaTek in 2013, 2015, and 2016. He is currently a Staff Engineer at the Camera Machine Learning Team, Qualcomm Technologies, Inc., San Diego, CA, USA. His research interests include image processing and computer vision.

Yen-Yu Lin (Senior Member, IEEE) received the B.B.A. degree in information management and the M.S. and Ph.D. degrees in computer science and information engineering from the National Taiwan University, Taipei, Taiwan, in 2001, 2003, and 2010, respectively. He is currently a Professor with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. His research interests include computer vision, machine learning, and artificial intelligence.



Chia-Wen Lin (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2000. He is currently a Professor with the Department of Electrical Engineering and the Institute of Communications Engineering, NTHU. He is also a Research and Development Director with the Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Hsinchu. His research interests include image/video processing and computer vision. He has served as a Fellow Evaluating Committee Member from 2021 to 2022, a BoG Member-at-Large from 2022 to 2024, and a Distinguished Lecturer from 2018 to 2019 for IEEE Circuits and Systems Society. He received Two Best Paper Award from VCIP 2010 and VCIP 2015. He was the Chair of the IEEE ICME Steering Committee from 2020 to 2021. He has served as the TPC Co-Chair for IEEE ICIP 2019 and IEEE ICME 2010 and the General Co-Chair for IEEE VCIP 2018. He has also served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE MULTIMEDIA.