

# BA-Net: Blur-aware Attention Networks for Dynamic Scene Deblurring

Fu-Jen Tsai<sup>\*1</sup>, Yan-Tsung Peng<sup>\*2</sup>, Yen-Yu Lin<sup>3</sup>, Chung-Chi Tsai<sup>4</sup>, and Chia-Wen Lin<sup>1</sup>

<sup>1</sup>National Tsing Hua University <sup>2</sup>National Chengchi University

<sup>3</sup>National Chiao Tung University <sup>4</sup>Qualcomm Technologies, Inc.

\*equal contribution

## Abstract

*Image motion blur usually results from moving objects or camera shakes. Such blur is generally directional and non-uniform. Previous research efforts attempt to solve non-uniform blur by using self-recurrent multi-scale or multi-patch architectures accompanying with self-attention. 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 blur-aware attention networks (BA-Net) that accomplish accurate and efficient deblurring via a single forward pass. Our BA-Net utilizes region-based self-attention with multi-kernel strip pooling to disentangle blur patterns of different degrees and with cascaded parallel dilated convolution to aggregate multi-scale content features. Extensive experimental results on the GoPro and HIDE benchmarks demonstrate that the proposed BA-Net performs favorably against the state-of-the-art in blurred image restoration and can provide deblurred results in real-time.*

## 1. Introduction

Dynamic scene deblurring or blind motion deblurring aims to restore a blurred image with little knowledge about the blur kernel. Scene blurring caused by camera shakes, moving objects, low shutter speeds, or low frame rates not only degrades the quality of the taken images/videos but also results in information loss. Therefore, removing such blurring artifacts to recover image details becomes essential to many subsequent vision applications where clean and sharp images are appreciated. Although significant progress has been made in both conventional and deep-learning-based approaches, we observe a compromise between accuracy and speed. Owing to this observation, we target at developing an efficient and effective algorithm in this paper

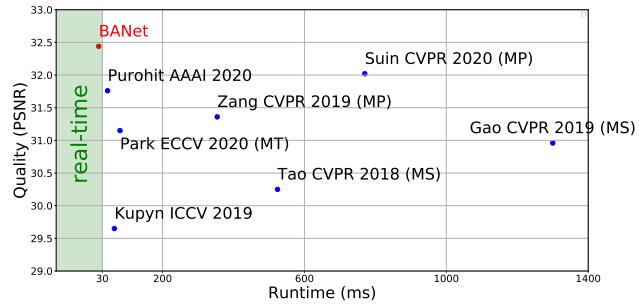


Figure 1. Performance comparison on the GoPro test dataset in deblurring quality and runtime. The proposed BA-Net performs favorably against the state-of-the-art methods in both accuracy and efficiency.

for blurred image restoration with its current performance in accuracy and speed shown in Figure 1.

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, the self-recurrent module is widely used to leverage blurred image repeatability in either *multiple scales* (MS) [3, 12, 20], *multiple patch levels* (MP) [19, 24], or *multiple temporal behaviors* (MT) [15], as shown in Figure 2(a)–(c). MS models distill multi-scale blur information in a self-recurrent manner and restore blurred images using the resultant coarse-to-fine features [3, 12, 20]. However, scaling a blurred image to a lower resolution often results in blur information loss, so it may not help deblurring much. MP models split an input blurred image into multiple patches to estimate and remove motion blur of different scales [19, 24]. Yet, patches heuristically separated are sub-optimal for handling non-uniform blur in dynamic scenes. In [15], an MT structure is proposed to eliminate non-uniform blur and can generate better results progressively. Despite their effectiveness, existing recurrent models are generally far from real-time deblurring.

In addition to model architectures, research studies [17,

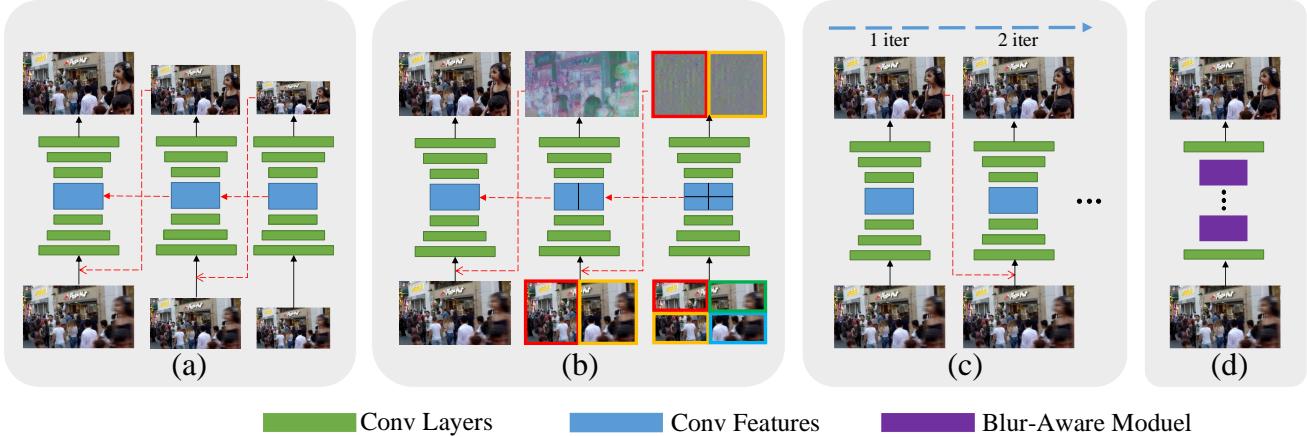


Figure 2. Network architecture comparisons among (a) MS, (b) MP, (c) MT, and (d) our BANet. Recurrent models are typically less efficient. BANet completes deblurring via a single forward pass.

[19] further exploit self-attention to address blur non-uniformity. Suin *et al.* [19] utilize MP-based processing with self-attention to extract features for areas with global and local motion. However, using the self-recurrent mechanism to generate multi-scale features often leads to a longer inference time. To shorten the latency, Purohit and Rajagopalan [17] selectively aggregate features through learnable attention which is enabled by deformable convolutions for modeling local blur 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 [17, 19]. Furthermore, motion blur coming from moving objects manifests smeared effects and produces directional and local averaging artifacts that cannot be handled by inter-pixel/channel correlations.

This paper proposes a *blur-aware attention networks* (BANet) to overcome the aforementioned issues. BANet is an efficient yet effective single-forward-pass model, as illustrated in Figure 2(d), which achieves the state-of-the-art deblurring performance and works in real time, as shown in Figure 1. Specifically, our model stacks multiple layers of the *blur-aware attention module* (BAM) for removing motion blur. The proposed BAM computes region-based attention by using fast and computationally inexpensive local averaging to globally and locally capture blurred patterns of different degrees, and leverages cascaded parallel dilated convolution to extract features without suffering from blur information loss as an MS model does. As a result, the proposed model possesses a better deblurring capability and can support subsequent real-time applications. In short, our contributions are two-fold: First, we design a novel BAM module which is capable of disentangling blur contents of different degrees in dynamic scenes. Second, our efficient single-forward-pass deep networks perform favor-

ably against the state-of-the-art methods while running 27x faster than the best competitor [19].

## 2. Related Work

**Conventional Methods.** Single-image deblurring is highly ill-posed. Conventional image deblurring studies often make different assumptions, *e.g.*, uniform, non-uniform, or depth-aware, to model blur characteristics. Namely, these methods [2, 7, 10, 13, 14] apply different constraints to the estimated blur kernels, latent images, or both with handcrafted regularization terms for blur removal. Unfortunately, these methods often lead to solving a non-convex optimization problem and involve heuristic parameter tuning that is entangled with the camera pipeline; thus, they cannot generalize well for complex real-world examples.

**Deblurring via Learning.** Learning-based approaches with self-recurrent modules gain great success in single-image deblurring; namely, the *coarse-to-fine* scheme can gradually restore the sharp image on different resolutions (MS) [3, 12, 20], fields of view (MP) [19, 24], or temporal characteristics (MT) [15]. Despite the success, self-recurrent models usually lead to longer runtime. Recently, non-recurrent based methods [8, 9, 17, 23] are proposed for efficient deblurring. For instance, Kupyn *et al.* [8, 9] suggest using conditional generative adversarial networks to restore blurred images. However, their methods do not well address non-uniform blur in dynamic scenes, often causing blur artifacts in the deblurring results. Yuan *et al.* [23] propose a spatially variant deconvolution network with optical flow estimation to guide deformable convolutions and capture moving objects during model training. However, optical flow may not always correlate with blur, which may cause ineffective deblurring.

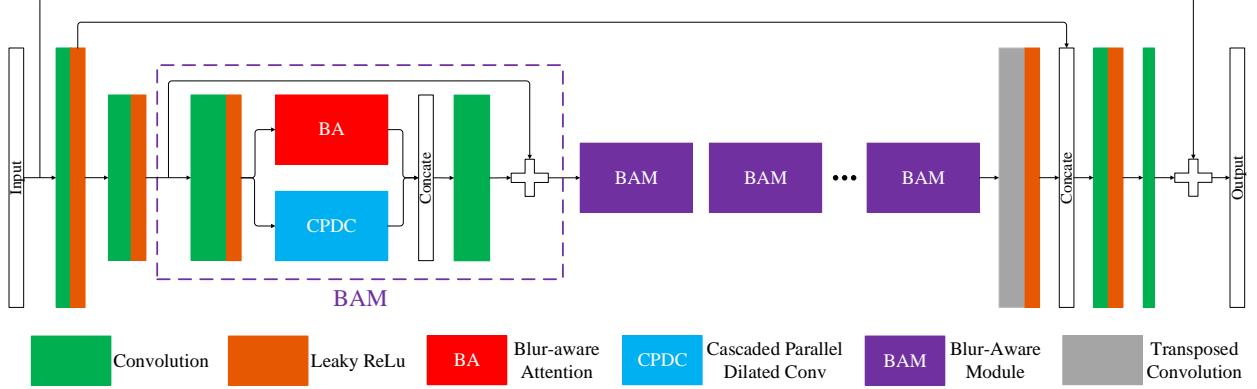


Figure 3. Network architecture of the proposed blur-aware attention networks (BANet). The blur-aware modules (BAM) serve as the building blocks of BANet. The first BAM is detailed in the purple dotted box while the rest are represented by solid purple boxes.

**Self-attention.** Self-attention (SA) [21] has been widely adopted to advance the fields of image processing [16, 25] and computer vision [5, 22]. Recent studies [17, 19] show that attention is beneficial for learning inter-pixel correlations to emphasize different local features for removing non-uniform blur. Purohit *et al.* [17] propose to deblur using SA to explore pixel-wise correlation for a non-local feature adaptation. However, since SA requires much memory in  $\mathcal{O}((HW)^2)$  space, where  $H$  and  $W$  are the height and width of the input to SA. Thus, their method only applies SA to the smallest-scale features (from a  $1280 \times 720$  blurred input to  $160 \times 90$  SA’s input), limiting the efficacy of SA. Also, motion blur causes directional and local averaging artifacts, which may not be well addressed by merely pixel-wise SA. Suin *et al.* [19] employ an MP architecture with less memory-intensive SA by using global average pooling, resulting in the space complexity  $\mathcal{O}(d_a d_c HW)$  where  $d_a$  is the channel dimension of the components *key* and *value* in SA;  $d_c$  is the dimension of the component *query* and  $d_a d_c \ll HW$ . Although their method produces high-quality deblurring results, it suffers from longer inference time. In contrast, we propose an efficient and low-memory-complexity regional averaging SA to capture blur information more accurately, which can be applied to input images of full resolution and gets better performance in real time.

### 3. Proposed Approach

We present the blur-aware attention networks (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 the stacked blur-aware

modules (BAM), which serve as the building blocks, to effectively disentangle different degrees of blurriness.

As illustrated in Figure 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: blur-aware attention (BA) and cascaded parallel dilated convolution (CPDC). The BAM is also a residual-like architecture that concatenates two output to capture both global and local blurring features in a learnable manner. We detail the two key components, BA and CPDC, in the following.

#### 3.1. Blur-aware Attention (BA)

To accurately restore the motion area displaying directional and averaging artifacts caused by moving objects and camera shakes, we propose a region-based self-attention module, called blur-aware attention (BA), to capture such effects in global (image) and local (patch) scales. As shown in Figure 4, BA contains two parts including multi-kernel strip pooling (MKSP) and attention refinement (AR). We describe their details as follows.

**Multi-Kernel Strip Pooling (MKSP).** Strip pooling (SP) has been broadly used to capture region-based information. Hou *et al.* [4] present an SP method that uses horizontal and vertical one-pixel long kernels to extract long-range 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 map be a three-dimensional (3D) tensor  $\mathbf{x} \in R^{H \times W \times C}$ , where  $C$  denotes the number of channels. Applying SP to  $\mathbf{x}$  generates a vertical and a horizontal tensor followed by a 1D convolution layer with the kernel size of 3. This produces  $\mathbf{y}^v \in R^{H \times C}$

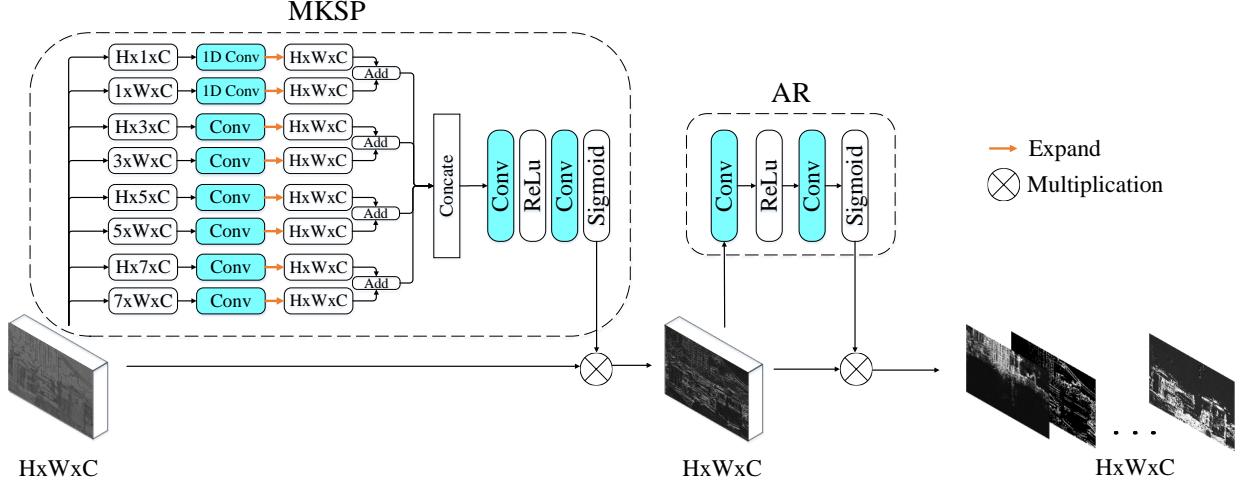


Figure 4. Architecture of blur-aware attention (BA). It cascades two parts, including multi-kernel strip pooling (MKSP) and attention refinement (AR). It is developed to disentangle blurred contents in an efficient way. See text for details.

and  $\mathbf{y}^h \in R^{W \times C}$ , where  $y_{i,c}^v = \frac{1}{W} \sum_{j=0}^{W-1} x_{i,j,c}$  and  $y_{j,c}^h = \frac{1}{H} \sum_{i=0}^{H-1} x_{i,j,c}$ . The SP operation fuses the two tensors into a 2D tensor  $\mathbf{y} \in R^{H \times W \times C}$ , where  $y_{i,j,c} = y_{i,c}^v + y_{j,c}^h$ , and then turns the fused tensor into an attention mask  $\mathbf{M}_{sp}$  as

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

where  $f_1$  is a  $1 \times 1$  convolution and  $\sigma_{sig}$  is the sigmoid function.

Motivated by SP, we propose MKSP that adopts strip pooling with different kernel sizes to filter out irrelevant parts from attended features more precisely. MKSP constructs horizontal and vertical  $n$ -pixel long kernels to average input features within rows or columns, where  $n \in \{1, 3, 5, 7\}$ . Thus, MKSP generates four pairs of tensors, each of which has a vertical and a horizontal tensor followed by a 1D (for the special case  $n = 1$ ) or 2D (for the rest cases) convolution layer with the kernel size of 3 or  $3 \times 3$ , respectively. This produces  $\mathbf{y}^{v,n} \in R^{H \times n \times C}$  and  $\mathbf{y}^{h,n} \in R^{n \times W \times C}$ , where the vertical tensor is

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

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

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

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

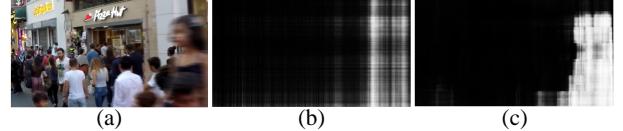


Figure 5. Visualization of the attention masks in our deblurring model. (a) An input blurred image. (b) & (c) The attention masks obtained using (b) strip pooling (SP) and (c) our MKSP.

MKSP then fuses each pair of tensors  $(\mathbf{y}^{v,n}, \mathbf{y}^{h,n})$  into a 2D tensor  $\mathbf{y}^n \in R^{H \times W \times C}$  by

$$y_{i,j,c}^n = y_{i,\lfloor \frac{n \times i}{W} \rfloor,c}^{v,n} + y_{\lfloor \frac{n \times i}{H} \rfloor,j,c}^{h,n}. \quad (4)$$

Similar to SP, we concatenate all the fused tensors to yield an attention mask as  $\mathbf{M}_{mksp} = f_{AR}(\mathbf{y}^1 \oplus \mathbf{y}^3 \oplus \mathbf{y}^5 \oplus \mathbf{y}^7)$ , where  $\oplus$  stands for concatenation operation, and  $f_{AR}(\cdot)$  represents two  $3 \times 3$  convolutions with a ReLU function in between followed by a sigmoid function.

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. Figure 5 shows an example of mask comparison between SP and the proposed MKSP.

**Attention Refinement (AR).** After obtaining the attended features by element-wise multiplication of the attention masks  $\mathbf{M}_{mksp}$  and the input tensor  $\mathbf{x}$ , we further refine these features 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

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

where  $\otimes$  represents element-wise multiplication, and  $\tilde{\mathbf{x}} = \mathbf{M}_{mksp} \otimes \mathbf{x}$ . Figure 6 demonstrates that cascading MKSP

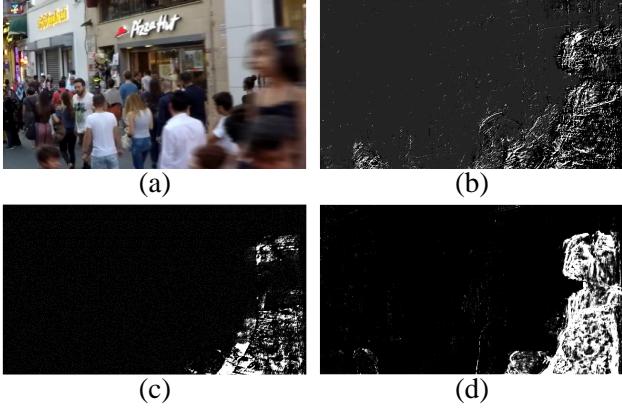


Figure 6. (a) Input blurred image. (b)-(d) Comparisons among the attended feature maps by using different components of the proposed BA including (b) AR, (c) MKSP, and (d) MKSP + AR.

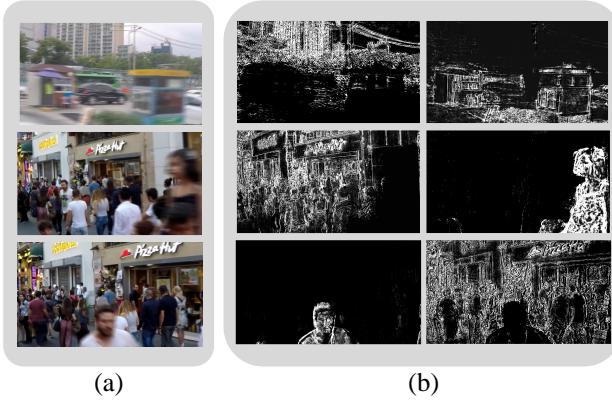


Figure 7. Three disentanglement examples of blurred patterns of different degrees using our BA. (a) Input blurred images and (b) attended feature maps on different regions.

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 channel dimensions, than those adopted in [17, 19]. It disentangles blurred contents with different degrees. Figure 7 showcases three examples of blur content disentanglement by using the proposed BA, where we witness that background scenes are differentiated from the foreground scenes because objects closer to the camera move faster, thus more blurred. Figure 9 shows more examples of attention maps yielded by BA, which implicitly acts as a gate for propagating relevant blur contents.

### 3.2. Cascaded Parallel Dilated Convolution (CPDC)

Atrous convolution, also called *dilated convolution*, has been widely applied to computer-vision tasks [1, 11] 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 con-*

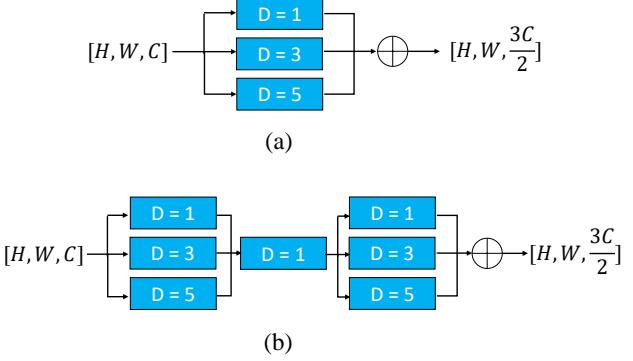


Figure 8. Architectures of (a) parallel dilated convolution (PDC) and (b) cascaded parallel dilated convolution (CPDC).

*volution* (CPDC) block with multiple dilation rates to capture multi-scale blurred objects. Instead of stacking dilated convolution layers with different rates in parallel, which we call *parallel dilated convolution* (PDC), our CPDC block cascades two sets of PDC with a convolutional layer. As an example, Figure 8(a) shows a PDC block that has three  $3 \times 3$  dilated convolutional layers, each of which has a dilation rate  $D$ , where  $D = 1, 3$ , and  $5$ , and each of which outputs features with half the number of input channels. After concatenation, the number of the output channel of the PDC block increases by 1.5 times. As shown in Figure 8(b), our CPDC block consists of two PDC blocks bridged by a  $3 \times 3$  convolutional layer, which would be more effective in aggregating multi-scale content information for deblurring.

## 4. Experiments

### 4.1. Experimental Setup

We evaluate the blur-aware attention networks (BANet) on two image deblurring benchmark datasets: 1) GoPro dataset [12] consists of 3214 pairs of blurred and sharp images of resolution  $720 \times 1280$ , where 2103 pairs are used for training and the rest are for testing, and 2) HIDE dataset [18] that contains 2025 pairs of HD images, all for testing. We train our model for 3,000 epochs using Adam optimizer with parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We set the initial learning rate to  $10^{-4}$ . After the first 50 epochs, it starts to linearly decay until  $10^{-7}$  after another 150 epochs. Following [15, 23], we randomly crop the input into  $256 \times 256$  patches, along with random flipping or rotation for data augmentation. Lastly, we implement our model with PyTorch library on the Intel Xeon Silver 4210 CPU and NVIDIA 2080ti GPU.

### 4.2. Experimental Results

**Quantitative Analysis:** We compare our method with eleven recently published works [12, 20, 3, 9, 24, 6, 23, 24, 15, 17, 19] that also handle dynamic deblurring on the



Figure 9. Visualization of the blur-aware attention masks where moving objects in the blurred images are highlighted while background is mostly excluded. These blur-aware masks are crucial for handling blurry images with diverse blur patterns.

Table 1. Evaluation results on the benchmark GoPro test dataset. A value in bold is the best score in its column and underlined the second best. Symbol \* represents that the code was not released; thus we cite the results from the original papers or evaluate on the released deblurring images.

Method	PSNR↑	SSIM↑	Time <sub>(ms)</sub>
MSCNN [12]	30.40	0.936	943
SRN [20]	30.25	0.934	650
DSD [3]	30.96	0.942	1300
DeblurGanv2 [9]	29.55	0.934	42
DMPHN [24]	31.36	0.947	354
LEBMD* [6]	31.79	0.949	—
EDSD* [23]	29.81	0.934	<b>10</b>
DBGAN+* [26]	31.10	0.942	—
MTRNN [15]	31.13	0.944	53
RADN* [17]	31.85	0.953	38
SAPHN* [19]	32.02	0.953	770
BANet (stack-8)	<u>32.23</u>	<u>0.955</u>	23
BANet (stack-10)	<b>32.44</b>	<b>0.957</b>	28

GoPro [12] test dataset. For HIDE [18] dataset, we chose seven recent deblurring [9, 20, 18, 3, 24, 15, 19] methods depending on compared methods’ availability. Table 1 lists the metric scores *i.e.* PSNR, SSIM and running time obtained on the GoPro dataset from different approaches, including two variants of our model: one stacking eight (stack-8) and the other stacking ten (stack-10) BAMs, respectively. We can observe that the self-recurrent types of models [3, 12, 15, 20, 19] take comparatively longer times than non-recurrent ones [9, 17], including ours. We recorded the average runtime of all the models using a single GPU. As reported in Table 1, our BANet runs significantly faster than the others while achieving the best metric scores. In particular, BANet outperforms the nearest competitor [19] by 0.42 in terms of PSNR while running about 27x faster. Similarly, in Table 2, the BANet works significantly better than all the compared methods.

Table 2. Evaluation results on the benchmark HIDE dataset. A value in bold is the best score in its column and underlined the second best. \* represents that the code was not released; thus we cite the results from the original papers or evaluate on the released deblurring images.

Method	PSNR↑	SSIM↑	Time <sub>(ms)</sub>
DeblurGanv2 [9]	27.40	0.882	42
SRN [20]	28.36	0.904	424
HAdeblur* [18]	28.87	0.930	—
DSD [3]	29.10	0.913	1.200
DMPHN [24]	29.10	0.918	341
MTRNN [15]	29.15	0.918	53
SAPHN* [19]	29.98	0.930	—
BANet (stack-8)	<u>30.16</u>	<u>0.930</u>	<u>22</u>
BANet (stack-10)	<b>30.27</b>	<b>0.931</b>	<b>26</b>

**Qualitative Analysis:** Figure 10 shows the qualitative comparisons on the GoPro test dataset with previous state-of-the-art [24, 15, 17], and Figure 11 on the HIDE dataset with [15, 24]. Thanks to the designed blur-aware module, our model can effectively restore results with sharper edges and richer details in dynamic scenes containing camera shake and object motion. For instance, MTRNN [15] and RADN [17] can not well recover the text region in the first example and the brick pattern in the second example in Figure 10. In Figure 11, both MTRNN [15] and DMPHN [24] do not work well on deblurring the text in the first example and the faces in the second and third examples. In contrast, BANet can recover those regions well with an even shorter runtime.

### 4.3. Ablation Study

**BAM with Different Components:** In Table 3, we explore the performance change with different component unions in our Blur-Aware Module (BAM) based on the GoPro test dataset. Adding a simple attention mechanism (AR) to both PDC (Net1 vs. Net2) or PDC+MKSP (Net3 vs. Net4)

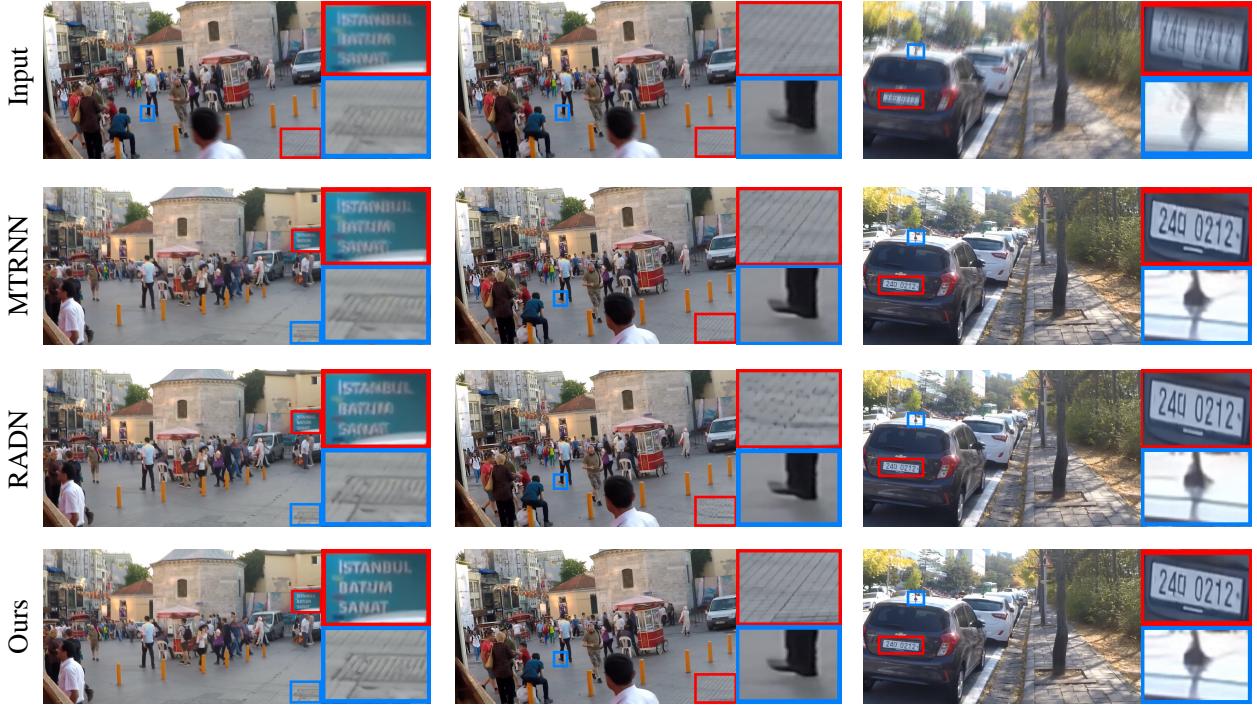


Figure 10. Qualitative comparisons of deblurring results on testing images from the GoPro [12] dataset. The first row shows the input blurred images and the following rows show results from MTRNN [15], RADN [17], and ours.



Figure 11. Qualitative comparisons of deblurring results on testing images from the HIDE [18] dataset. The first row contains the input blurred images. The second row contains MTRNN [15]'s deblurred results, the third DMPHN [24]'s and the last our BANet's.

Table 3. Ablation study based on the GoPro test dataset for using different component combinations in the BAM of our BANet (stack-8).

Method	PDC	AR	MKSP	CPDC	PSNR	Time (ms)
Net1	✓				31.29	7
Net2	✓	✓			31.47	9
Net3	✓		✓		31.83	13
Net4	✓	✓	✓		32.02	16
Net5	✓	✓	✓	✓	32.23	23

Table 4. Performance comparisons of BANet stacking different numbers of BAMs based on the GoPro test dataset.

Method	stack-4	stack-8	stack-10	stack-12
PSNR	31.44	32.23	32.44	32.46
Time (ms)	12	23	28	35

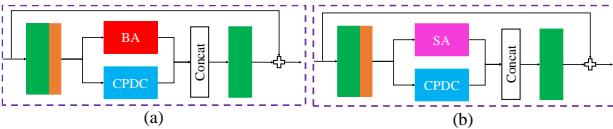


Figure 12. Architecture comparisons between (a) our original BAM and (b) BA replaced by SA [25] in BAM.

for deblurring increases PSNR by around 0.2. Substituting PDC in Net4 with CPDC (Net5), our proposed version of the BAM, leads to a significant performance gain. Our proposed Net5 consists of using global attention and local convolution for locating blur regions; thus, this combination can generate the best performance while achieving real-time practical significance.

**Numbers of Stacked BAMs:** Using more layers to enlarges the receptive field may improve performance for computer vision or image processing tasks. However, for deblurring, stacking more layers does not guarantee better performance [19], and might additionally cost 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 4, we show performance comparisons with varying numbers of BAMs stacked in our model based on the GoPro test dataset. We list four versions: stack-4, stack-8, stack-10, and stack-12, corresponding four, eight, ten, and twelve BAMs used in the BANet. Although the quantitative results improved with the increase in the number of BAMs, the improvement became marginal after 12. Therefore, we choose 10 res-blocks for the excellent balance between efficiency and performance.

#### 4.4. Blur-aware Attention vs. Self-Attention:

Purohit *et al.* [17] utilized a similar self-attention (SA) mechanism proposed in [25] for deblurring. It helps connect

Table 5. A performance comparison between BA with SA [25] using BANet (stack-4) as a backbone based on the GoPro test dataset. “\*” represents deblurring eight sub-images instead of an entire image.

SA*	BA*	BA	PSNR	Time (ms)
✓			30.59	1240
	✓		31.30	670
		✓	31.44	12

regions with similar blur to facilitate global access to relevant features across the entire input feature maps. However, its high memory usage makes applying it to images of high resolution infeasible. Thus, SA can only be employed in network layers on a smaller scale like [17], where important blur information would be lost due to down-sampling. In contrast, our proposed region-based attention is more suitable to correlate regions with similar blur characteristics. Moreover, it can be applied to images with a larger resolution thanks to its low memory consumption.

To further demonstrate our BA’s efficacy, we compare the SA [25] with BA using our BANet (stack-4) as a backbone network, shown in Figure 12 (b). Due to high memory demand for SA ( $\mathcal{O}((HW)^2)$ ) to process  $720 \times 1280$  images, we adopted our stack-4 model for training. When testing the networks, we separated the input image into eight sub-images for SA to process each with a single 2080ti GPU. We provided deblurring results using BA with or without splitting the input image into eight sub-images for a fair comparison. Table 5 shows the scenario of the input divided into eight sub-images; BA still works better than SA for deblurring based on our BANet. Also, dividing the input image for deblurring using a single GPU increases the runtime, and adopting BA runs significantly faster. Last but not least, since our BA has lower memory usage, we can process the original input image without division to attain better quality (PSNR).

## 5. Conclusions

This paper proposes novel blur-aware attention networks (BANet), which consists of the stacked blur-aware modules (BAMs) to disentangle blur contents of different degrees and the cascaded parallel dilated convolution (CPDC) module to aggregate multi-scale content features, for more accurate and efficient dynamic scene deblurring. We investigate and examine our design through demonstrations of attention masks and attended feature maps as well as extensive ablation studies and performance comparisons. It turns out that the proposed BANet achieves real-time deblurring and performs favorably against the state-of-the-art deblurring methods on the GoPro and HIDE benchmarks.

## References

- [1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5
- [2] S. Cho and S. Lee. Fast motion deblurring. 2009. 2
- [3] H. Gao, X. Tao, X. Shen, and J. Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2019. 1, 2, 5, 6
- [4] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng. Strip Pooling: Rethinking spatial pooling for scene parsing. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2020. 3
- [5] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2018. 3
- [6] Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu. Learning event-based motion deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2020. 5, 6
- [7] T. H. Kim and K. M. Lee. Segmentation-free dynamic scene deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2014. 2
- [8] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. 2018. 2
- [9] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proc. Int'l Conf. Computer Vision*, 2019. 2, 5, 6
- [10] X. L and J. J. Two-phase kernel estimation for robust motion deblurring. In *Proc. Euro. Conf. Computer Vision*, 2010. 2
- [11] S. Liu, D. Huang, and a. Wang. Receptive field block net for accurate and fast object detection. In *Proc. Euro. Conf. Computer Vision*, 2018. 5
- [12] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2017. 1, 2, 5, 6, 7
- [13] J. Pan, Z. Hu, Z. Su, H. Lee, and M. Yang. Soft-segmentation guided object motion deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2016. 2
- [14] J. Pan, Z. Hu, Z. Su, and M. Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2014. 2
- [15] 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. Euro. Conf. Computer Vision*, 2020. 1, 2, 5, 6, 7
- [16] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. *arXiv preprint arXiv:1802.05751*, 2018. 3
- [17] K. Purohit and A. N. Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proc. Nat'l Conf. Artificial Intelligence*, 2020. 2, 3, 5, 6, 7, 8
- [18] Z. Shen, W. Wang, J. Shen, H. Ling, T. Xu, and L. Shao. Human-aware motion deblurring. In *Proc. Int'l Conf. Computer Vision*, 2019. 5, 6, 7
- [19] M. Suin\*, K. Purohit\*, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2020. 1, 2, 3, 5, 6, 8
- [20] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. Neural Information Processing Systems*, 2017. 3
- [22] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2018. 3
- [23] Y. Yuan, W. Su, and D. Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2020. 2, 5, 6
- [24] H. Zhang, Y. Dai, H. Li, and P. Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2019. 1, 2, 5, 6, 7
- [25] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. 2019. 3, 8
- [26] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li. Deblurring by realistic blurring. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2020. 6