Perceptual Variousness Motion Deblurring with Light Global Context Refinement

Jichun Li Weimin Tan Bo Yan*
School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, China
lijc19@fudan.edu.cn wmtan14@fudan.edu.cn byan@fudan.edu.cn*

Abstract

Deep learning algorithms have made significant progress in dynamic scene deblurring. However, several challenges are still unsettled: 1) The degree and scale of blur in different regions of a blurred image can have a considerable variation in a large range. However, the traditional input pyramid or downscaling-upscaling, is designed to have limited and inflexible perceptual variousness to cope with large blur scale variation. 2) The nonlocal block is proved to be effective in the image enhancement tasks, but it requires high computation and memory cost. In this paper, we are the first to propose a light-weight globally-analyzing module into the image deblurring field, named Light Global Context Refinement (LGCR) module. With exponentially lower cost, it achieves even better performance than the nonlocal unit. Moreover, we propose the Perceptual Variousness Block (PVB) and PVB-piling strategy. By placing PVB repeatedly, the whole method possesses abundant reception field spectrum to be aware of the blur with various degrees and scales. Comprehensive experimental results from the different benchmarks and assessment metrics show that our method achieves excellent performance to set a new state-of-the-art in motion deblurring. [1]

1. Introduction

The restoration of the latent sharp image from the blur input in dynamic scene has long been an important task in computer vision and image processing. Deep learning methods for single image deblurring, particularly convolutional neural networks (CNNs), have obtained remarkable success [9, 24, 1, 5, 31, 6, 21]. Nah *et al.* propose the method [20] recovering the blurred image with the input pyramid on 3 scales, in a coarse-to-fine manner. Tao *et al.* deeply investigate the coarse-to-fine strategy and propose a Scale-Recurrent Network [25]. With the adoption of the convLSTM [30], multi-scale, and weight-sharing, SRN achieves high PSNR with fewer parameters. Recently, the state-of-the-art methods [16, 4, 34, 17, 22, 33, 35] have further revealed the potential of CNNs in deblurring task.

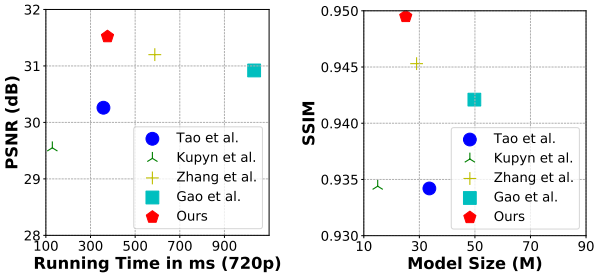

Figure 1: The comparison of different methods, in terms of the accuracy and cost. Our approach is better than the other state-of-the-art methods. "720p" indicates that the test image is in the size of 1280× 720.

One of the biggest challenges of deblurring comes from the fact that the blur pattern's degree and scale vary widely. Traditionally, multi-scale input pyramid and downsampling-upsampling layers inside the network are common strategies to release the difficulty brought by the complicated blur pattern [20, 16]. More recent methods focus on other handcrafted strategies to deal with the wide range of blur scale variation. [25] utilizes a recurrent network with weight sharing in a coarse-to-fine manner. [34] proposes a multi-patch methodology to exploit multi-scale information. [22] even puts forward a multi-temporal idea that deblurs the image from hard to easy progressively.

Unfortunately, their adopted multi-scale, multi-patch, and multi-temporal strategies augment their models' perceptual scale variousness with only limited times, e.g., there are only two scales or five temporal intervals considered in their designed methods. In other words, the final reception

[1]This work is supported by NSFC (Grant No.: U2001209, 61902076, 61772137) and Natural Science Foundation of Shanghai (21ZR1406600).
* Corresponding author: Bo Yan.

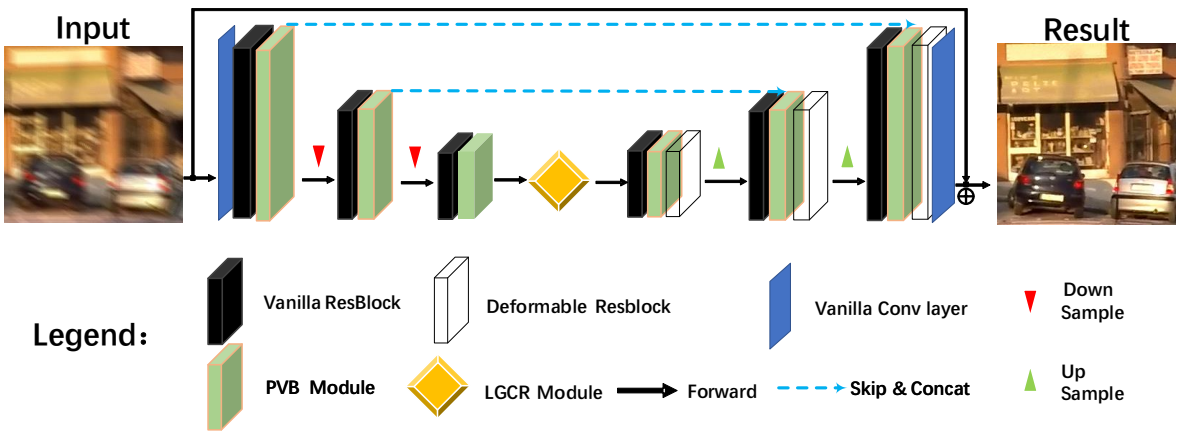

Figure 2: The architecture of the SimpleNet. Since it grasps the essence of deblurring with the help of the PVB and LGCR, it is designed in an auto-encoder fashion, which is easy to implement and follow. The vanilla ResBlock [8] has three convolution layers. The "Down Sample" is a conv layer, stride=2; the "Up Sample" is deconvolution layer, stride=2.

fields obtained in the information flow are only augmented by limited times. However, the blur's degree has a considerable variation in a relatively wide range. Thus, these discrete and handcrafted strategies are not satisfactory to equip CNN with enough ability to percept the complex blur patterns whose scale is widely distributed.

Moreover, non-localized neural operation with lower cost is in good demand for deblurring task. The CNN design is based on a localized filtering operation, which processes one local neighborhood at a time. It is unfavorable for the task that requires a broader reference range or even full-image self-reference, such as image segmentation, pose estimation, and severe motion blur recovery. The nonlocal proposed in [28] is an excellent, classic yet expensive solution for the caption of long-range dependencies. Recently, inspired by tensor canonical-polyadic decomposition theory, Chen *et al.* propose a tensor generation module and a tensor reconstruction module, named "TGM+TRM" (T+T) in semantic segmentation [2]. It computes the global information while tackling the high-rank difficulty. However, T+T's structure is good at high-level semantic reasoning, yet bad at detail recovery, which is essential in deblurring task. Moreover, its non-linearity of the 1-rank tensors is not sufficient; its global context is not well learned and utilized.

In this paper, we work on the deficiencies mentioned above, and propose our deblurring method, SimpleNet. We propose a new light-weight non-localized module, named Light Global Context Refinement (LGCR). It is the first time that such a light-weight non-localized module is proposed in deblurring task to enrich global detail instead of pixel-wise reasoning, with better performance and a much lower cost than the nonlocal module. Moreover, we propose the Perceptual Variousness Block (PVB) and PVB-piling

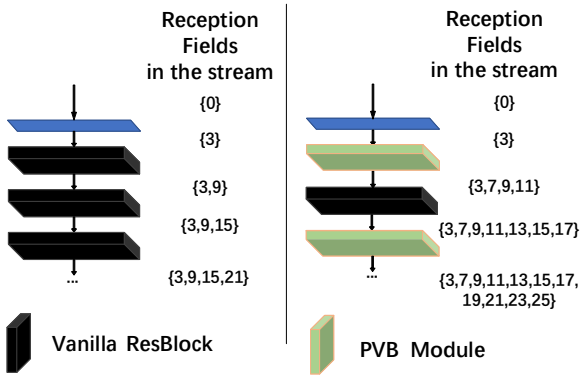

Figure 3: PVB greatly expands the variety of reception fields that a network can perceive. The architecture of PVB is illustrated in Figure 5. The "Vanilla ResBlock" is an ordinary three-layered. The braced statistics stands for reception field spectrum. With larger reception field spectrum, network has the better perceptual variousness and ability.

strategy. PVB provides abundant adaptive multi-scale reception ability with broad reception spectrum. Unlike the traditional "multi" methodology, PVB-piling strategy can greatly broaden the variousness of the network's reception scales and perceptual ability, as shown in Figure 3, facing the challenge of wide range blur variation. Finally, we examine our method with the state-of-the-art methods, in GoPro, RealBlur-J and RWBI benchmarks. Comprehensive experiments show that our method achieves the best performance to set a new state-of-the-art.

We name our network "SimpleNet" because its architecture is straightforward, easy to implement.

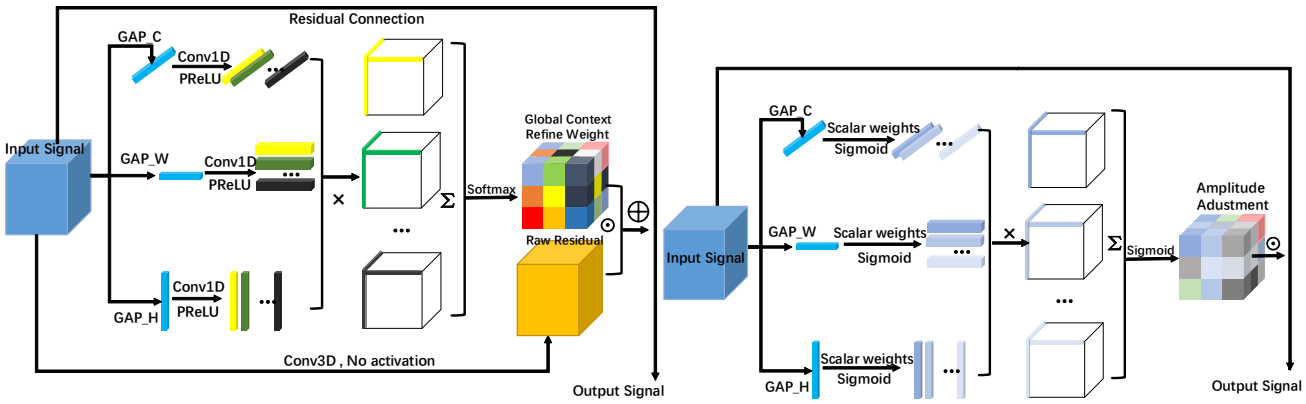

Figure 4: The detailed architecture of the LGCR module (col. #1), in comparison with TGM+TRM (T+T) module (col. #2). Although both modules adopt the high-rank-to-low-rank decomposition theory[15], their main aims and detailed design are totally different. Experiments demonstrate that LGCR is far more effective than T+T in deblurring task.

In conclusion, our contributions are as follows:

- We are the first to propose a novel light-weight non-localized module into the image deblurring, named Light Global Context Refinement (LGCR). It outperforms both nonlocal and T+T methods, in a detail enhancement manner, instead of pixel-wise reasoning.
- We proposed Perceptual Variousness Block (PVB) and PVB-piling strategy. PVB provides abundant adaptive multi-scale reception ability. PVB-piling strategy can greatly enrich the variousness of the network's reception spectrum and perceptual ability, challenging the wide range of blur variations.
- We put forward a robust and effective deblur network, named SimpleNet. It has a simple encoder-decoder structure, and it is easy to implement and follow.
- Comprehensive experiments are conducted, not only on the prevalent GoPro benchmark, but also on the newly proposed RealBlur-J, RWBI benchmarks, with comprehensive assessment metrics.

2. Related Work

2.1. Single image deblurring in dynamic scene

As we discussed in Section 1, the motion blur in the dynamic scene is introduced by the fast relative motion, between the scene and camera, during the short period of the shutter's exposure. It is not generated by a sharp static image with a specific blur kernel. Thus, the blur pattern in the dynamic scene is blind and non-uniform.

Recently, CNN-based algorithms have made remarkable success in the task. Gao *et al.* [4] put forward a more complex network with a new parameter selective sharing strategy and high order nested skip connections. Zhang *et al.* [34] propose multi-patch methodology to obtain fine-to-coarse hierarchical representation for deblurring. Kupyn *et al.* [17] propose a model composed of backbone-fpn generator with global and local discriminators. Methods [33, 22] start to exploit the latent temporal information to recover the blurred image, since the GoPro dataset can be regarded as video clips in some aspect.

When dealing with the blur pattern with high variation, these methods use handcrafted and discrete multi-scale, multi-patch, and multi-temporal mechanisms. These mechanisms are fixed and inflexible, and they only augment the reception variousness and perceptual ability of the models with limited times. Since the degree and scale of blur can have considerable variation, some effective module or strategy is required to equip CNN methods with abundant reception variousness or perceptual variousness.

2.2. Assessing metrics for dynamic scene deblurring

Dynamic scene deblurring task is a special case of image enhancement. Thus, theoretically, every assessment metric in image enhancement can be adopted in the motion deblurring field. Peak Signal Noise Ratio (PSNR) and Structural Similarity (SSIM) [29] are the most commonly used ones. The latter has better consistency with HVS than the former. Recently, the Learned Perceptual Image Patch Similarity (LPIPS) [37] has been adopted as a full-reference metric in many works [26, 13, 12, 10, 18]. It calculates the pixel-wise perceptual similarity between the input images. It is trained by the proposed BAPPS Dataset. Experiments prove that it performs much better than the traditional full-reference similarity metrics.

For a blurred image obtained in the wild, there is no pixel-aligned sharp ground truth on hand. Only No-reference metrics can be used. Li *et al.* [19] propose a CNN-based non-reference deblurring quality assessment

method. To our knowledge, it is the first and the only CNN-based IQA specially designed for deblurring tasks. Without a latent sharp image, it provides a quality assessment score accordant with the human vision system (HVS).

In our work, we will comprehensively evaluate the performance of our method using the metrics mentioned above.

3. Our method

The network architecture of the SimpleNet is demonstrated in Fig. 2. LGCR module enriches the feature's detail based on long-range dependencies, which is helpful to deblurring, with better performance yet lower cost. The proposed PVBs and PVB-piling strategy equip the network with perceptual variousness to conquer blur variation. With the above, the Deformable ResBlock and skip connections, SimpleNet achieves the best performance with a simple auto-encoder structure.

3.1. Light Global Context Refinement (LGCR)

For dynamic scene deblurring, long-range dependencies are crucial when the blur pattern is severe, or in a large scale. The nonlocal module by [28] is a possible solution, but it requires considerably large computation and memory expenses. Inspired by the works of [2, 15], we are the first to propose a light-weight long-range dependencies enriching module in the deblurring field, named Light Global Context Refinement module (LGCR), as shown in Figure 4. Please note that LGCR is designed for detail enrichment, while T+T is designed for semantic pixel reasoning.

According to the tensor decomposition theory, a tensor can be represented as the linear combination of its low-rank principal components.

Formally, given the input tensor $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$ and the CP tensor reconstruction rank r, the axes-based pooled vectors $\boldsymbol{v}_c \in \mathbb{R}^{C\times 1\times 1}$,$\boldsymbol{v}_h \in \mathbb{R}^{1\times H\times 1}$,$\boldsymbol{v}_w \in \mathbb{R}^{1\times 1\times W}$ are obtained by global average pooling (GAP) of $\boldsymbol{I}$, along the channel-axis, height-axis and width-axis. Then, the context fragments are generated by a Conv-PReLU sequence:

$$\begin{aligned}
\boldsymbol{v}_{ci} &= PReLU(Conv1D(\boldsymbol{v}_c, \boldsymbol{W}c_i)), \boldsymbol{v}_{ci} \in \mathbb{R}^{C\times 1\times 1},\\
\boldsymbol{v}_{hi} &= PReLU(Conv1D(\boldsymbol{v}_h, \boldsymbol{W}h_i)), \boldsymbol{v}_{hi} \in \mathbb{R}^{1\times H\times 1},\\
\boldsymbol{v}_{wi} &= PReLU(Conv1D(\boldsymbol{v}_w, \boldsymbol{W}w_i)), \boldsymbol{v}_{wi} \in \mathbb{R}^{1\times 1\times W}.
\end{aligned} \tag{1}$$

where i indicates the rank-1 tensor index, $0 \le i \le r$; $Conv1D$ indicates the 1D convolution operator; $\boldsymbol{W}m_i, m \in \{c, h, w\}$ indicates the learned weights with respect to each axis, with the kernel size of 1×3; $PReLU$ indicates the activation function proposed in [7]. Then CP rank-r reconstructed "Global Context Refine Weight" ($\boldsymbol{GCRW} \in \mathbb{R}^{C\times H\times W}$) is calculated by:

$$\boldsymbol{GCRW_{raw}} = \sum_{i=1}^{r} \boldsymbol{v}_{ci} \otimes \boldsymbol{v}_{hi} \otimes \boldsymbol{v}_{wi}, \tag{2}$$

$$\begin{aligned}
\boldsymbol{GCRW} &= softmax(\boldsymbol{GCRW_{raw}})\\
&= \frac{exp^{\boldsymbol{GCRW_{raw}}}}{\sum_{c=1}^{C}(\sum_{h=1}^{H}(\sum_{w=1}^{W}(exp(\boldsymbol{GCRW_{raw}}(c,h,w)))))},
\end{aligned} \tag{3}$$

Next, the "Raw Residual"($\boldsymbol{Rr}$) tensor is simply computed with 3D convolution without activation:

$$\boldsymbol{Rr} = Conv3D(\boldsymbol{I}, \boldsymbol{W}_G), \boldsymbol{Rr} \in \mathbb{R}^{C\times H\times W}, \tag{4}$$

where $Conv3D$ indicates 3D convolution operator, the $\boldsymbol{W}_G$ is the kernel with size of $3\times 3\times 3$. Finally, the output of LGCR is:

$$\boldsymbol{Out} = \boldsymbol{Rr} \odot \boldsymbol{GCRW} + \boldsymbol{I}, \boldsymbol{Out} \in \mathbb{R}^{C\times H\times W}, \tag{5}$$

The detailed comparison between our method and TGM+TRM (T+T) module is shown in Figure 4. Their main aims and detailed design are totally different. 1) LGCR aims to enrich feature details, providing refined global information as a residual to "enrich" (add details to) the input. Meanwhile, T+T is designed to perform pixel semantic reasoning. It reasons out a global amplitude adjustment weight to multiply the input, tune-up the positive semantic pixels, and suppress negative ones. 2) For detailed design, firstly, with "Raw Residual", LGCR can compensate GAP's information loss, while T+T ignores it. Secondly, before activation functions, LGCR calculates context fragments using Conv1D with kernel size 1×3, while T+T uses simple multiplication with $3\times r$ scalars. The nonlinearity of the LGCR context fragments is thus better than that of the T+T context fragments, which is essential to the reconstructed high-rank tensor's representation ability. Experiments also show that our LGCR outperforms the nonlocal module, while T+T decreases the performance. Related results and discussions are presented in Section 4.4.

3.2. Perceptual Variousness Block (PVB)

One of the biggest challenges of deblurring is that blur pattern's degrees and scales vary widely. However, the traditional multi-scale mechanisms are designed fixed and inflexible. Thus, they augment the models' reception variousness and perceptual ability with only limited times, while the complex blur patterns' scales are widely distributed.

We put forward the PVB module and corresponding PVB-piling strategy, as shown in Fig. 2 and 5. PVB provides abundant adaptive multi-scale reception ability. PVB-piling strategy is to simply apply PVB in every scale of the

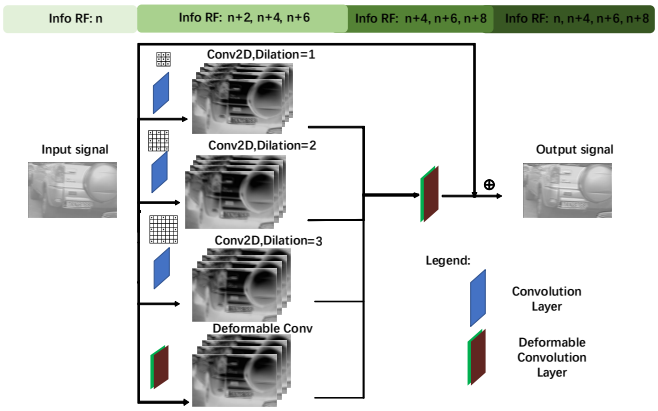

Figure 5: A real example that PVB module extracts the information from reception field of 3 scales with different dilation. "Info RF: n" indicates the information from the reception field sized n. The conv layers with different dilations ensure PVB's variousness of reception range. Deformable conv provides PVB with an adaptive reception range as a flexible reception supplement.

SimpleNet, which can significantly enrich the variousness of the network's reception scales and perceptual ability.

Formally, given the input tensor $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$, PVB extract a comprehensive feature from 3 conv layers with different reception scales and 1 deformable conv layer. The dilation rates of conv layers[32] are 1, 2, 3. Formulation is:

$$\begin{aligned}\boldsymbol{Feat_c} = Concat(ReLU(Conv2D(\boldsymbol{I}, \boldsymbol{W}d_1))+\\ ReLU(Conv2D(\boldsymbol{I}, \boldsymbol{W}d_2)) + ReLU(Conv2D(\boldsymbol{I}, \boldsymbol{W}d_3))\\ + DeformConv(\boldsymbol{I}, \boldsymbol{W}df_1)),\end{aligned} \tag{6}$$

where $Concat$ is the feature concatenation along channel axis, $\boldsymbol{W}d_i, i \in \{1, 2, 3\}$ represent the weight that has 3×3 non-zero parameters with the dilation rate of i. Then the comprehensive feature is fused by another deformable convolution layer, obtaining the fused residual feature:

$$\boldsymbol{Feat_{fused}} = DeformConv(\boldsymbol{Feat_c}, \boldsymbol{W}f), \tag{7}$$

Finally, the output is:

$$\boldsymbol{Out} = \boldsymbol{I} + \boldsymbol{Feat_{fused}}, \tag{8}$$

3 conv layers with different fixed-sized reception fields ensure PVB's variousness of reception range. The deformable conv provides PVB with an adaptive (learnable) reception range as a flexible reception supplement. Thus, the perceptual range of PVB, from small to large, from fixed to flexible, is reasonably sufficient. Therefore, with its perceptual variousness, PVBs can perceive and adapt for various blur patterns with large distribution scales. Piling PVB several times can significantly broaden the diversity of the network's reception scales and perceptual ability, which is beneficial to deblurring.

3.3. Deformable ResBlock (DR)

As we discussed above, the degree of the blur pattern is in a considerable variation. In many cases, the useful pixels to recover a certain pattern are located irregularly in a somewhat distorted spatial distribution. Fortunately, the deformable convolution [3, 27, 38] has a flexible spatial sampling point of the filter, which the network can learn itself.

We propose Deformable ResBlock (DR). Given the input $\boldsymbol{I} \in \mathbb{R}^{C\times H\times W}$, the calculation of DR is formulated as

$$Conv2D(ReLU(DeformConv(\boldsymbol{I}, \boldsymbol{W}df_2)), \boldsymbol{W}_l) + \boldsymbol{I}, \tag{9}$$

It captures the shape of irregularly distributed blur patterns, as well as enriches the perception scales that our network aware of. Therefore, we place one DR after each PVB in the decoder of SimpleNet.

3.4. SimpleNet

The architecture of our SimpleNet is based on a simple auto-encoder. As shown in Fig.2, it is composed of six Resblocks, six PVBs, three DRs, and one LGCR. These blocks are all based on residual methodology, and they are carefully designed and deployed in the SimpleNet. It is easy to implement and follow, without bells and whistles.

3.5. Optimization and implementation

The loss function we choose is L1 loss, and ADAM optimizer [14] is adopted to train SimpleNet, with β_1=0.9,β_2=0.999. The batch size is 8. Learning rate is 1e-4, exponentially decayed every 630k iterations, with the decay rate $\frac{\sqrt{10}}{10}$, for totally 2,200k iterations. The CP tensor decomposition rank r in our LGCR is 64, following [2].

Our SimpleNet is implemented in Pytorch [23], on Ubuntu 16.04 desktop. The training set is the proposed training set in GoPro benchmark.

4. Experiments

4.1. Platform and benchmark

All our experiments are conducted on the Ubuntu 16.04 desktop PC with Intel i7-7700k, 32GB RAM, GTX-1080Ti. All the PSNR and SSIM results are obtained by running the built-in functions in MATLAB R2019b.

The benchmarks we adopt are:

GoPro The most prevalent dataset, consists of 3214 pairs, 2103 for training, 1111 for testing. Ground truth images are obtained by a GoPro high-speed camera with a frame rate of 240, while the blur input images are gained by the average of the neighboring 7 to 13 frames.

Table 1: Performance comparisons with existing algorithms on GoPro[20]. SimpleNet achieves the best performance.

Methods	Tao *et al.* [25]	Kupyn *et al.* [17]	Zhang *et al.* [34]	Gao *et al.* [4]	Yuan *et al.* [33]	Park *et al.* [22]	Zhang *et al.* [35]	Our SimpleNet
PSNR (dB)	30.26	29.55	31.20	30.92	29.81	31.15	30.43	**31.52**
SSIM	0.9342	0.9344	0.9453	0.9421	0.9368	0.9454	0.9372	**0.9495**

Table 2: Comprehensive analysis of all the current competitive deblurring algorithms. **Red** refers to the best performance of its item, while blue is the second. The competition is fierce. The introduction of LPIPS and Deblur-IQA is in Section 2.2.

	GoPro [20]			RealBlur-J [11]			RWBI [36]	Model Size	Time (ms)
	PSNR↑	SSIM↑	LPIPS [37]↓	PSNR↑	SSIM↑	LPIPS [37]↓	Deblur-IQA[19]↑		
Tao *et al.*[25]	30.26	0.9342	0.12706	26.58	0.8630	0.16042	-8.4581	33.6M	358
Kupyn *et al.*[17]	29.55	0.9344	0.11728	26.68	0.8622	0.14295	**-7.7350**	**15.0M**	**129**
Zhang *et al.*[34]	31.20	0.9453	0.12800	25.84	0.8459	0.17838	-9.0433	29.0M	588
Gao *et al.*[4]	30.92	0.9421	0.12220	26.35	0.8552	0.19132	-8.1785	49.8M	1033
Ours (SimpleNet)	**31.52**	**0.9495**	**0.10788**	**26.95**	**0.8641**	**0.14126**	-7.9188	25.1M	376

RealBlur-J [11] A newly proposed benchmark with 3758 training pairs and 980 testing pairs. The image pairs are obtained by a beam splitter and two cameras with different exposure and the post-processing procedure.

RWBI [36] A brand new benchmark named "Real-World Blur Image dataset". There are 3112 blur images that are taken in the real world with several types of devices, without sharp ground truth.

4.2. Quantitative Evaluation on the benchmarks

Firstly, we evaluate SimpleNet in GoPro benchmark, with all the current state-of-the-art methods. From Tab. 1, we find that our method achieves the best performance.

To further comprehensively reveal our algorithm's strengths and weaknesses, we selected the most competitive, representative, and available deblurring algorithms at present to conduct the experiments in Tab. 2 and Figure 6. All the methods involved are only trained by GoPro's training set by the relative authors. It is solid and persuasive to announce that SimpleNet is the winner in the three metrics. In Fig.6, SimpleNet has the best performance, in challenging cases such as big blur scale by a close object (row #1), severe blur by fast motion (row #2), or structured patterns (the rest rows). It indicates that LCGR learns the long-range dependencies well, while PVB-piling brings perceptual variousness to cope with a wide range of blur patterns.

In RealBlur-J, we bring those methods directly to run only the test set, to evaluate their deblurring accuracy and transferability. We can observe that these methods' performances are close and low because of the domain gap introduced by the different ways to generate data samples. Even so, our method still excels in others.

In RWBI, a real-world dataset, SimpleNet also achieves great results. Since there is no ground truth for RWBI, originally, it is hard to assess the algorithms except for the visual results in Figure 6. Thanks to [19], we use deblur-IQA model to test the no-reference quality score, as shown in the eighth column in Tab. 2. The discriminator-trained method [17] outperforms SimpleNet, because GAN mechanism can greatly improve the perceptual quality by introducing details, but sometimes unwanted artifacts.

Our model has a rather small model size, with good execution efficiency. Comprehensively, SimpleNet is the most competitive deblurring method in current art.

Table 3: The ablation study. All the proposed modules are contributive to the final SimpleNet. The ablated module are replaced by traditional 3-layered Resblocks. Results are PSNR (dB) and SSIM.

	PVB	LGCR	DR	Results in GoPro[20]
Baseline1	✓	✓		31.24 dB, 0.9455
Baseline2	✓		✓	31.19 dB, 0.9459
Baseline3		✓	✓	31.04 dB, 0.9443
Ours	✓	✓	✓	**31.52 dB, 0.9495**

Table 4: Time consumption of each module, with 720p input, on the average of 1000 run.

	Convs/Deconvs	ResBlocks	LGCR	PVBs	DRs
Time (ms)	36.55	46.58	57.75	137.01	98.10
Proportion	9.72%	12.39%	15.36%	36.44%	26.09%

4.3. Ablation Study

We test the contribution and consumption of each modules in SimpleNet, and results are in Tab. 3, 4. The ablated modules are replaced by traditional 3-layered Resblocks.

Our proposed PVB, LGCR, even the DR module all make non-negligible contributions to the final SimpleNet performance. PVB is the most helpful to the SimpleNet,

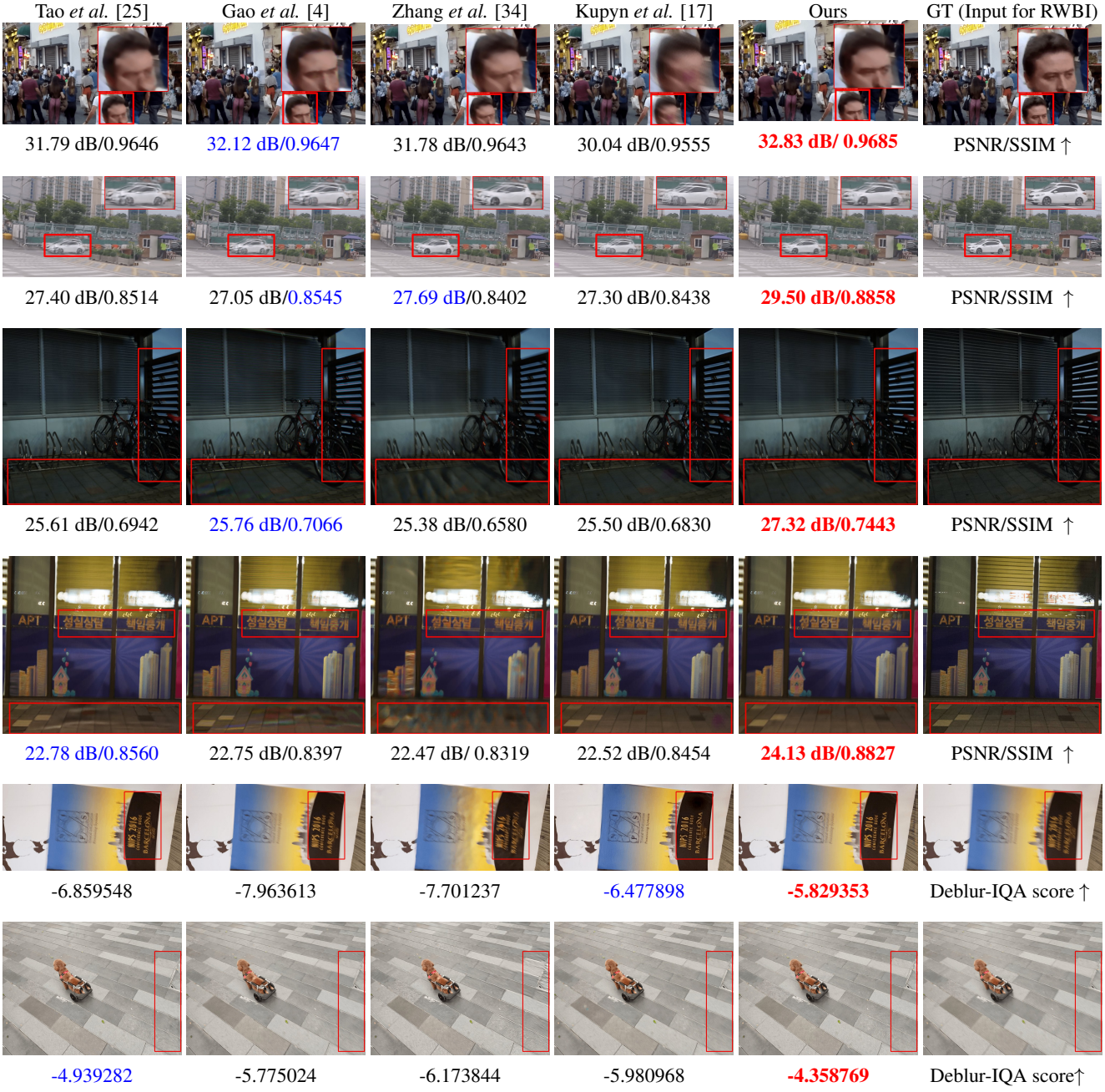

Figure 6: Visual results. The first 3 rows are from GoPro, the next two rows are from RealBlur-J, the last two rows are real world results (from RWBI). In cases for big blur scale by near object (row #1), fast moving object (row #2), or structured patterns (row #3,4,5,6), SimpleNet shows its strength. Zoom in for detail. More results are given in supplementary materials.

as shown in row 3, 4 in Tab. 3. Moreover, it also brings performance gain when transferred in each scale in [25] as shown in the last row of Tab. 5. Yet, PVB's concat and fusing cost time. The LGCR also effectively brings performance gain, as can be observed in the statistics of Baseline3 and Ours. DR costs some time for warping the sampled feature map. Visual results are in Fig. 8. Due to limited space, more ablation results are in the supplementary file.

4.4. LGCR Effectiveness

To further prove the effectiveness and efficiency of LGCR module, experiments are shown in Tab. 5, 6.

In Tab. 5, it is evident that LGCR brings more performance improvement than the others. We also find: 1) T+T

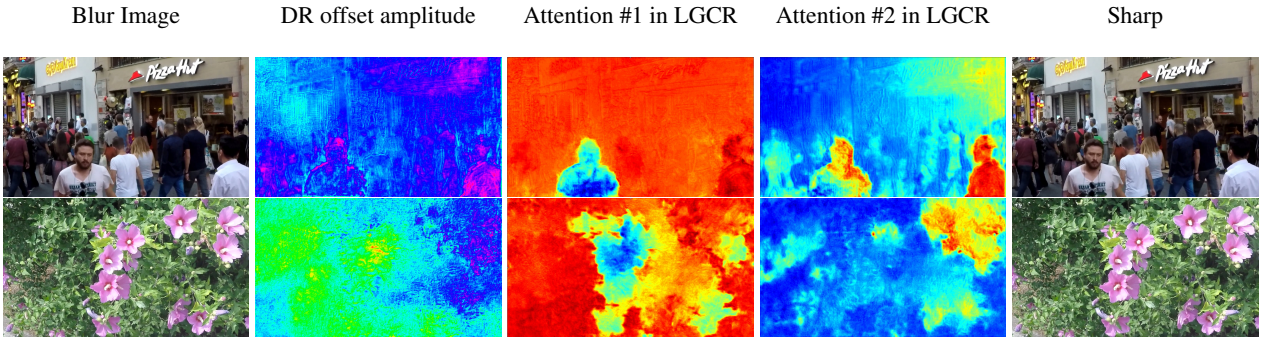

Figure 7: Visualization for the offset amplitude of a DR, and the spatial attention from 2 channels of the GCRW in LGCR.

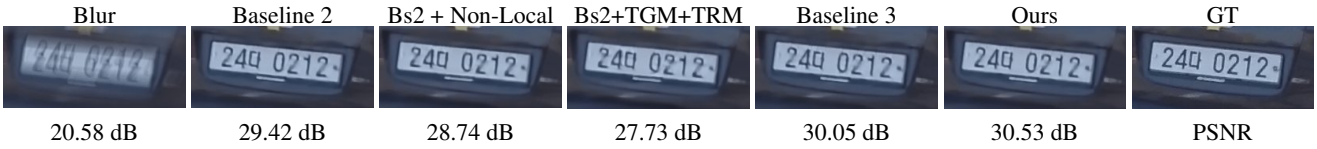

Figure 8: Visualization of ablation study. "Bs2" is short for "Baseline 2". More results are shown in supplementary materials.

Table 5: The effectiveness of LGCR and PVB. LGCR is effective, and it brings more performance improvement than Non-Local and T+T, where T+T degrades the performance. Other algorithms can also benefit from LGCR and PVB.

Method	Non-localized Module	Results in GoPro[20]
Baseline2	w/o	31.19 dB, 0.9459
Baseline2+ Non-Local	Non-Local	31.39 dB, 0.9478
Baseline2+ T+T	T+T	30.83 dB, 0.9158
Our	LGCR	**31.52 dB, 0.9495**
Tao *et al.* [25]	w/o	30.26 dB, 0.9342
[25] + LGCR	LGCR	30.38 dB, 0.9362
[25] + PVB	w/o	30.41 dB, 0.9368

Table 6: A simple memory cost and running time analysis for non-localized modules. With the input patch sized 180×180 with 3 channels, the memory and time cost are evaluated. The result is an average of 1000 run.

Non-localized Module	Memory Cost	Running Time (ms)
Non-Local	8018M	68.16
TGM+TRM	30M	0.302
LGCR	34M	0.466

severely degrades the performance, mainly because it is for pixel reasoning that tunes up/suppresses the input, not enriches it. Its lost details by GAP are not compensated either. It is also because T+T has less representation ability brought by less non-linearity in the context fragments. In summary, such design brings adverse effects to the sharp recovery. 2) The sixth record in Tab. 5 are obtained by insertion of LGCR into the encoder's end in Tao *et al.*'s method (fifth record). It proves the effectiveness and transferability of LGCR. Yet, LGCR does not provide a big performance boost than in our backbone, because Tao *et al.*'s method has already partly conquered long-range dependencies by convLSTM and multi-scale recurrent learning.

Tab. 6 shows the time and memory cost of these non-localized modules. LGCR achieves better performance than nonlocal with much less memory and time consumption.

4.5. Visualization of SimpleNet

To illustrate the correctness and learning ability of the SimpleNet, we show the offset amplitude of a DR, and the spatial attention from 2 channels of the GCRW in LGCR, in Figure 7. DR tends to learn the moving contours while the LCGR tends to potentially pay attention to the global distribution of the blur pattern.

5. Conclusion

Facing the challenge of various blur scales in deblurring tasks, we are the first to propose a light-weight globally-analyzing module, LGCR, in deblurring field. With low cost, it achieves better performance than the nonlocal and T+T units. Moreover, we propose PVB and PVB-piling strategy that enriches the variousness of the network's reception scales and perceptual ability, which helps restore images with a wide range of blur scales. Comprehensive experiments on both prevalent and new benchmarks prove the excellence of our SimpleNet.

References

[1] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *European conference on computer vision*, pages 221–235. Springer, 2016.

[2] Wanli Chen, Xinge Zhu, Ruoqi Sun, Junjun He, Ruiyu Li, Xiaoyong Shen, and Bei Yu. Tensor low-rank reconstruction for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Li Yi, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019.

[5] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2017.

[6] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: a deep learning solution for removing heterogeneous motion blur. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2319–2328, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Sun Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the International Conference on Computer Vision*, 2015.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Michal Hradiš, Jan Kotera, Pavel Zemcık, and Filip Šroubek. Convolutional neural networks for direct text deblurring. In *Proceedings of BMVC*, volume 10, page 2, 2015.

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[11] Jucheol Won Sunghyun Cho Jaesung Rim, Haeyun Lee. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[15] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[16] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2018.

[17] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8878–8887, 2019.

[18] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[19] Jichun Li, Bo Yan, Qing Lin, Ang Li, and Chenxi Ma. Motion blur removal with quality assessment guidance. *IEEE Transactions on Multimedia*, 2021.

[20] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[21] Thekke Madam Nimisha, Akash Kumar Singh, and Ambasamudram N Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4752–4760, 2017.

[22] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[24] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[25] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018.

[26] Jianyi Wang, Xin Deng, Mai Xu, Congyong Chen, and Yuhang Song. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[27] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced

deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[28] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[31] Xiangyu Xu, Jinshan Pan, Yu-Jin Zhang, and Ming-Hsuan Yang. Motion blur kernel estimation via deep learning. *IEEE Transactions on Image Processing*, 27(1):194–205, 2017.

[32] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[33] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3555–3564, 2020.

[34] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5978–5986, 2019.

[35] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020.

[36] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020.

[37] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[38] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019.