

Uformer: A General U-Shaped Transformer for Image Restoration

Zhendong Wang¹, Xiaodong Cun^{2*}, Jianmin Bao, Wengang Zhou¹, Jianzhuang Liu³, Houqiang Li¹

¹ University of Science and Technology of China, ² University of Macau,

³ University of Chinese Academy of Sciences

Abstract

In this paper, we present Uformer, an effective and efficient Transformer-based architecture for image restoration, in which we build a hierarchical encoder-decoder network using the Transformer block. In Uformer, there are two core designs. First, we introduce a novel locally-enhanced window (LeWin) Transformer block, which performs non-overlapping window-based self-attention instead of global self-attention. It significantly reduces the computational complexity on high resolution feature map while capturing local context. Second, we propose a learnable multi-scale restoration modulator in the form of a multi-scale spatial bias to adjust features in multiple layers of the Uformer decoder. Our modulator demonstrates superior capability for restoring details for various image restoration tasks while introducing marginal extra parameters and computational cost. Powered by these two designs, Uformer enjoys a high capability for capturing both local and global dependencies for image restoration. To evaluate our approach, extensive experiments are conducted on several image restoration tasks, including image denoising, motion deblurring, defocus deblurring and deraining. Without bells and whistles, our Uformer achieves superior or comparable performance compared with the state-of-the-art algorithms. The code and models are available at <https://github.com/ZhendongWang6/Uformer>.

1. Introduction

With the rapid development of consumer and industry cameras and smartphones, the requirements of removing undesired degradation (*e.g.*, noise, blur, rain, and so on) in images are constantly growing. Recovering genuine images from their degraded versions, *i.e.*, image restoration, is a classic task in computer vision. Recent state-of-the-art methods [9, 44, 71, 73, 74] are mostly ConvNets-based, which achieve impressive results but show a limitation in capturing long-range dependencies. To address this problem, several

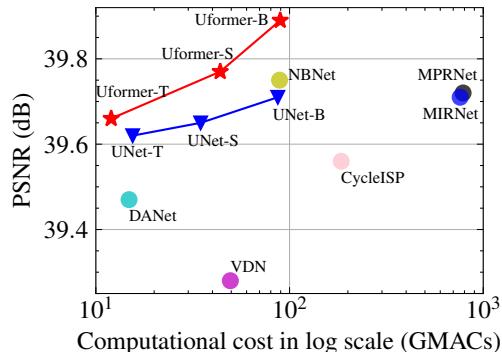


Figure 1. PSNR vs. computational cost on the SIDD dataset [1].

recent works [30, 38, 80] start to employ single or few self-attention layers in low resolution feature maps due to the self-attention computational complexity being quadratic to the feature map size.

In this paper, we aim to leverage the capability of self-attention in feature maps at multi-scale resolutions to recover more image details. To this end, we present Uformer, an effective and efficient Transformer-based structure for image restoration. Uformer is built upon an elegant architecture UNet [46], where we modify the convolution layers to Transformer blocks while keeping the same overall hierarchical encoder-decoder structure and the skip-connections.

We propose two core designs to make Uformer suitable for image restoration tasks. First, we propose the Locally-enhanced Window (LeWin) Transformer block, which is an efficient and effective basic component. The LeWin Transformer block performs non-overlapping window-based self-attention instead of global self-attention, which significantly reduces the computational complexity on high resolution feature maps. Since we build hierarchical feature maps and keep the window size unchanged, the window-based self-attention at low resolution is able to capture much more global dependencies. On the other hand, local context is essential for image restoration, we further introduce a depth-wise convolutional layer between two fully-connected layers of the feed-forward network in the Transformer block to better capture local context. We also notice that recent works [34, 69] use the similar design for different tasks.

*Corresponding author

Second, we propose a learnable multi-scale restoration modulator to handle various image degradations. The modulator is formulated as a multi-scale spatial bias to adjust features in multiple layers of the Uformer decoder. Specifically, a learnable window-based tensor is added to features in each LeWin Transformer block to adapt the features for restoring more details. Benefiting from the simple operator and window-based mechanism, it can be flexibly applied for various image restoration tasks in different frameworks.

Based on the above two designs, without bells and whistles, *e.g.*, the multi-stage or multi-scale framework [74, 75] and the advanced loss function [27, 28], our simple U-shaped Transformer structure achieves state-of-the-art performance on multiple image restoration tasks. For denoising, Uformer outperforms the previous state-of-the-art method (NBNet [9]) by 0.14 dB and 0.09 dB on the SIDD [1] and DND [43] benchmarks, respectively. For the motion blur removal task, Uformer achieves the best (Go-Pro [42], RealBlur-R [45], and RealBlur-J [45]) or competitive (HIDE [49]) performance, displaying its strong capability of deblurring. Uformer also shows the potential on the defocus deblurring task [3] and outperforms the previous best model [51] by 1.04 dB. Also, on the SPAD dataset [58] for deraining, it obtains 47.84 dB on PSNR, an improvement of 3.74 dB over the previous state-of-the-art method [44]. We expect our work will encourage further research to explore Transformer-based architectures for image restoration.

Overall, we summarize the contributions of this paper as follows:

- We present Uformer, a general and superior U-shaped Transformer for various image restoration tasks. Uformer is built on the basic LeWin Transformer block that is both efficient and effective.
- We present an extra light-weight learnable multi-scale restoration modulator to adjust on multi-scale features. This simple design significantly improves the restoration quality.
- Extensive experiments show that Uformer establishes new state-of-the-arts on various datasets for image restoration tasks.

2. Related Work

Image Restoration Architectures Image restoration aims to restore the clean image from its degraded version. A popular solution is to learn effective models using the U-shaped structures with skip-connection to capture multi-scale information hierarchically for various image restoration tasks, including image denoising [9, 71, 74], deblurring [3, 27, 28], and demoiréing [37, 52]. Some image restoration methods are inspired by the key insight from the rapid development of image classification [17, 26]. For example,

ResNet-based structure has been widely used for general image restoration [39, 80] as well as for specific tasks in image restoration such as super-resolution [36, 81] and image denoising [15, 78]. More CNN-based image restoration architectures can be found in the recent surveys [31, 54, 62] and the NTIRE Challenges [2].

Until recently, some works start to explore the attention mechanism to boost the performance. For example, squeeze-and-excitation networks [20] and non-local neural networks [60] inspire a branch of methods for different image restoration tasks, such as super-resolution [35, 79], deraining [32, 74], and denoising [73, 74]. Our Uformer also applies the hierarchical structure to build multi-scale features while using the newly introduced LeWin Transformer block as the basic building block.

Vision Transformers Transformer [56] shows a significant performance in natural language processing (NLP). Different from the design of CNNs, Transformer-based network structures are naturally good at capturing long-range dependencies in the data by the global self-attention. The success of Transformer in the NLP domain also inspires the computer vision researchers. The pioneering work of ViT [14] directly trains a pure Transformer-based architecture on the medium-size (16×16) flattened patches. With large-scale data pre-training (*i.e.*, JFT-300M), ViT gets excellent results compared to state-of-the-art CNNs on image classification.

Since the introduction of ViT, many efforts have been made to reduce the quadratic computational cost of global self-attention for making Transformer more suitable for vision tasks. Some works [19, 59] focus on establishing a pyramid Transformer architecture similar to ConvNet-based structure. To overcome the quadratic complexity of original self-attention, self-attention is performed on local windows with the halo operation or window shift [40, 55] to help cross-window interaction, and get promising results. Rather than focusing on image classification, recent works [10, 13, 22, 67, 82] propose a branch of Transformer-based backbones for more general high-level vision tasks.

Besides high-level discriminative tasks, there are also some Transformer-based works [24, 65, 83] for generative tasks. While there are a lot of explorations in the vision area, introducing Transformer to low-level vision still lacks exploration. Early work [66] makes use of self-attention mechanism to learn texture for super-resolution. As for image restoration tasks, IPT [8] first applies standard Transformer blocks within a multi-task learning framework. However, IPT relies on pretraining on a large-scale synthesized dataset and multi-task learning for good performance. In contrast, we design a general U-shaped Transformer-based structure, which proves to be efficient and effective for image restoration.

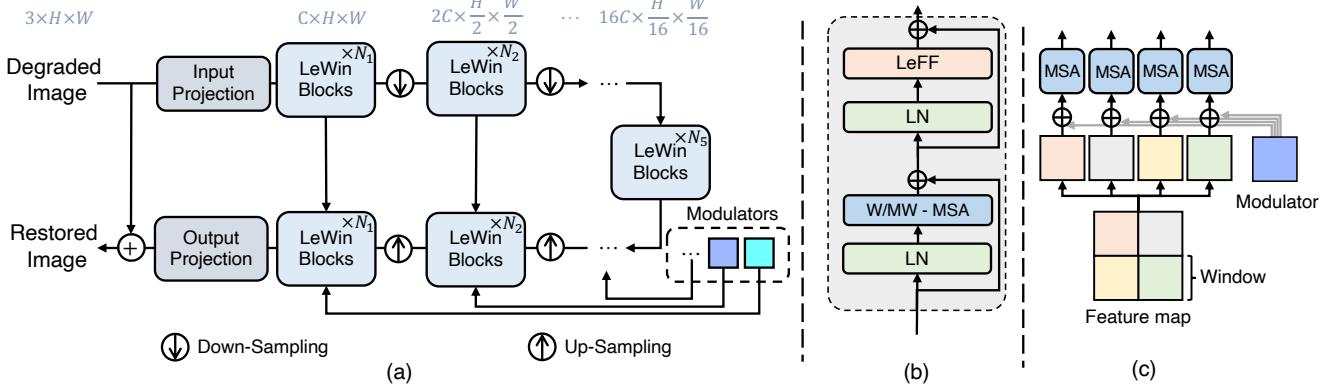


Figure 2. (a) Overview of the Uformer structure. (b) LeWin Transformer block. (c) Illustration of how the modulators modulate the W-MSAs in each LeWin Transformer block which is named MW-MSA in (b).

3. Method

In this section, we first describe the overall pipeline and the hierarchical structure of Uformer for image restoration. Then, we provide the details of the *LeWin Transformer block* which is the basic component of Uformer. After that, we present the *multi-scale restoration modulator*.

3.1. Overall Pipeline

As shown in Figure 2(a), the overall structure of the proposed Uformer is a U-shaped hierarchical network with skip-connections between the encoder and the decoder. To be specific, given a degraded image $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$, Uformer firstly applies a 3×3 convolutional layer with LeakyReLU to extract low-level features $\mathbf{X}_0 \in \mathbb{R}^{C \times H \times W}$. Next, following the design of the U-shaped structures [23, 46], the feature maps \mathbf{X}_0 are passed through K encoder stages. Each stage contains a stack of the proposed LeWin Transformer blocks and one down-sampling layer. The LeWin Transformer block takes advantage of the self-attention mechanism for capturing long-range dependencies, and also cuts the computational cost due to the usage of self-attention through non-overlapping windows on the feature maps. In the down-sampling layer, we first reshape the flattened features into 2D spatial feature maps, and then down-sample the maps, double the channels using 4×4 convolution with stride 2. For example, given the input feature maps $\mathbf{X}_0 \in \mathbb{R}^{C \times H \times W}$, the l -th stage of the encoder produces the feature maps $\mathbf{X}_l \in \mathbb{R}^{2^l C \times \frac{H}{2^l} \times \frac{W}{2^l}}$.

Then, a bottleneck stage with a stack of LeWin Transformer blocks is added at the end of the encoder. In this stage, thanks to the hierarchical structure, the Transformer blocks capture longer (even global when the window size equals the feature map size) dependencies.

For feature reconstruction, the proposed decoder also contains K stages. Each consists of an up-sampling layer and a stack of LeWin Transformer blocks similar to the encoder.

We use 2×2 transposed convolution with stride 2 for the up-sampling. This layer reduces half of the feature channels and doubles the size of the feature maps. After that, the features input to the LeWin Transformer blocks are concatenation of the up-sampled features and the corresponding features from the encoder through skip-connection. Next, the LeWin Transformer blocks are utilized to restore the image. After the K decoder stages, we reshape the flattened features to 2D feature maps and apply a 3×3 convolution layer to obtain a residual image $\mathbf{R} \in \mathbb{R}^{3 \times H \times W}$. Finally, the restored image is obtained by $\mathbf{I}' = \mathbf{I} + \mathbf{R}$. We train Uformer using the Charbonnier loss [7, 73]:

$$\ell(\mathbf{I}', \hat{\mathbf{I}}) = \sqrt{\|\mathbf{I}' - \hat{\mathbf{I}}\|^2 + \epsilon^2}, \quad (1)$$

where $\hat{\mathbf{I}}$ is the ground-truth image, and $\epsilon = 10^{-3}$ is a constant in all the experiments.

3.2. LeWin Transformer Block

There are two main challenges to apply Transformer for image restoration. First, the standard Transformer architecture [14, 56] computes self-attention globally between all tokens, which contributes to the quadratic computation cost with respect to the number of tokens. It is unsuitable to apply global self-attention on high-resolution feature maps. Second, the local context information is essential for image restoration tasks since the neighborhood of a degraded pixel can be leveraged to restore its clean version, but previous works [34, 63] suggest that Transformer shows a limitation in capturing local dependencies.

To address the above mentioned two issues, we propose a Locally-enhanced Window (LeWin) Transformer block, as shown in Figure 2(b), which benefits from the self-attention in Transformer to capture long-range dependencies, and also involves the convolution operator into Transformer to capture useful local context. Specifically, given the features

at the $(l-1)$ -th block \mathbf{X}_{l-1} , we build the block with two core designs: (1) non-overlapping Window-based Multi-head Self-Attention (W-MSA) and (2) Locally-enhanced Feed-Forward Network (LeFF). The computation of a LeWin Transformer block is represented as:

$$\begin{aligned}\mathbf{X}'_l &= \text{W-MSA}(\text{LN}(\mathbf{X}_{l-1})) + \mathbf{X}_{l-1}, \\ \mathbf{X}_l &= \text{LeFF}(\text{LN}(\mathbf{X}'_l)) + \mathbf{X}'_l,\end{aligned}\quad (2)$$

where \mathbf{X}'_l and \mathbf{X}_l are the outputs of the W-MSA module and LeFF module, respectively. LN represents the layer normalization [5]. In the following, we elaborate W-MSA and LeFF separately.

Window-based Multi-head Self-Attention (W-MSA). Instead of using global self-attention like the vanilla Transformer, we perform the self-attention within non-overlapping local windows, which reduces the computational cost significantly. Given the 2D feature maps $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ with H and W being the height and width of the maps, we split \mathbf{X} into non-overlapping windows with the window size of $M \times M$, and then get the flattened and transposed features $\mathbf{X}^i \in \mathbb{R}^{M^2 \times C}$ from each window i . Next, we perform self-attention on the flattened features in each window.

Suppose the head number is k and the head dimension is $d_k = C/k$. Then computing the k -th head self-attention in the non-overlapping windows can be formulated as follows,

$$\begin{aligned}\mathbf{X} &= \{\mathbf{X}^1, \mathbf{X}^2, \dots, \mathbf{X}^N\}, \quad N = HW/M^2, \\ \mathbf{Y}_k^i &= \text{Attention}(\mathbf{X}^i \mathbf{W}_k^Q, \mathbf{X}^i \mathbf{W}_k^K, \mathbf{X}^i \mathbf{W}_k^V), \quad i = 1, \dots, N, \\ \hat{\mathbf{X}}_k &= \{\mathbf{Y}_k^1, \mathbf{Y}_k^2, \dots, \mathbf{Y}_k^M\},\end{aligned}\quad (3)$$

where $\mathbf{W}_k^Q, \mathbf{W}_k^K, \mathbf{W}_k^V \in \mathbb{R}^{C \times d_k}$ represent the projection matrices of the queries, keys, and values for the k -th head, respectively. $\hat{\mathbf{X}}_k$ is the output of the k -th head. Then the outputs for all heads $\{1, 2, \dots, k\}$ are concatenated and then linearly projected to get the final result. Inspired by previous works [40, 48], we also apply the relative position encoding into the attention module, so the attention calculation can be formulated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{B}\right)\mathbf{V}, \quad (4)$$

where \mathbf{B} is the relative position bias, whose values are taken from $\hat{\mathbf{B}} \in \mathbb{R}^{(2M-1) \times (2M-1)}$ with learnable parameters [40, 48].

Window-based self-attention can significantly reduce the computational cost compared with global self-attention. Given the feature maps $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, the computational complexity drops from $O(H^2 W^2 C)$ to $O(\frac{HW}{M^2} M^4 C) = O(M^2 HWC)$. Since we design Uformer as a hierarchical architecture, our window-based self-attention at low resolution feature maps works on larger receptive fields and is sufficient to learn long-range dependencies. We also try the

shifted-window strategy [40] in the even LeWin Transformer block of each stage in our framework, which gives only slightly better results.

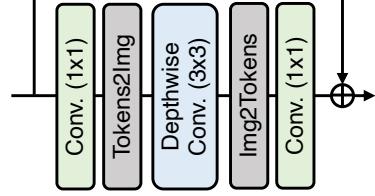


Figure 3. Locally-enhanced feed-forward network.

Locally-enhanced Feed-Forward Network (LeFF). As pointed out by previous works [63, 69], the Feed-Forward Network (FFN) in the standard Transformer suffers limited capability to leverage local context. Actually, neighboring pixels are crucial references for image restoration [6, 21]. To overcome this issue, we add a depth-wise convolutional block to the FFN in our Transformer-based structure following the recent works [34, 47, 69]. As shown in Figure 3, we first apply a linear projection layer to each token to increase its feature dimension. Next, we reshape the tokens to 2D feature maps, and use a 3×3 depth-wise convolution to capture local information. Then we flatten back the features to tokens and shrink the channels via another linear layer to match the dimension of the input channels. We use GELU [18] as the activation function after each linear/convolution layer.

3.3. Multi-Scale Restoration Modulator

Different types of image degradation (*e.g.* blur, noise, rain, etc.) have their own distinctive perturbed patterns to be handled or restored. To further boost the capability of Uformer for approaching various perturbations, we propose a light-weight multi-scale restoration modulator to calibrate the features and encourage more details recovered.

As shown in Figure 2(a) and 2(c), the multi-scale restoration modulator applies multiple modulators in the Uformer decoder. Specially in each LeWin Transformer block, a modulator is formulated as a learnable tensor with a shape of $M \times M \times C$, where M is the window size and C is the channel dimension of current feature map. Each modulator is simply served as a shared bias term that is added into all non-overlapping windows before self-attention module. Due to this light-weight addition operation and window-sized shape, the multi-scale restoration modulator introduces marginal extra parameters and computational cost.

We prove the effectiveness of the multi-scale restoration modulator on two typical image restoration tasks: image deblurring and image denoising. The visualization comparisons are presented in Figure 4. We observe that adding the multi-scale restoration modulator makes more motion blur/noising patterns removed and yields a much cleaner

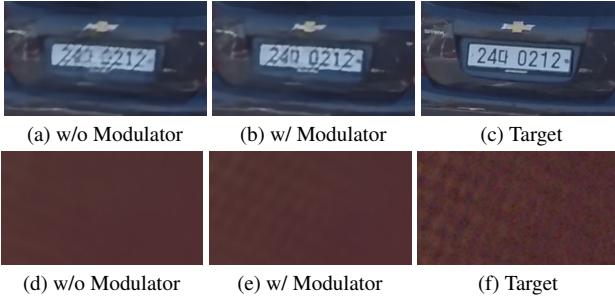


Figure 4. Effect of the multi-scale restoration modulator on image deblurring (top samples from GoPro [42]) and denoising (bottom samples from SIDD [1]). Compared with (a), Uformer w/ Modulator (b) can remove much more blur and recover the numbers accurately. Compared with (d), the image restored by Uformer w/ Modulator (e) is closer to the target with more details.

image. These results show that our multi-scale restoration modulator truly helps to recover restoration details with little computation cost. One possible explanation is that adding modulators at each stage of the decoder enables a flexible adjustment of the feature maps that boosts the performance for restoring details. This is consistent with the previous work StyleGAN [25] using a multi-scale noise term adding to the convolution features, which realizes stochastic variation for generating photo-realistic images.

4. Experiments

In this section, we first discuss the experimental setup. After that, we verify the effectiveness and efficiency of Uformer on various image restoration tasks on eight datasets. Finally, we perform comprehensive ablation studies to evaluate each component of our proposed Uformer.

4.1. Experimental Setup

Basic settings. Following the common training strategy of Transformer [56], we train our framework using the AdamW optimizer [41] with the momentum terms of $(0.9, 0.999)$ and the weight decay of 0.02. We randomly augment the training samples using the horizontal flipping and rotate the images by 90° , 180° , or 270° . We use the cosine decay strategy to decrease the learning rate to $1e-6$ with the initial learning rate $2e-4$. We set the window size to 8×8 in all LeWin Transformer blocks. The number of Uformer encoder/decoder stages K equals 4 by default. And the dimension of each head in Transformer block d_k equals C . More dataset-specific experimental settings can be found in the supplementary materials.

Architecture variants. For a concise description, we introduce three Uformer variants in our experiments, Uformer-T (Tiny), Uformer-S (Small), and Uformer-B (Base) by setting different Transformer feature channels C and the numbers of the Transformer blocks in each encoder and decoder

stages. The details are listed as follows:

- Uformer-T: $C = 16$, depths of Encoder = $\{2, 2, 2, 2\}$,
- Uformer-S: $C = 32$, depths of Encoder = $\{2, 2, 2, 2\}$,
- Uformer-B: $C = 32$, depths of Encoder = $\{1, 2, 8, 8\}$,

and the depths of Decoder are mirrored depths of Encoder.

Evaluation metrics. We adopt the commonly-used PSNR and SSIM [61] metrics to evaluate the restoration performance. These metrics are calculated in the RGB color space except for deraining where we evaluate the PSNR and SSIM on the Y channel in the YCbCr color space, following the previous work [57].

Method	SIDD		DND	
	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
BM3D [12]	25.65	0.685	34.51	0.851
RIDNet [4]	38.71	0.914	39.26	0.953
VDN [70]	39.28	0.909	39.38	0.952
DANet [71]	39.47	0.918	39.59	0.955
CycleISP [72]	39.52	0.957	39.56	0.956
MIRNet [73]	39.72	<u>0.959</u>	39.88	0.956
MPRNet [74]	39.71	0.958	39.80	0.954
NBNet [9]	<u>39.75</u>	<u>0.959</u>	<u>39.89</u>	<u>0.955</u>
Uformer-B	39.89	0.960	39.98	0.955

Table 1. Denoising results on the SIDD [1] and DND [43] datasets.

4.2. Real Noise Removal

Table 1 reports the results of real noise removal on the SIDD [1] and DND [43] datasets. We compare Uformer with 8 state-of-the-art denoising methods, including the feature-based BM3D [12] and seven learning-based methods: RIDNet [4], VDN [70], CycleISP [72], NBNet [9], DANet [71], MIRNet [73], and MPRNet [74]. Our Uformer-B achieves 39.89 dB on PSNR, surpassing all the other methods by at least 0.14 dB. As for the DND dataset, we follow the common evaluation strategy and test our model trained on SIDD via the online server testing. Uformer outperforms the previous state-of-the-art method NBNet [9] by 0.09 dB. To verify whether the gains benefit from more computation cost, we present the results of PSNR vs. computational cost in Figure 1. We notice that our Uformer-T can achieve a better performance than most models but with the least computation cost, which demonstrates the efficiency and effectiveness of Uformer. We also show the qualitative results on the SIDD and DND datasets in Figure 5, in which Uformer can not only successfully remove the noise but also keep the texture details.

4.3. Motion Blur Removal

For motion blur removal, Uformer also shows state-of-the-art performance. We follow the previous method [74] to

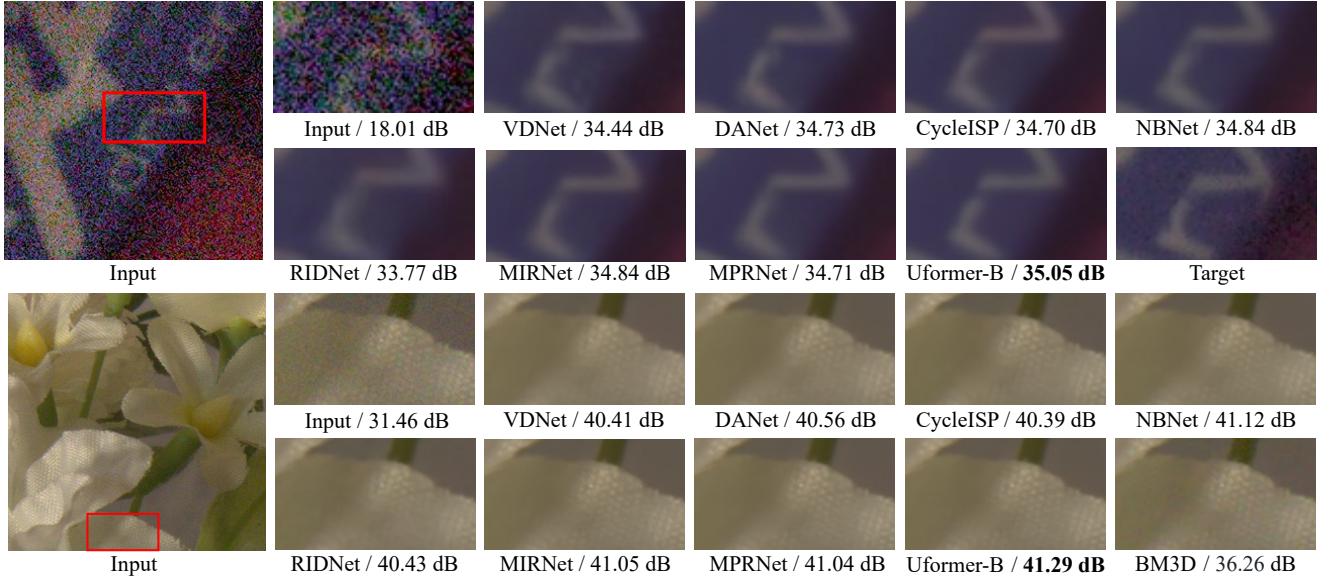


Figure 5. Visual comparisons with state-of-the-art methods on real noise removal. The top sample comes from SIDD while the bottom one is from DND.

Method	GoPro		HIDE		RealBlur-R		RealBlur-J	
	PSNR ↑	SSIM ↑						
Nah <i>et al.</i> [42]	29.08	0.914	25.73	0.874	32.51	0.841	27.87	0.827
DeblurGAN [27]	28.70	0.858	24.51	0.871	33.79	0.903	27.97	0.834
Xu <i>et al.</i> [64]	21.00	0.741	-	-	34.46	0.937	27.14	0.830
DeblurGAN-v2 [28]	29.55	0.934	26.61	0.875	35.26	0.944	28.70	0.866
DBGAN [77]	31.10	0.942	28.94	0.915	-	-	-	-
SPAIR [44]	32.06	0.953	30.29	0.931	-	-	<u>28.81</u>	0.875
†Zhang <i>et al.</i> [76]	29.19	0.931	-	-	35.48	0.947	27.80	0.847
†SRN [53]	30.26	0.934	28.36	0.915	35.66	0.947	28.56	0.867
†DMPHN [75]	31.20	0.940	29.09	0.924	35.70	0.948	28.42	0.860
†MPRNet [74]	32.66	<u>0.959</u>	30.96	<u>0.939</u>	35.99	<u>0.952</u>	28.70	<u>0.873</u>
Uformer-B	32.97	0.967	<u>30.83</u>	0.952	36.22	0.957	29.06	0.884

Table 2. Results on motion deblurring. Following previous works [27, 28, 74], our Uformer is only trained on the GoPro dataset [42]. Then we apply our GoPro trained model directly on the HIDE dataset [49] and the RealBlur dataset [45] to evaluate the generalization on real scenes. † denotes recurrent/multi-stage designs for better performance.

train Uformer on the GoPro dataset and test it on the four datasets: two synthesized datasets (HIDE [49] and the test set of GoPro [42]), and two real-world datasets (RealBlur-R/J from the RealBlur dataset [45]). We compare Uformer with ten state-of-the-art methods: Nah *et al.* [42], DeblurGAN [27], Xu *et al.* [64], DeblurGAN-v2 [28], DBGAN [77], SPAIR [44], Zhang *et al.* [76], SRN [53], DMPHN [75], and MPRNet [74]. The results are reported in Table 2. For synthetic deblurring, Uformer gets significant better performance on GoPro than previous state-of-the-art methods and shows a comparable result on the HIDE dataset. As for real-world deblurring, the causes of blur are compli-

cated so the task is usually more challenging. Our Uformer outperforms other methods by at least 0.23 dB and 0.36 dB on RealBlur-R and RealBlur-J, respectively, showing a strong generalization ability. Besides, we show some visual results in Figure 6. Compared with other methods, the images restored by Uformer are more clear and closer to their ground truth.

4.4. Defocus Blur Removal

We perform defocus blur removal on the DPD dataset [3]. Table 3 and Figure 7 report the quantitative and qualitative results, respectively. Uformer achieves a better per-

formance (1.04 dB, 1.15 dB, 1.44 dB, and 1.87 dB) over previous state-of-the-art methods KPAC [51], DPDNet [3], JNB [50], and DMENet [29], respectively. From the visualization results, we observe that the images recovered by Uformer are sharper and closer to the ground-truth images.

	DMENet [29]	JNB [50]	DPDNet [3]	KPAC [51]	Uformer-B
PSNR \uparrow	23.41	23.84	25.13	25.24	26.28
SSIM \uparrow	0.714	0.715	0.786	0.842	0.891

Table 3. Results on the DPD dataset [3] for defocus blur removal .

4.5. Real Rain Removal

We conduct the deraining experiments on SPAD [58] and compare with 6 deraining methods: GMM [33], RESCAN [32], SPANet [58], JORDER-E [68], RCDNet [57], and SPAIR [44]. As shown in Table 4, Uformer presents a significantly better performance, achieving 3.74 dB improvement over the previous best work [44]. This indicates the strong capability of Uformer for deraining on this real derain dataset. We also provide the visual results in Figure 7 where Uformer can remove the rain more successfully while introducing fewer artifacts.

	GMM [33]	RESCAN [32]	SPANet [58]	JORDER-E [68]	RCDNet [57]	SPAIR [44]	Uformer-B
PSNR \uparrow	34.30	38.11	40.24	40.78	41.47	44.10	47.84
SSIM \uparrow	0.9428	0.9707	0.9811	0.9811	0.9834	0.9872	0.9925

Table 4. Results on the SPAD dataset [58] for real rain removal.

4.6. Ablation Study

In this section, we analyze the effect of each component of Uformer in detail. The evaluations are conducted on image denoising (SIDD [1]), deblurring (GoPro [42], RealBlur [45]), and deraining (SPAD [58]) using different variants. The ablation results are reported in Tables 5, 6, and 7.

Transformer vs. convolution. We replace all the LeWin Transformer blocks in Uformer with the convolution-based ResBlocks [9], resulting in the so-called "UNet", while keeping all others unchanged. Similar to the Uformer variants, we design UNet-T/S-B:

- UNet-T: $C = 32$, depths of Encoder = {2, 2, 2, 2},
- UNet-S: $C = 48$, depths of Encoder = {2, 2, 2, 2},
- UNet-B: $C = 76$, depths of Encoder = {2, 2, 2, 2},

and the depths of Decoder are mirrored depths of Encoder.

Table 5 reports the comparison results. We observe that Uformer-T achieves 39.66 dB and outperforms UNet-T by 0.04 dB with fewer parameters and less computation.

	GMACs	# Param	PSNR \uparrow
UNet-T	15.49G	9.50M	39.62
UNet-S	34.76G	21.38M	39.65
UNet-B	86.97G	53.58M	39.71
ViT	8.83G	14.86M	38.51
Uformer-T	12.00G	5.23M	39.66
Uformer-S	43.86G	20.63M	39.77
Uformer-B	89.46G	50.88M	39.89

Table 5. Comparison of different network architectures for denoising on the SIDD dataset [1].

	W-MSA	FFN	GMACs	# Param	PSNR \uparrow
Uformer-S (SIDD [1])	-	-	43.00G	20.47M	39.74
	✓	-	43.64G	20.59M	39.72
	-	✓	43.86G	20.63M	39.77
Uformer-B (RealBlur-R/J [45])	-	-	88.31G	50.45M	36.15/28.99
	✓	✓	90.31G	51.20M	36.19/28.85
	-	✓	89.46G	50.88M	36.22/29.06

Table 6. Effect of enhancing locality in different modules.

Uformer-S achieves 39.77 dB and outperforms UNet-S by 0.12 dB with fewer parameters and a slightly higher computation cost. And Uformer-B achieves 39.89 dB which outperforms UNet-B by 0.18 dB. This study indicates the effectiveness of the proposed LeWin Transformer block, compared with the original convolutional block.

Hierarchical structure vs. single scale. We further build a ViT-based architecture which only contains a single scale of the feature maps for image denoising. This architecture employs a head of two convolution layers for extracting features from the input image and also a tail of two convolution layers for the output. 12 standard Transformer blocks are used between the head and the tail. We train the ViT with the hidden dimension of 256 on patch size 16×16 . The results are presented in Table 5. We observe that the vanilla ViT structure gets an unsatisfactory result compared with UNet, while our Uformer significantly outperforms both the ViT-based and UNet architectures, which demonstrates the effectiveness of hierarchical structure for image restoration.

Where to enhance locality? Table 6 compares the results of no locality enhancement and enhancing locality in the self-attention calculation [63] or the feed-forward network based on Uformer-S and Uformer-B. We observe that introducing locality into the feed-forward network yields 0.03 dB (SIDD), 0.07 dB (RealBlur-R)/0.07 dB (RealBlur-J) over the baseline (no locality enhancement), while introducing locality into the self-attention yields -0.02 dB (SIDD). Further, we combine introducing locality into the feed-forward network and introducing into the self-attention. The results on RealBlur-R/J also drop from 36.22 dB/29.06 dB to 36.19 dB/28.85 dB, indicating that compared to involving locality into self-attention, introducing locality into the feed-forward network is more suitable for image restoration

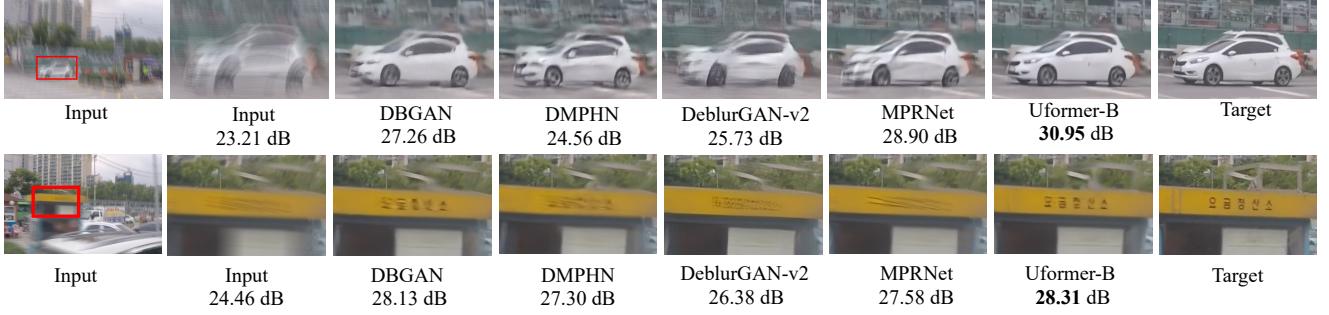


Figure 6. Visual comparisons with state-of-the-art methods on the GoPro dataset [42] for motion blur removal.

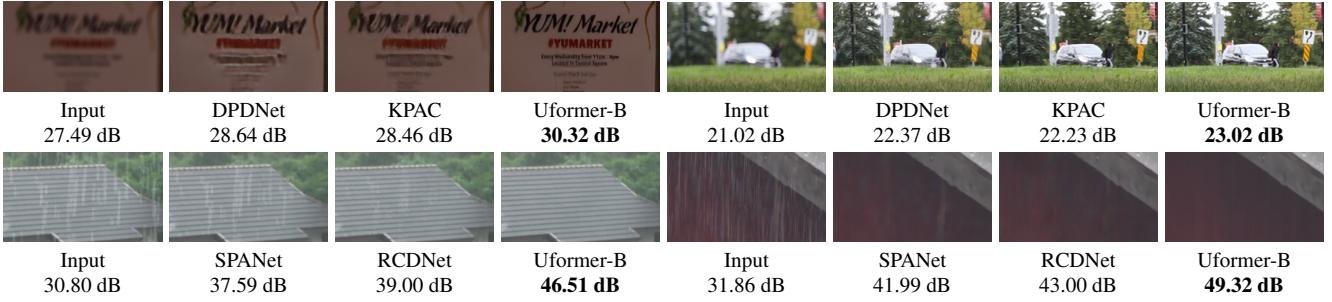


Figure 7. Top row: Visual comparisons with state-of-the-art methods on the DPD dataset [3] for defocus blur removal. Bottom row: Visual comparisons with state-of-the-art methods on the SPAD dataset [58] for real rain removal.

tasks.

	GoPro [42] Uformer-T	SIDD [1] Uformer-B	SPAD [58] Uformer-B
Modulator	- ✓	- ✓	- ✓
PSNR ↑	29.11 29.57	39.86 39.89	47.43 47.84

Table 7. Effect of the multi-scale restoration modulator.

Effect of the multi-scale restoration modulator. In Table 7, to verify the effect of the modulator, we conduct experiments on GoPro for image deblurring, SIDD for image denoising, and SPAD for deraining. For deblurring, we observe that w/ modulator can bring a performance improvement of 0.46 dB, which reveals the effectiveness of the modulator for deblurring. We also compare the results of Uformer-B with/without the modulator on SIDD and SPAD, and the comparisons indicate that the proposed modulator introduces 0.03 dB improvement (SIDD)/0.41 dB improvement (SPAD). In Figure 4, we have provided visual comparisons of Uformer w/ and wo/ the modulator. This study validates the proposed modulator can bring extra ability of restoring more details.

5. Discussion and Conclusion

In this paper, we have presented an alternative architecture Uformer for image restoration tasks by introducing the

Transformer block. In contrast to existing ConvNet-based structures, our Uformer builds upon the main component LeWin Transformer block, which can not only handle local context but also capture long-range dependencies efficiently. To handle various image restoration degradation and enhance restoration quality, we propose a learnable multi-scale restoration modulator inserted into the Uformer decoder. Extensive experiments demonstrate that Uformer achieves state-of-the art performance on several tasks, including denoising, motion deblurring, defocus deblurring, and deraining. Uformer also surpasses the UNet family by a large margin with less computation cost and fewer model parameters.

Limitation and broader impacts. Thanks to the proposed architecture, Uformer achieves the state-of-the-art performance on a variety of image restoration tasks (image denoising, deblurring, and deraining). But we have not evaluated Uformer for more vision tasks such as image-to-image translation, image super-resolution, and so on. We look forward to investigating Uformer for more applications. Meanwhile, we notice that there are several negative impacts caused by abusing image restoration techniques. For example, it may cause human privacy issue with the restored images in surveillance. The techniques may destroy the original patterns for camera identification and multi-media copyright [11], which hurts the authenticity for image forensics.

References

- [1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A High-Quality Denoising Dataset for Smartphone Cameras. In *CVPR*, 2018. 1, 2, 5, 7, 8, 12, 14
- [2] Abdelrahman Abdelhamed, Radu Timofte, and Michael S. Brown. NTIRE 2019 Challenge on Real Image Denoising: Methods and Results. In *CVPR Workshop*, 2019. 2
- [3] Abdullah Abuolaim and Michael S Brown. Defocus Deblurring Using Dual-Pixel Data. In *ECCV*. Springer, 2020. 2, 6, 7, 8, 13, 16
- [4] Saeed Anwar and Nick Barnes. Real Image Denoising with Feature Attention. In *ICCV*, 2019. 5
- [5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4
- [6] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *CVPR*, 2005. 4
- [7] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two Deterministic Half-Quadratic Regularization Algorithms for Computed Imaging. In *ICIP*, 1994. 3
- [8] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-Trained Image Processing Transformer. In *CVPR*, 2021. 2
- [9] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Hao-jiang Fan, and Shuaicheng Liu. NBNet: Noise Basis Learning for Image Denoising with Subspace Projection. In *CVPR*, 2021. 1, 2, 5, 7, 12, 13
- [10] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. In *NeurIPS*, 2021. 2
- [11] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a CNN-based camera model fingerprint. *arXiv preprint arXiv:1808.08396*, 2018. 8
- [12] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *TIP*, 16(8):2080–2095, 2007. 5
- [13] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. *arXiv preprint arXiv:2107.00652*, 2021. 2
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2020. 2, 3
- [15] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-Guided Network for Fast Image Denoising. In *ICCV*, 2019. 2
- [16] Bin He, Ce Wang, Boxin Shi, and Ling-Yu Duan. Mop Moire Patterns Using MopNet. In *ICCV*, 2019. 12, 13
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2
- [18] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 4
- [19] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking Spatial Dimensions of Vision Transformers. In *ICCV*, 2021. 2
- [20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2018. 2
- [21] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2Neighbor: Self-Supervised Denoising from Single Noisy Images. In *CVPR*, 2021. 4
- [22] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. *arXiv preprint arXiv:2106.03650*, 2021. 2
- [23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 2017. 3
- [24] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. *arXiv preprint arXiv:2102.07074*, 2021. 2
- [25] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019. 5
- [26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NeurIPS*, 2012. 2
- [27] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. *ArXiv e-prints*, 2017. 2, 6
- [28] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In *ICCV*, 2019. 2, 6
- [29] Junyoung Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep Defocus Map Estimation using Domain Adaptation. In *CVPR*, 2019. 7
- [30] Guanbin Li, Xiang He, Wei Zhang, Huiyou Chang, Le Dong, and Liang Lin. Non-locally Enhanced Encoder-Decoder Network for Single Image De-raining. In *ACMMM*, 2018. 1
- [31] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, 2019. 2
- [32] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent Squeeze-and-Excitation Context Aggregation Net for Single Image Deraining. In *ECCV*, 2018. 2, 7
- [33] Yu Li, Robby T. Tan, Xiaojie Guo, Jiangbo Lu, and Michael S. Brown. Rain Streak Removal Using Layer Priors. In *CVPR*, 2016. 7
- [34] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. LocalViT: Bringing Locality to Vision Transformers. *arXiv preprint arXiv:2104.05707*, 2021. 1, 3, 4

- [35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image Restoration Using Swin Transformer. In *ICCV Workshops*, 2021. 2
- [36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *CVPR Workshop*, 2017. 2
- [37] Bolin Liu, Xiao Shu, and Xiaolin Wu. Demoiréing of Camera-Captured Screen Images Using Deep Convolutional Neural Network. *arXiv preprint arXiv:1804.03809*, 2018. 2, 12
- [38] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-Local Recurrent Network for Image Restoration. In *NeurIPS*, 2018. 1
- [39] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual Residual Networks Leveraging the Potential of Paired Operations for Image Restoration. In *CVPR*, 2019. 2
- [40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 4, 12
- [41] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [42] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. In *CVPR*, 2017. 2, 5, 6, 7, 8, 13
- [43] Tobias Plotz and Stefan Roth. Benchmarking Denoising Algorithms with Real Photographs. In *CVPR*, 2017. 2, 5, 13
- [44] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-Adaptive Image Restoration using Distortion-Guided Networks. In *ICCV*, 2021. 1, 2, 6, 7
- [45] Jaesung Rim, Haeyun Lee, Juchol Won, and Sunghyun Cho. Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms. In *ECCV*, 2020. 2, 6, 7, 13
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*. Springer, 2015. 1, 3, 12
- [47] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *CVPR*, 2018. 4
- [48] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. *arXiv preprint arXiv:1803.02155*, 2018. 4
- [49] Ziyi Shen, Wenguan Wang, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-Aware Motion Deblurring. In *ICCV*, 2019. 2, 6, 13
- [50] Jianping Shi, Li Xu, and Jiaya Jia. Just Noticeable Defocus Blur Detection and Estimation. In *CVPR*, 2015. 7
- [51] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single Image Defocus Deblurring Using Kernel-Sharing Parallel Atrous Convolutions. In *ICCV*, 2021. 2, 7
- [52] Yujing Sun, Yizhou Yu, and Wenping Wang. Moiré photo restoration using multiresolution convolutional neural networks. *TIP*, 27(8):4160–4172, 2018. 2, 12, 13
- [53] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent Network for Deep Image Deblurring. In *CVPR*, 2018. 6
- [54] Chunwei Tian, Yong Xu, Lunke Fei, and Ke Yan. Deep Learning for Image Denoising: A Survey. In *International Conference on Genetic and Evolutionary Computing*, 2018. 2
- [55] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling Local Self-Attention for Parameter Efficient Visual Backbones. In *CVPR*, 2021. 2
- [56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017. 2, 3, 5, 12
- [57] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A Model-Driven Deep Neural Network for Single Image Rain Removal. In *CVPR*, 2020. 5, 7
- [58] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial Attentive Single-Image Deraining with a High Quality Real Rain Dataset. In *CVPR*, 2019. 2, 7, 8, 13, 15, 16, 17
- [59] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *ICCV*, 2021. 2
- [60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local Neural Networks. In *CVPR*, 2018. 2
- [61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *TIP*, 13(4):600–612, 2004. 5
- [62] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep Learning for Image Super-resolution: A Survey. *TPAMI*, 2020. 2
- [63] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021. 3, 4, 7
- [64] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural L0 Sparse Representation for Natural Image Deblurring. In *CVPR*, 2013. 6
- [65] Rui Xu, Xiangyu Xu, Kai Chen, Bolei Zhou, and Chen Change Loy. STransGAN: An Empirical Study on Transformer in GANs. *arXiv preprint arXiv:2110.13107*, 2021. 2
- [66] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning Texture Transformer Network for Image Super-Resolution. In *CVPR*, 2020. 2
- [67] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, and Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. *arXiv preprint arXiv:2107.00641*, 2021. 2
- [68] Wenhan Yang, Robby T. Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint Rain Detection and Removal from a Single Image with Contextualized Deep Networks. *TPAMI*, 42(6):1377–1393, 2020. 7
- [69] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating Convolution Designs into Visual Transformers. *arXiv preprint arXiv:2103.11816*, 2021. 1, 4
- [70] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational Denoising Network: Toward Blind Noise Modeling and Removal. In *NeurIPS*, 2019. 5

- [71] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual Adversarial Network: Toward Real-world Noise Removal and Noise Generation. In *ECCV*, 2020. 1, 2, 5, 12, 13
- [72] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *CVPR*, 2020. 5
- [73] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning Enriched Features for Real Image Restoration and Enhancement. In *ECCV*, 2020. 1, 2, 3, 5, 13
- [74] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-Stage Progressive Image Restoration. In *CVPR*, 2021. 1, 2, 5, 6, 13
- [75] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep Stacked Hierarchical Multi-patch Network for Image Deblurring. In *CVPR*, 2019. 2, 6, 13
- [76] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic Scene Deblurring Using Spatially Variant Recurrent Neural Networks. In *CVPR*, 2018. 6
- [77] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by Realistic Blurring. In *CVPR*, 2020. 6
- [78] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *TIP*, 26(7):3142–3155, 2017. 2
- [79] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *ECCV*, 2018. 2
- [80] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual Non-local Attention Networks for Image Restoration. In *ICLR*, 2019. 1, 2
- [81] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual Dense Network for Image Super-Resolution. In *CVPR*, 2018. 2
- [82] Zizhao Zhang, Han Zhang, Long Zhao, Ting Chen, and Tomas Pfister. Aggregating Nested Transformers. In *arXiv preprint arXiv:2105.12723*, 2021. 2
- [83] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris N Metaxas, and Han Zhang. Improved Transformer for High-Resolution GANs. *arXiv preprint arXiv:2106.07631*, 2021. 2

A. Additional Ablation Study

A.1. Is Window Shift Important

Table 8 reports the results of whether to use the shifted window design [40] in Uformer. We observe that window shift brings an improvement of 0.01 dB for image denoising. We use the window shift as the default setting in our experiments.

Uformer-S	PSNR \uparrow
w/o window shift	39.76
w window shift	39.77

Table 8. Effect of window shift.

A.2. Variants of Skip-Connections

To investigate how to deliver the learned low-level features from the encoder to the decoder, considering the self-attention computing in Transformer, we present three different skip-connection schemes, including concatenation-based skip-connection, cross-attention as skip-connection, and concatenation-based cross-attention as skip-connection.

Concatenation-based Skip-connection (Concat-Skip). Concat-Skip is based on the widely-used skip-connection in UNet [9, 46, 71]. To build our network, firstly, we concatenate the l -th stage flattened features E_l and each encoder stage with the features D_{K-l+1} from the $(K-l+1)$ -th decoder stage channel-wisely. Here, K is the number of the encoder/decoder stages. Then, we feed the concatenated features to the W-MSA component of the first LeWin Transformer block in the decoder stage, as shown in Figure 8(a).

Cross-attention as Skip-connection (Cross-Skip). Instead of directly concatenating features from the encoder and the decoder, we design Cross-Skip inspired by the decoder structure in the language Transformer [56]. As shown in Figure 8(b), we first add an additional attention module into the first LeWin Transformer block in each decoder stage. The first self-attention module in this block (the shaded one) is used to seek the self-similarity pixel-wisely from the decoder features D_{K-l+1} , and the second attention module in this block takes the features E_l from the encoder as the *keys* and *values*, and uses the features from the first module as the *queries*.

Concatenation-based Cross-attention as Skip-connection (ConcatCross-Skip). Combining above two variants, we also design another skip-connection. As illustrated in Figure 8(c), we concatenate the features E_l from the encoder and D_{K-l+1} from the decoder as the *keys* and *values*, while the *queries* are only from the decoder.

Table 9 compares the results of using different skip-connections in our Uformer: concatenating features (*Concat*), cross-attention (*Cross*), and concatenating *keys* and

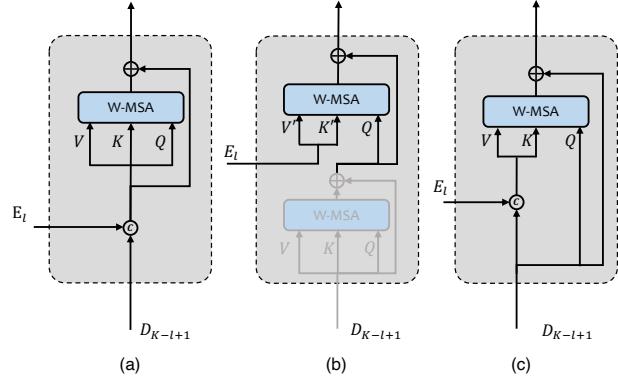


Figure 8. Three skip-connection schemes: (a) Concat-Skip, (b) Cross-Skip, and (c) ConcatCross-Skip.

	GMACs	# Param	PSNR \uparrow
Uformer-S- <i>Concat</i>	43.86G	20.63M	39.77
Uformer-S- <i>Cross</i>	44.78G	27.95M	39.75
Uformer-S- <i>ConcatCross</i>	42.75G	27.28M	39.73

Table 9. Different skip-connections.

values for cross-attention (*ConcatCross*). For a fair comparison, we increase the channels in Uformer-S from 32 to 44 in variants *Cross* and *ConcatCross*. These three skip-connections achieve similar results, and concatenating features gets slightly better performance. We adopt the feature concatenation as the default setting in Uformer.

B. Additional Experiment for Demoireing

We also conduct an experiment of moire pattern removal on the TIP18 dataset [52]. As shown in Table 10, Uformer outperforms previous methods MopNet [16], MSNet [52], CFNet [37], UNet [46] by 1.53 dB, 2.29 dB, 3.19 dB, and 2.79 dB, respectively. And in Figure 13, we show examples of visual comparisons with other methods. This experiment further demonstrates the superiority of Uformer.

	UNet [46]	CFNet [37]	MSNet [52]	MopNet [16]	Uformer-B
PSNR \uparrow	26.49	26.09	26.99	27.75	29.28
SSIM \uparrow	0.864	0.863	0.871	0.895	0.917

Table 10. Results on the TIP18 dataset [52] for demoireing.

C. Additional Experimental Settings for Different Tasks

Denoising. The training samples are randomly cropped from the original images in SIDD [1] with size 128×128 , which is also the common training strategy for image denoising in

recent works [9, 71, 73]. And the training process lasts for 250 epochs with batch size 32. Then, the trained model is evaluated on the 256×256 patches of SIDD and 512×512 patches of the DND test images [43], following [9, 73]. The results on DND are online evaluated.

Motion deblurring. Following previous methods [74, 75], we train Uformer only on the GoPro dataset [42], and evaluate it on the test set of GoPro, HIDE [49], and RealBlur-R/J [45]. The training patches are randomly cropped from the training set with size 256×256 . The batch size is set to 32. For validation, we use the central crop with size 256×256 . The number of training epochs is 3k. For evaluation, the trained model is tested on the full-size test images.

Defocus deblurring. Following the official patch segmentation algorithm [3] of DPD, we crop the training and validation samples to 60% overlapping 512×512 patches to train the model. We also discard 30% of the patches that have the lowest sharpness energy (by applying Sobel filter to the patches) as [3]. The whole training process lasts for 160 epochs with batch size 4. For evaluation, the trained model is tested on the full-size test images.

Deraining. We conduct deraining experiments on the SPAD dataset [58]. This dataset contains over 64k 256×256 images for training and 1k 512×512 images for evaluation. We train Uformer on two GPUs, with mini-batches of size 16 on the 256×256 samples. Since this dataset is large enough and the training process converges fast, we just train Uformer for 10 epochs in the experiment. Finally, we evaluate the performance on the test images following the default setting in [58].

Demoireing. We further validate the effectiveness of Uformer on the TIP18 dataset [52] for demoireing. Since the images in this dataset contain additional borders, following [16], we crop the central regions with the ratio of [0.15, 0.85] in all training/validation/testing splits and resize them to 256×256 for training and evaluation. Since this task is sensitive to the down-sampling operation, we choose the bilinear interpolation same as the previous work [16]¹. The training epochs are 250.

D. More Visual Comparisons

As shown in Figures 9-13 in this supplementary materials, we give more visual results of our Uformer and others on the five tasks (denoising, motion deblurring, defocus deblurring, deraining, and demoireing) as the supplement of the visualization in the main paper.

¹The dataset we used is also downloaded from the Github Page of [16].

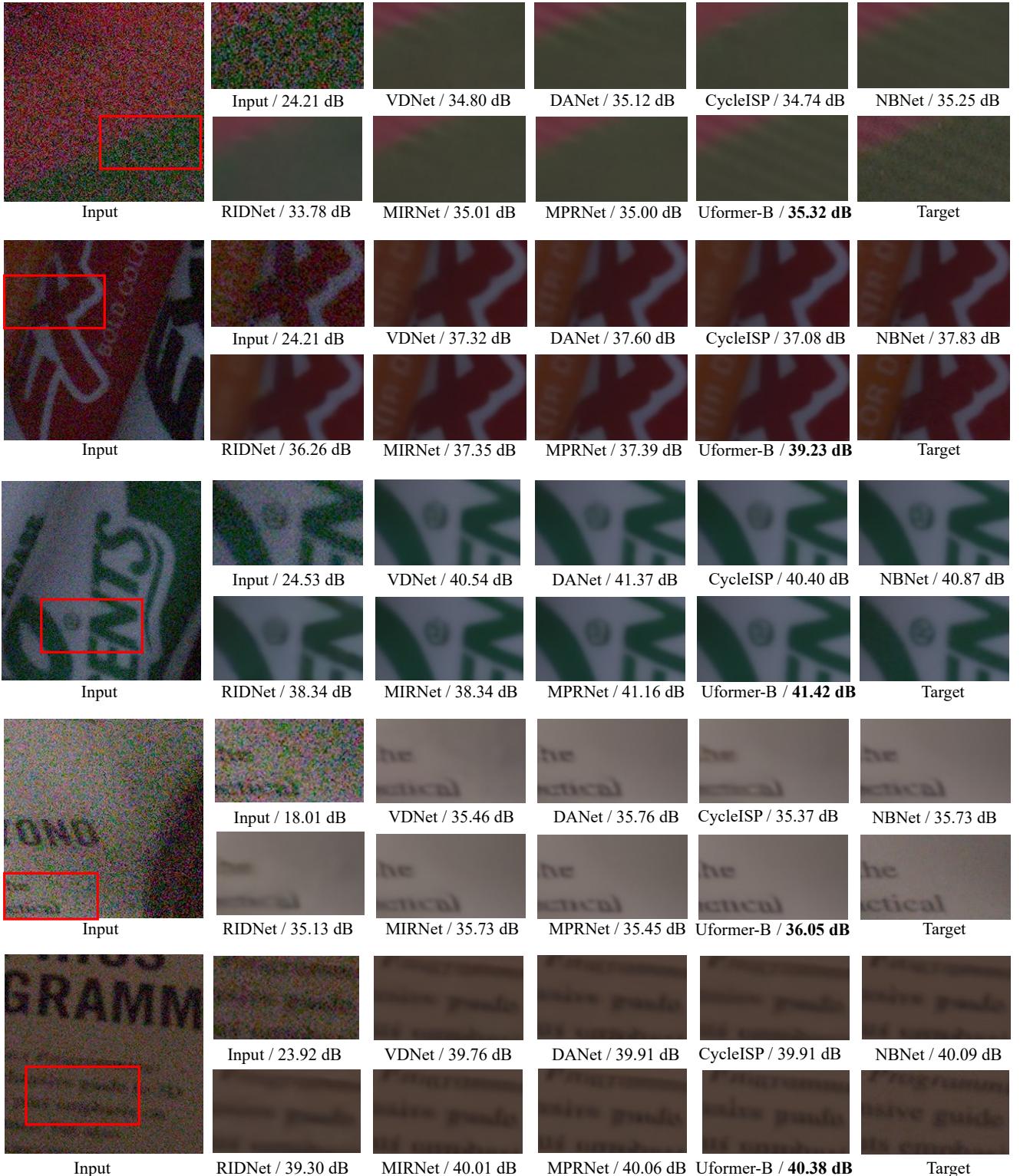


Figure 9. More visual results on the SIDD dataset [1] for image denoising. The PSNR value under each patch is computed on the corresponding whole image.

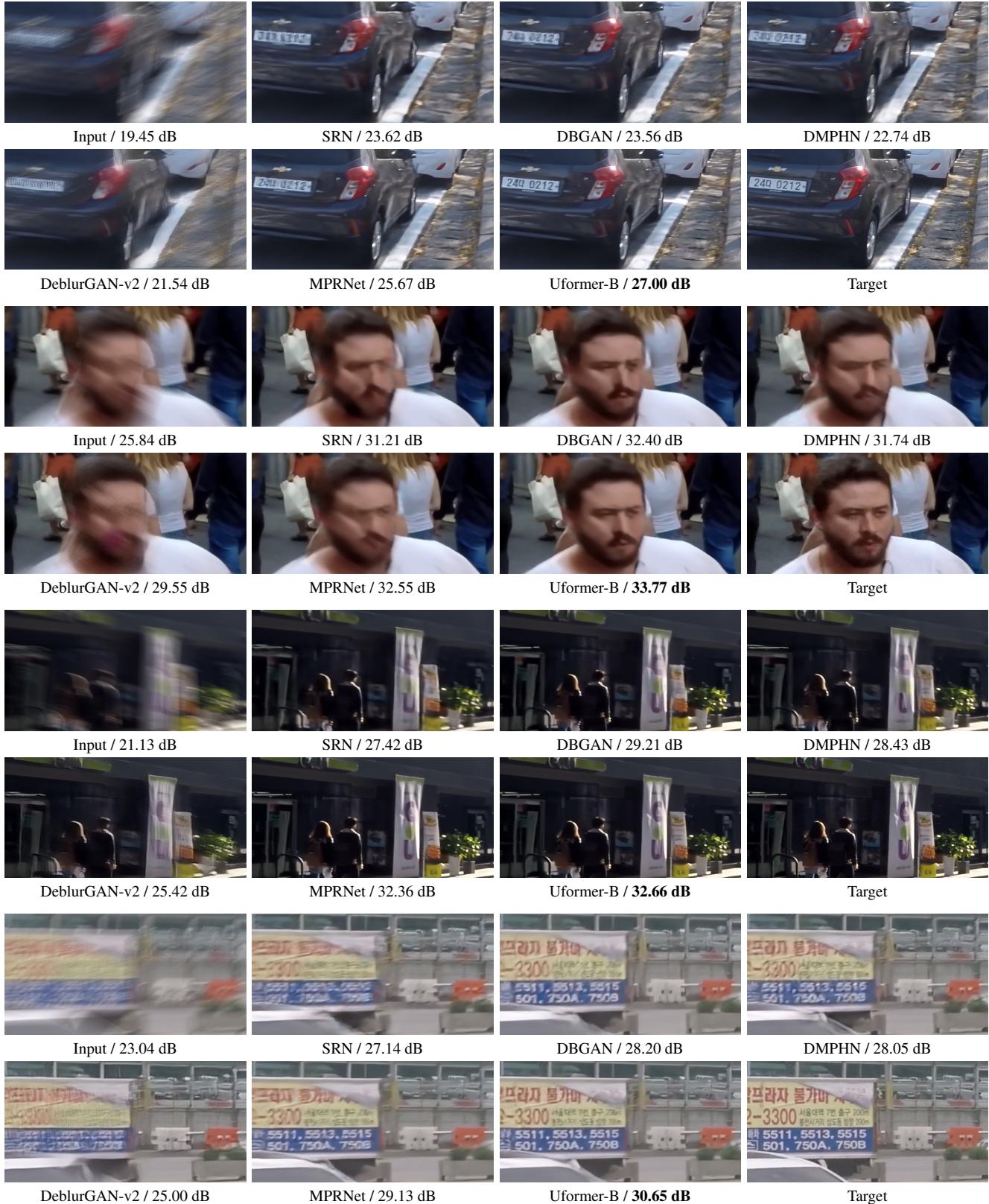


Figure 10. More results on GoPro [58] for image motion deblurring. The PSNR value under each patch is computed on the corresponding whole image.

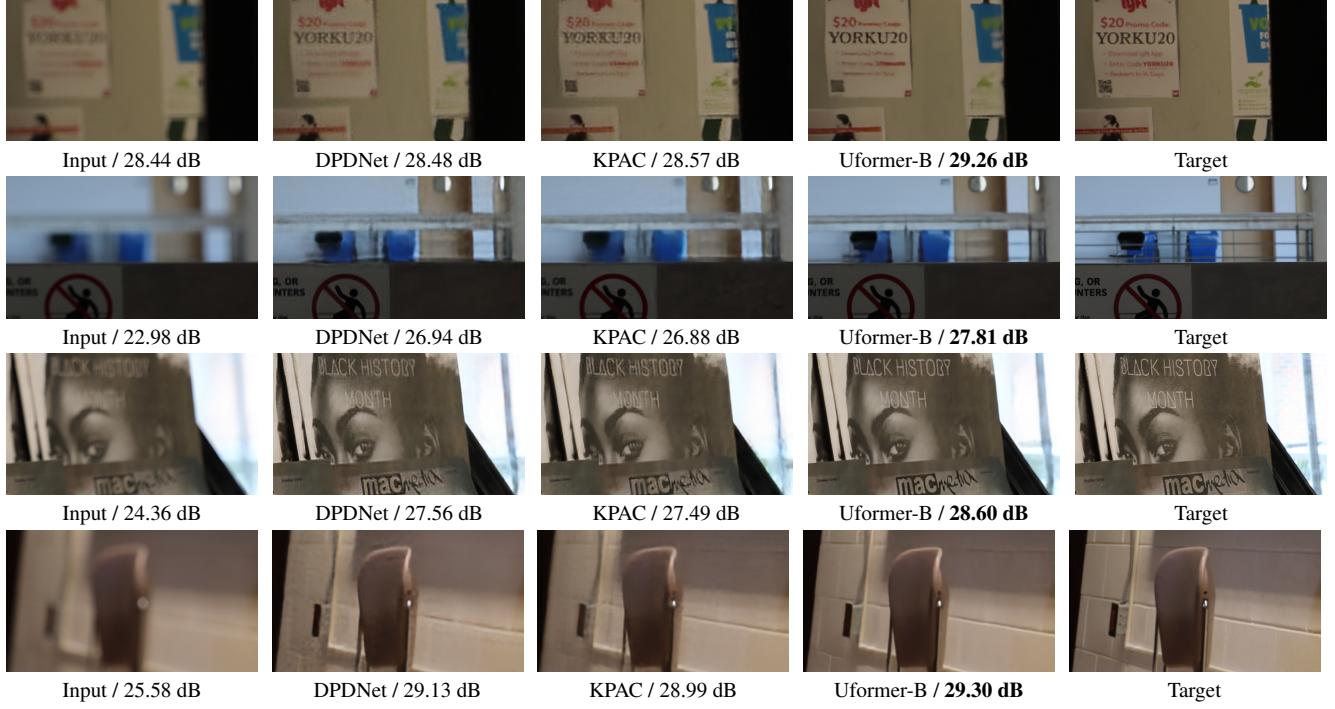


Figure 11. More results on DPD [3] for image defocus deblurring. We report the performance of PSNR on the whole test image and show the zoomed region only for visual comparison.

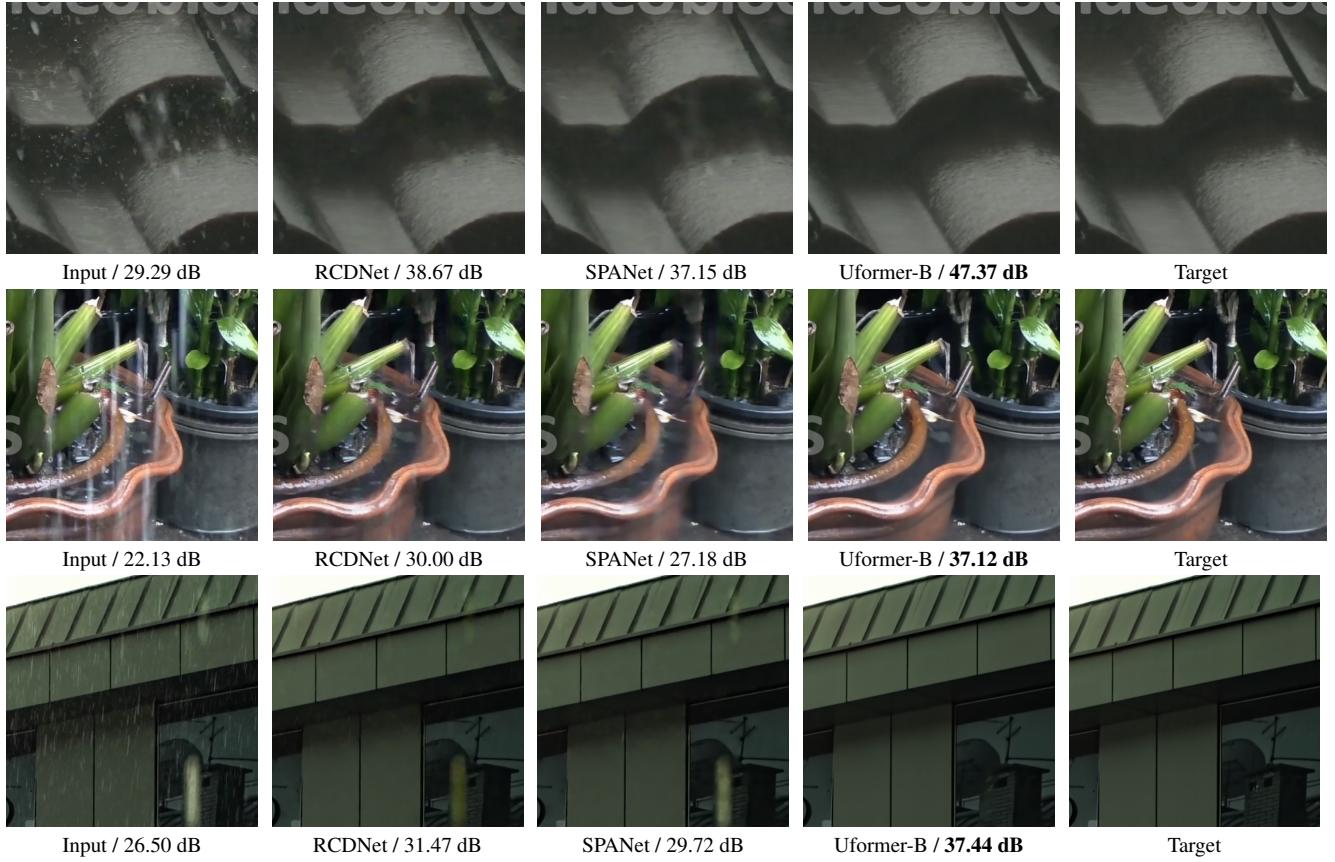


Figure 12. More results on SPAD [58] for image deraining.

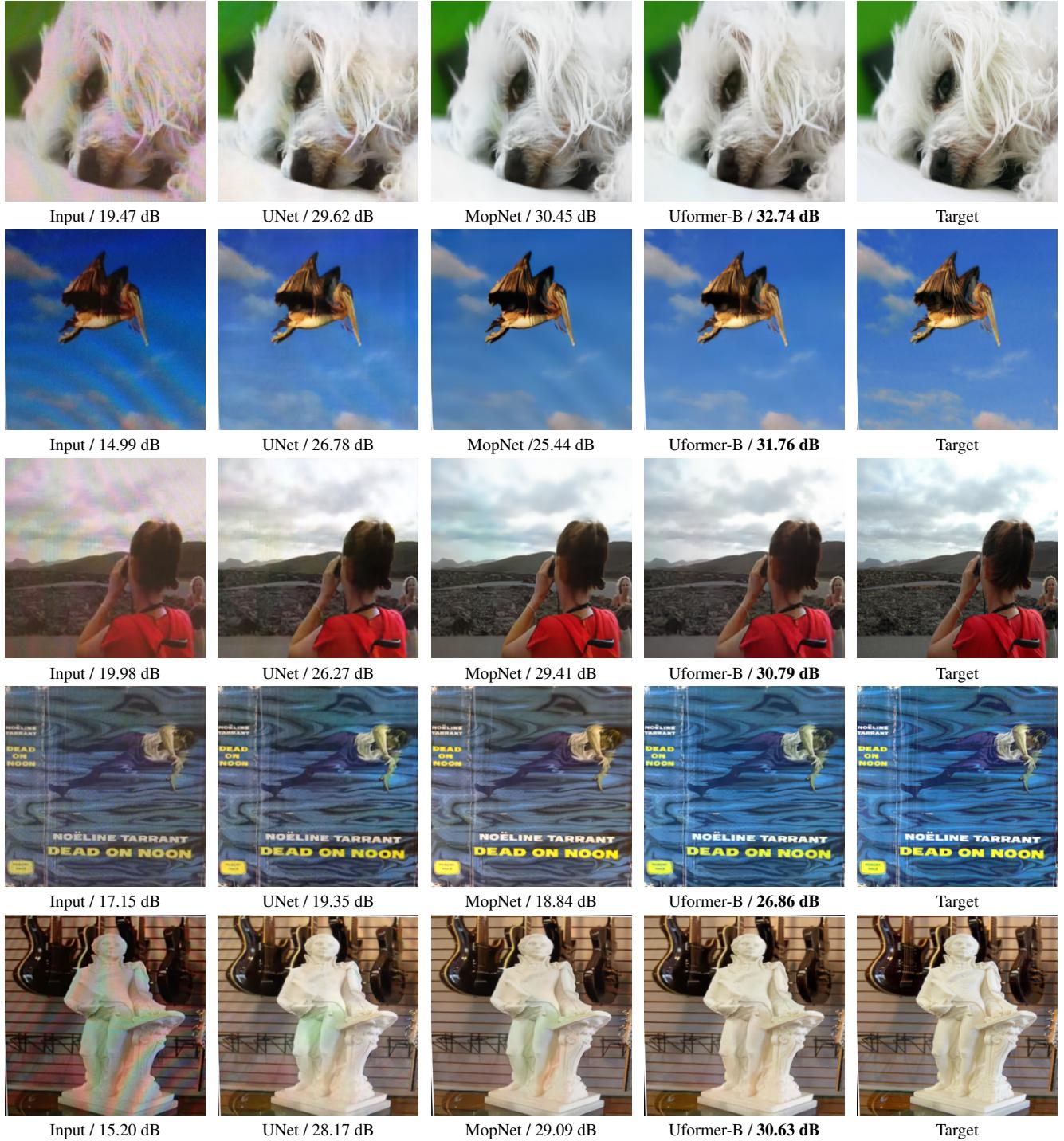


Figure 13. Results on the TIP18 dataset [58] for image demoiréing.