Motion-adaptive Separable Collaborative Filters for Blind Motion Deblurring

Chengxu Liu[1,4] Xuan Wang[2] Xiangyu Xu[1] Ruhao Tian[1]
Shuai Li[2] Xueming Qian[1,4] Ming-Hsuan Yang[3]
[1]Xi'an Jiaotong University [2]MEGVII Technology [3]University of California, Merced
[4]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd

Abstract

Eliminating image blur produced by various kinds of motion has been a challenging problem. Dominant approaches rely heavily on model capacity to remove blurring by reconstructing residual from blurry observation in feature space. These practices not only prevent the capture of spatially variable motion in the real world but also ignore the tailored handling of various motions in image space. In this paper, we propose a novel real-world deblurring filtering model called the Motion-adaptive Separable Collaborative (MISC) Filter. In particular, we use a motion estimation network to capture motion information from neighborhoods, thereby adaptively estimating spatially-variant motion flow, mask, kernels, weights, and offsets to obtain the MISC Filter. The MISC Filter first aligns the motion-induced blurring patterns to the motion middle along the predicted flow direction, and then collaboratively filters the aligned image through the predicted kernels, weights, and offsets to generate the output. This design can handle more generalized and complex motion in a spatially differentiated manner. Furthermore, we analyze the relationships between the motion estimation network and the residual reconstruction network. Extensive experiments on four widely used benchmarks demonstrate that our method provides an effective solution for real-world motion blur removal and achieves state-of-the-art performance. Code is available at https://github.com/ChengxuLiu/MISCFilter.

1. Introduction

Blind motion deblurring aims at recovering high-quality images from the counterparts of blurred images containing real-world motions. From the perspective of the imaging process, real-world motions that produce blur are diverse and spatially varying, which introduces significant challenges for the corresponding solutions.

To tackle these issues, recent years have witnessed a growing number of image deblurring approaches, which can be categorized into two paradigms. The former attempts to strengthen the model capacity and directly reconstruct the residuals used to obtain sharp image [7, 12, 26–28, 40, 57, 59, 62, 65, 66]. These methods use multi-scale learning and supervision [7, 12, 65], high efficiency attention mechanism [57, 59, 62, 66], generative adversarial learning [27, 28], and frequency domain learning [26, 33, 40] to optimize the model capability. Unfortunately, due to the shift-invariant of popularly used "black box" convolution networks, nearly all of these methods either lack interpretability to remove motion blur or the ability to handle spatially variable and large-scale motion. The latter attempts to learn the motion blur kernel or prior first, which then guides the reconstruction of residuals [2, 4, 8, 19, 49, 54]. These methods focus on designing different types of mathematical models to estimate motion prior, such as probabilistic distribution of motion blur [54], blur-field [2], latent representation of motion [8], and motion kernel [4, 19, 49].

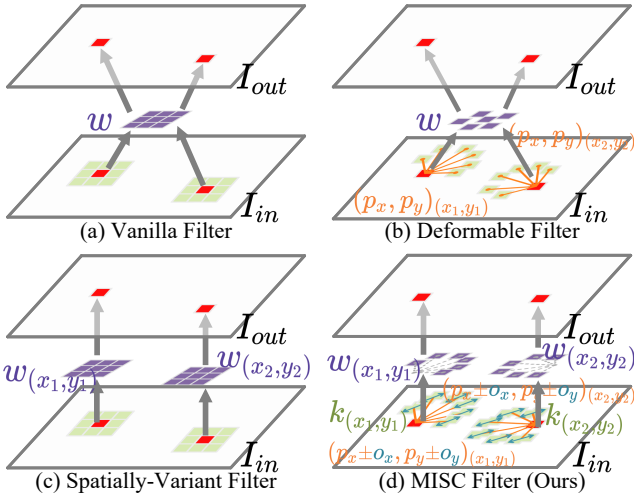

Figure 1. Illustration of other classical filters and our method. Red part is the center point of the filter in the image, and the green part is the reference point for generating it. Violet and orange parts are the weights and offsets in the filter, respectively, Blue part of our method indicates the motion-guided alignment along the motion direction (best viewed in color).

These methods focus on reconstructing the image residuals and ignore the tailored handling of motion in image space, although promising results have been achieved. Moreover, multiple factors in the real world (*e.g.*, camera shake, object large-scale movement, etc.) limit the accuracy of their predicted motion priors. Therefore, the image quality will be further improved if the complex motion in blurred images can be directly handled in the image space.

Different from reconstructing the residuals of sharp images in feature space via deep networks, reconstructing a high-quality image in image space usually involves various filtering operators [1]. From easy to hard, vanilla filters (in Fig. 1(a)) are usually hand-crafted with a fixed shape operator for the whole image, limiting the ability to capture motion at far distances and at critical pixel features. Deformable filters [16] (in Fig. 1(b)) and spatially-variant filters [51] (in Fig. 1(c)) further address these problems by adding spatial offsets and weights, respectively. We attempt to integrate the advantages of these filters to provide the ability for capturing various complex motions (in Fig. 1(d)). In addition, various filtering operators also have made significant progress in low-level vision recently, such as video artifact removal [39], video frame interpolation [10, 29], multiple degradation removal [45], and so on. In summary, these demonstrate the generalizability and great potential of filtering for low-level restoration tasks.

To address the shortcomings of focusing only on reconstructing image residuals, we study the removal of motion blur in image space. Specifically, inspired by the deformable separable convolution [16, 60, 70], we propose a novel Motion-adaptive Separable Collaborative (MISC) Filter for blind motion deblurring. As shown in Fig. 2, MISC Filter first aligns motion-induced blurring patterns to the middle along the predicted flow direction, and then collaboratively filters the alignment image to render the output. All parameters contained therein (*i.e.*, the motion field, mask, kernels, weights, and offsets) are predicted by a motion estimation network without affecting the reconstruction of residuals by existing methods. To improve the overall model efficiency, we further analyze the relationship between the motion estimation network and residual reconstruction network in an attempt to couple them.

The main contributions of this work are:

- We propose a novel real-world deblurring filtering model called the Motion-adaptive Separable Collaborative (MISC) Filter. It targets the shortcomings of existing methods that focus only on image residual reconstruction in feature space and can handle more generalized and complex motion in image space.
- We analyze the relationship between the motion estimation network for producing filter parameters and the residual reconstruction network to maximize model efficiency.
- We demonstrate the effectiveness of our method by extensive quantitative and qualitative evaluations, and provide an effective solution for blind motion deblurring.

2. Related Work

2.1. Blind Image Deblurring

Blind motion deblurring is always a popular topic in low-level vision. Recently, to handle the degradation caused by complex motion patterns, an increasing number of methods have been proposed with significant success [7, 8, 19, 26, 40, 57, 59, 62, 66]. According to whether the motion prior is required or not, these methods can be categorized into two groups, motion prior-free and motion prior-related.

Motion prior-free methods. These methods [6, 7, 12, 14, 15, 26–28, 31, 40, 44, 57, 59, 62, 65–68] focus attention on designing more robust feature learning networks that learn directly to remove various motion-induced blurring. In particular, these methods improve the model capacity from the perspectives of multi-scale learning and supervision [6, 7, 12, 44, 65, 67], high efficiency attention mechanism [14, 15, 31, 57, 59, 62, 66], generative adversarial learning [27, 28, 68], and frequency domain learning [26, 40], separately. Typical MPRNet [65] and MIMO-Unet [12] explore multi-scale and multi-stage constraint mechanisms, providing a robust framework for deblurring. The latest FFTformer [26] optimizes the matrix multiplication in the Transformer by element-wise product in the frequency domain, significantly increasing the model capacity. Although significant progress has been made, the shift-invariant of the convolution limits the ability of these methods to deal with complex motion in the real world.

Motion prior-related methods. These methods [2, 4, 8, 11, 17, 19, 20, 36, 46, 49, 54, 63] treat them as inverse problems for motion patterns. In particular, these methods first learn a spatially variable motion prior from the blurred appearance. The motion prior is then introduced into the reconstruction network to guide the network learning for motion blur removal. Typically, UFPNet [19] predicts the representations of motion blur through the Flow-based models and introduces them into residual reconstruction networks. Inspired by the recent progress of Diffusion model [21], HiDiff [8], GDP [20], and DiffIR [63] estimate the motion prior and introduce them into the denoising process. However, these methods emphasize the reconstruction of image residuals in feature space and ignore the direct handling of motions in image space. In addition, by adaptively estimating all parameters in the filter, our method is able to handle various complex motions.

2.2. Filters in Low-level Vision

Filtering algorithms have achieved widespread effect in various low-level tasks in recent years [10, 17, 29, 34, 39, 45].

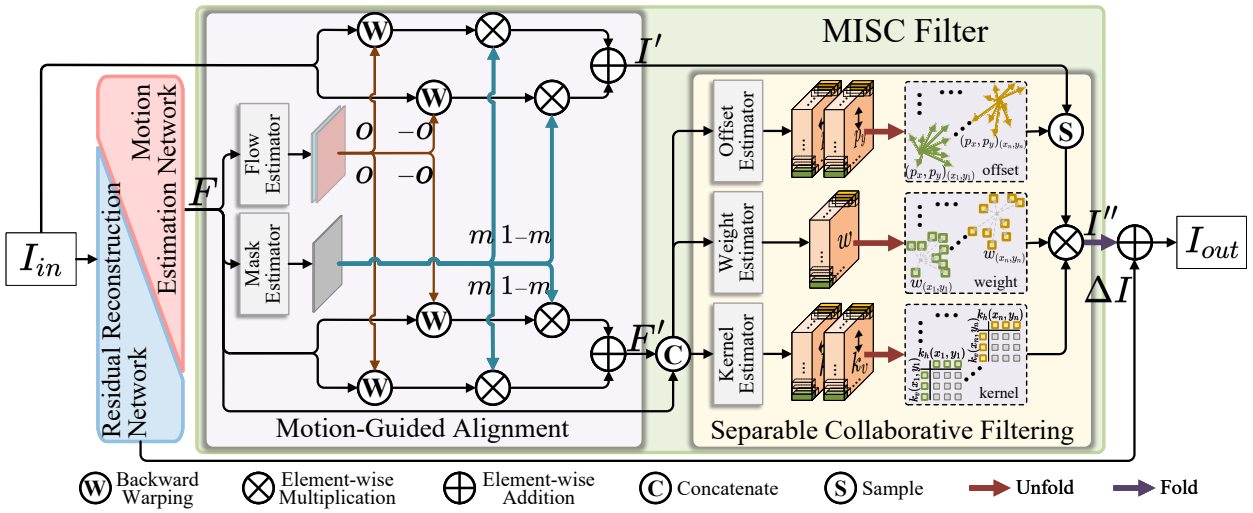

Figure 2. Overview of proposed Motion-adaptive Separable Collaborative (MISC) Filter. MISC Filter focuses on removing various motions in image space. It inputs image feature F obtained from a motion estimation network and generates filtered image I'' via a motion-guided alignment (MGA) module and a separable collaborative filtering (SCF) module.

Typically, Lu *et al*. [39] introduces the Kalman filtering process into video artifact removal and builds a recursive filtering scheme. In the video frame interpolation task, AdaCoF [29] and EDSC [10] propose an adaptive collaboration of flows and enhanced deformable separable convolution for intermediate frame synthesis by filtering operations, respectively. More recently, to deal with unknown multiple degradations in unpredictable realistic environments, ADMS [45] proposes an adaptive discriminative filter-based model and achieves promising results. In general, due to the strong ability to deal with spatially-variant degradations, filters demonstrate great potential in low-level tasks. Therefore, in this paper, we introduce a MISC Filter to improve the ability to remove motion blur in the image space.

3. Our Approach

As shown in Fig. 2, given a blurred input, our method uses a motion estimation network to obtain the parameters in the MISC Filter. The output sharp image is obtained from the combination of the filtered image by the MISC Filter and the image residuals from the reconstruction network.

In this section, we first briefly introduce the problem formulation of the blind motion deblurring in Sec. 3.1 and then elaborate on the proposed MISC Filter in Sec. 3.2. Finally, we analyze the relationship between the motion estimation and the residual reconstruction network in Sec. 3.3.

3.1. Problem Formulation

Blind motion deblurring aims to recover a sharp image from a blurred image without knowing the blur patterns. In this work, we define the blind motion deblurring task as an inverse problem for the blur kernel. Same as existing works [19, 48], we formulate a blurred image $\hat{y}$ as:

$$I_{in} = k * I_{GT} + \tilde{n}, \tag{1}$$

where I_{GT} is the sharp image. k is the blur kernel produced by motion, which usually differs in each region depending on the variety of motions. $*$ denotes the filtering operations. $\tilde{n}$ denotes the additive noise of the camera. Based on this assumption, solving 1) the inverse problem of finding k from a blurred image I_{in} and 2) filtering the image directly in the image space are essential to recovering a sharp image.

3.2. MISC Filter

Existing filters [9, 10, 16, 51] either fail to capture various spatially-variant motions or are limited by the degrees of freedom of multiple parameter choices (*e.g*., kernels, weights, offsets, etc.). Therefore, as illustrated in Fig. 2, we propose a Motion-adaptive Separable Collaborative (MISC) Filter to deal with these problems. Specifically, it first takes the image features learned from the motion estimation network as input. A motion-guided alignment (MGA) module then aligns the motion-induced blur to the motion middle along the estimated flow direction. A separable collaborative filtering (SCF) module finally uses the predicted filter parameters to filter the aligned image as output.

In terms of formula, given a blurred image I_{in}, we use a motion estimation network consisting of CNNs to obtain a feature denoted as $F \in \mathbb{R}^{C\times H\times W}$. H, W, and C represent the feature's height, width, and channel, respectively. The $\mathrm{MGA}(\cdot)$ and $\mathrm{SCF}(\cdot)$ denote the MGA module and SCF module, respectively. The aligned image $I' \in \mathbb{R}^{3\times H\times W}$ and feature $F' \in \mathbb{R}^{C\times H\times W}$ are formulated as:

$$I', F' = \mathrm{MGA}(I_{in}, F). \tag{2}$$

Then, the filtered image $I'' \in \mathbb{R}^{3\times H\times W}$ of our MISC Filter can be formulated as:

$$I'' = \mathrm{SCF}(I', F'). \tag{3}$$

The final output sharp image I_{out} is obtained by summing the filtered image I'' with the image residual $\Delta I \in \mathbb{R}^{3\times H\times W}$ obtained from the reconstruction network. Following, we detail two core modules $\mathrm{MGA}(\cdot)$ and $\mathrm{SCF}(\cdot)$.

Motion-guided alignment. Motion blur often occurs by the object displacement in a short period. To generate a sharper image, we propose the MGA module to localize the object's final position. First, a flow estimator is used to predict the motion field from motion start and end points in blurred images to the middle in latent sharp images. Then, bi-directional warping [5, 32] is utilized to align the blur to the middle moment, extending the range of handling blur and sharpening texture details. In addition, to avoid the pixel occlusion [3, 35] in different directions during warping, we incorporate a estimator to generate mask as a modulation mechanism to optimize bi-directional pixel synthesis.

Specifically, for the input feature F, we first obtain the motion flow and mask using the flow estimator $E_f(\cdot)$ and the mask estimator $E_m(\cdot)$, respectively, formulated as:

$$\begin{aligned} o &= E_f(F), \\ m &= E_m(F), \end{aligned} \tag{4}$$

where $o \in \mathbb{R}^{2\times H\times W}$ is the motion flow, indicating the offset of each pixel along the direction from the motion center to the motion end. It is obtained from a flow estimator consisting of only one convolution layer. $m \in \mathbb{R}^{1\times H\times W}$ denotes the mask to adaptively adjust the weighted summation of pixels that are aligned to the motion center from the motion start and end. Different from the flow estimator, the output of the mask estimator is fed to a sigmoid function. Such a design not only enables the automatic parameter update of the estimator during training but also lighter the motion estimation in existing works [18, 22, 53].

We then use bi-directional warping to obtain the aligned image I', and the aligned feature F', formulated as:

$$\begin{aligned} I' &= m \otimes \mathrm{W}(o, I_{in}) \oplus (1-m) \otimes \mathrm{W}(-o, I_{in}), \\ F' &= m \otimes \mathrm{W}(o, F) \oplus (1-m) \otimes \mathrm{W}(-o, F), \end{aligned} \tag{5}$$

where $\mathrm{W}(o, \cdot)$ denotes the backward warping [23] according to the motion flow o. $\otimes$ and $\oplus$ denote the element-wise multiplication and element-wise addition, respectively. Following this, the obtained aligned feature are fed to the $\mathrm{SCF}(\cdot)$ module to further estimate the parameters that are used to filter the aligned image.

Separable collaborative filtering. To alleviate the limitations of multiple degrees of freedom in filter parameter settings and find reference pixels when capturing motions, we collaboratively obtain the filter parameters by using kernel, weight, and offset estimators. The obtained parameters serve on the aligned image I' to output a sharper result I''.

Specifically, for the input feature F and aligned feature F', we separately use three estimators to predict the kernel, offset, and weight of the filter, formulated as:

$$\begin{aligned} k_v, k_h &= E_k(C(F, F')), \\ p_x, p_y &= E_o(C(F, F')), \\ w &= E_w(C(F, F')). \end{aligned} \tag{6}$$

Among them, inspired by the existing works [9, 42, 47], we approximate a 2D kernel with a pair of 1D kernels. This design encodes an $n \times n$ kernel with only $2n$ variables. $k_v \in \mathbb{R}^{n\times H\times W}$ and $k_h \in \mathbb{R}^{n\times H\times W}$ are the corresponding 1D filter kernels in the vertical and horizontal directions, respectively. n denotes the kernel size. $w \in \mathbb{R}^{n^2\times H\times W}$ is the weight used to differentially aggregate different referenced pixels. $p_x \in \mathbb{R}^{n^2\times H\times W}$ and $p_y \in \mathbb{R}^{n^2\times H\times W}$ are the offset of referenced pixels on the x-axis and y-axis, respectively. $C(\cdot)$ is the concatenation operation. $E_k(\cdot)$, $E_o(\cdot)$, and $E_w(\cdot)$ are lightweight estimators consisting of only one convolution layer with the same structure. The weight estimator additionally incorporates a softmax function to ensure that the pixel value after filtering is normalized to a value between 0 and 1. All parameters in the estimators are automatically learned during training.

Taking the pixel with coordinates (a, b) in the filtered image I'' as example, the filtering process is formulated as:

$$\begin{aligned} I''(a,b) = \sum_{(p_x(a,b), p_y(a,b))} \Big(& w(a,b) \otimes k(a,b) \\ & \otimes S\Big(I', \big(a+p_x(a,b), b+p_y(a,b)\big)\Big)\Big), \end{aligned} \tag{7}$$

where $S\Big(I', \big(a+p_x(a,b), b+p_y(a,b)\big)\Big)$ represents the spatial sampling operation from the pixel in I' with coordinates $\big(a+p_x(a,b), b+p_y(a,b)\big)$. I'' is the output filtered image. $k(a,b) = k_v(b) \cdot k_h^T(a)$ denotes the 2D kernel which can be approximated with two 1D kernels, effectively reducing the computational cost to a linear magnitude.

In summary, "Kernel" is the initial filter weight, which is generated adaptively depending on the blurred content of different regions. "Offset" is the vector direction of the motion that induces the local blur, and it locates the blurred boundary around the reconstructed pixel and captures the blur's shape. "Weight" implies a quadratic adjustment of content aggregation, increasing the model's non-linear representation. This design with collaborative parameter estimation significantly improves the ability to capture complex motions while avoiding cumbersome parameter settings.

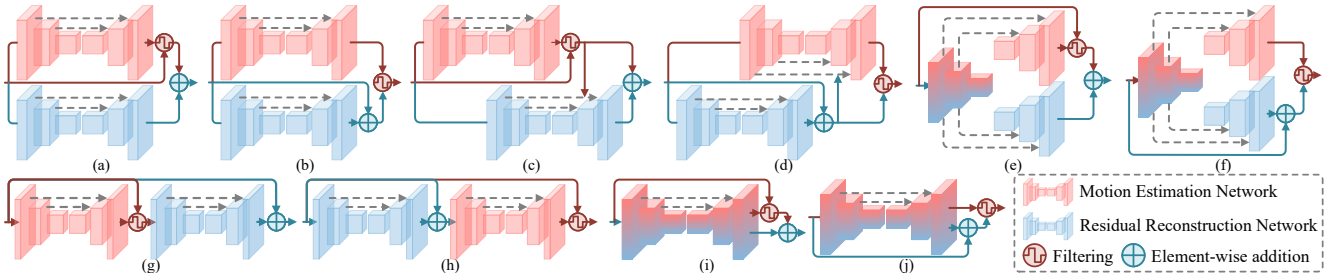

Figure 3. Design of different coupling strategies for motion estimation and residual reconstruction networks. In the ablation studies of Tab. 3, we explore different strategies to generate filters and residuals.

3.3. Network Coupling Strategies

To estimate the filter parameters, we introduce a motion estimation network similar to the existing encoder-decoder structure [12, 40] in our model. However, with a fixed model size, a large motion estimation network will improve the filtering performance but degrade the performance of the residual reconstruction, and vice versa. In addition, since filtering and residual reconstruction are completely independent, it is necessary to analyze the order of their execution. Therefore, to trade-off between the motion estimation network and the residual reconstruction network, we analyze the relationship between them with different coupling strategies in an attempt to maximize the model efficacy with minimal model size.

Specifically, we analyze them in detail regarding 1) the coupling strategies of the two networks and 2) the order of filtering and reconstruction. As shown in Fig. 3(a)-(j), we build a total of ten frameworks. Among them, we set up the following five coupling strategies:

- Parallel-based (Figs. 3(a) and (b)), where the feature learning of the two networks is completely independent.
- Semi-parallel-based (Figs. 3(c) and (d)), where the output of one is fed into the other as latent encoding.
- Serial-based (Figs. 3(g) and (h)), where the output of one network serves as an input to the other one.
- Semi-shared-based (Figs. 3(e) and (f)), where the encoders in both networks are shared.
- Shared-based (Figs. 3(i) and (j)), where both networks share the same encoder-decoder structure and parameters.

For each strategy, the order of filtering and residual reconstruction can be exchanged.

For fair comparisons, we keep them with the same size to study the performance. Among them, the highest performance is achieved when based on the shared-based strategy and filtering first (in Fig. 3(i)). This is due to the mutual facilitation of the two parts of features used for filter parameters estimation and residual reconstruction. Specific experimental results can be obtained from Sec. 4.4.

4. Experiments

4.1. Datasets and Metrics

We evaluate the our method on widely-used datasets: **RealBlur** [48], **GoPro** [41], and **HIDE** [50]. **RealBlur** [48] dataset contains two subsets: **RealBlur-R** and **RealBlur-J**. Each subset contains 3,758 image pairs for training and 980 pairs for testing. **GoPro** [41] dataset includes 2,103 pairs for training and 1,111 pairs for testing. **HIDE** [50] dataset only includes 2,025 images pairs for testing. For fair comparisons, we follow previous works [8, 26, 40] to 1) train the model with the RealBlur-R and RealBlur-J training sets and test the model with their test sets, and 2) train the model with the GoPro training set and test the model with the test set of GoPro and HIDE. Same as previous works [26, 65], we keep the same evaluation metrics:1) Peak Signal-to-Noise Ratio (PSNR) and 2) Structural Similarity Index (SSIM) [61].

4.2. Implementation Details

Similar to existing works [12, 40, 65], the proposed filters are learned with supervision at multiple scales. To strengthen the capacity of the motion estimation network and the residual reconstruction network, we use the same U-Net structure, frequency-domain learning scheme, and loss function as prior works [26, 40]. In SCF, the kernel size n is set to 7. During training, we use Cosine Annealing scheme [38] and Adam [24] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is reduced from the initial 2×10^{-4} to 1×10^{-6}. We set the batch size as 16 and the input patch size as 256×256 and augment the data with random horizontal flips, vertical flips, and 90° rotations. The total number of epochs is 6K. All experiments were based on PyTorch and trained on 2 Nvidia V100 GPUs.

4.3. Comparisons with State-of-the-art Methods

We compare our method with 15 classical start-of-the-art methods. We summarize these methods into two categories: Motion prior-free [7, 12, 13, 26, 28, 30, 40, 56, 57, 59, 62, 65, 66] and Motion prior-related [19, 58]. For fair com-

Method	#P(M)	RT(s)	RealBlur-R		RealBlur-J		GoPro		HIDE	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRN [56]	6.8	0.07	38.65	0.965	31.38	0.909	30.26	0.934	28.36	0.904
DeblurGANv2 [28]	60.9	0.04	36.44	0.934	29.69	0.870	29.55	0.934	26.61	0.875
SAPHN [52]	23	0.77	-	-	-	-	31.85	0.948	29.98	0.930
MPRNet [65]	20.1	0.09	39.31	0.972	31.76	0.922	32.66	0.958	30.96	0.939
MIMO-Unet+ [12]	16.1	0.02	-	-	31.92	0.919	32.45	0.956	29.99	0.930
MAXIM [59]	-	-	39.45	0.962	32.84	0.935	32.86	0.961	32.83	0.956
Uformer-B [62]	50.9	0.07	-	-	-	-	33.06	0.967	30.90	0.953
Restormer [66]	26.1	0.08	-	-	-	-	32.92	0.961	31.22	0.942
Restormer+local [13]	26.1	0.42	-	-	-	-	33.57	0.965	31.49	0.944
MSDI-Net [30]	135.4	-	-	-	-	-	33.28	0.964	31.02	0.940
Stripformer [57]	26.1	0.04	39.84	0.973	32.48	0.929	33.08	0.962	31.03	0.939
NAFNet [7]	67.9	0.04	-	-	-	-	33.71	0.966	31.31	0.942
DeepRFT+ [40]	23	0.09	39.84	0.972	32.19	0.930	33.23	0.963	31.42	0.944
UFPNet [19]	80.3	-	40.61	0.974	33.35	0.934	34.06	0.968	31.74	0.947
FFTformer [26]	16.6	0.13	40.11	0.973	32.62	0.932	34.21	0.969	31.62	0.945
MISC Filter (Ours)	16.0	0.07	41.23	0.978	33.88	0.938	34.10	0.969	31.66	0.946

Table 1. Quantitative comparison on the RealBlur-R [48], RealBlur-J [48], GoPro [41], and HIDE [50] dataset. We follow existing works [26, 40] using a model trained on the RealBlur-R and RealBlur-J training sets for testing on the RealBlur-R and RealBlur-J test sets, respectively, and using a model trained on the GoPro training set for testing on the GoPro and HIDE test sets. #P(M) and RT(s) indicate the parameter and runtime, respectively. Runtime is computed on images with the size of 256×256 with an Nvidia RTX 3090 GPU. Red and blue indicate the best and the second best performance, respectively (best viewed in color).

parisons, we report the results from their original papers or reproduce the results through the published models.

Quantitative comparison. The performance comparison of our method with other SOTA methods is shown in Tab. 1. Compared to the latest motion prior-free method FFTformer [26], our method achieves significant performance gains on the more complex real-world motion blur datasets RealBlur-R [48] and RealBlur-J [48]. While our method achieves comparable performance on GoPro [41] and HIDE [50], it only spends half the runtime. It proves the robustness of our method to handle complex real-world motions. Compared to the latest motion prior-related method UFPNet [19], our method achieves higher performance on the GoPro [41], RealBlur-R [48], and RealBlur-J [48] with less number of parameters and runtime. It proves the superiority of our method in handling various motions and verifies the generalization ability on various motion blurs.

Qualitative comparison. To further compare the visual quality of different methods, we present their de-blurring results in Fig. 4. For fair comparisons, we directly use author-released models to get results for fairness. The results of our method have better visual quality, especially when complex motion is involved. For example, in the first case of Fig. 4, our method recovers the clear text with complex movement. In the second case, our method recovers the texture details of a fast-moving car. More visual results can refer to the supplementary materials.

4.4. Ablation study

In this section, we conduct the ablation on the proposed MISC Filter and analyze the network coupling strategies. In addition, we further analyze the effect of the different filtering methods and kernel sizes. All experiments are performed on our lightweight version model on GoPro [41].

MISC Filter. To demonstrate the superiority of each component in our MISC Filter, we analyze their effect on performance in Tab. 2 and Fig. 5. We try to keep the same model size in each comparison for fairness. We directly remove the MISC Filter as the "Base" model and progressively add each component for comparisons. The addition of MGA improves PSNR (from 32.40 dB to 32.51 dB) and visual quality, verifying its ability to eliminate motion blur through bi-directional alignment. Adding the kernel, weight, and offset estimators separately in the SCF can all lead to a 0.1 dB improvement, demonstrating the superiority of each component in our MISC Filter. With all of them added, the model performance improves from 32.51dB to 32.83dB, and the visual quality is further improved. It proves that our method alleviates the limitations of multiple degrees of freedom in filter parameters setting and enhances the ability to handle various complex motions.

To demonstrate the model ability of the filtering process to capture complex motions, we further visualize the aligned image following MGA, motion flow, mask, weight, and offset in the MISC Filter. As shown in Fig. 6, the mo-

Figure 4. Visual results on **Realblur-J** [48], **GoPro** [41], and **HIDE** [50] datasets. The method is shown at the bottom of each case. Zoom in to see better visualization.

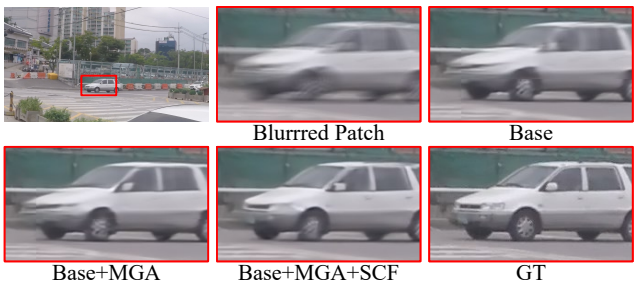

Figure 5. Visualization of ablation studies on MISC Filter.

Base	MGA	SCF			PSNR	SSIM
		kernel	weight	offset		
✓					32.40	0.955
✓	✓				32.51	0.956
✓	✓	✓			32.55	0.957
✓	✓	✓	✓		32.67	0.957
✓	✓	✓		✓	32.71	0.958
✓		✓	✓	✓	32.67	0.958
✓	✓	✓	✓	✓	**32.83**	**0.960**

Table 2. Results of ablation studies on MISC Filter. MGA: motion-guided alignment. SCF: separable collaborative filtering. our MISC Filter can be interpreted as "Base+MGA+SCF".

tion flow and mask can effectively capture the moving car, and the aligned image following MGA removes the blurring induced by fast motion. In addition, we chose pixels of the wheels and back of the car (indicated by red dots) to visualize the weights and offsets, where stronger signals indicate higher weights. It demonstrates the strong ability of the filter to capture non-local spatial motions.

Network coupling strategies. To analyze the relationship between the motion estimation network and the residual reconstruction network in Sec. 3.3, we compare all ten frameworks presented in Fig. 3. In Tab. 3, the highest performance is achieved when the two networks share the same structure and when filtering is performed before adding the residuals (Group (i) in Fig. 3). This is because sharing the same network would facilitate interactive learning of the two-part features. Moreover, prioritizing filtering blurred images in image space is more straightforward for resolving motion blur. Therefore, we use the framework (i) to

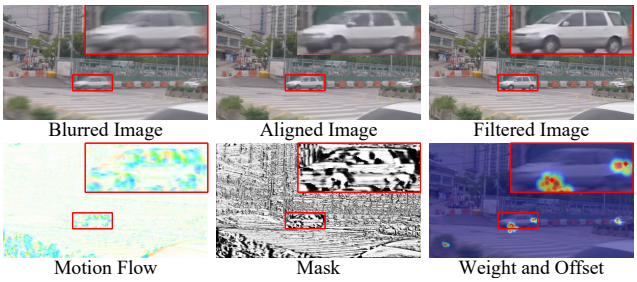

Figure 6. Visualization of the aligned image following MGA, motion flow, mask, weight, and offset in MISC Filter.

Group	a	b	c	d	e
PSNR	32.53	32.42	32.56	32.38	32.63
SSIM	0.958	0.956	0.958	0.956	0.959
Group	f	g	h	i	j
PSNR	32.59	32.55	32.46	**32.83**	32.39
SSIM	0.958	0.957	0.957	**0.960**	0.956

Table 3. Results of ablation studies on frameworks in Fig. 3.

Method	PSNR	SSIM
Vanilla Filter	32.21	0.952
Deformable Filter [16]	32.48	0.956
Separable Filter [51]	32.51	0.957
Deformable Separable Filter [10]	32.54	0.957
MISC Filter (Ours)	**32.83**	**0.960**

Table 4. Results of ablation studies on different filtering methods.

maximize model efficacy with minimal model size.

Filtering. To demonstrate the reliability of our filtering process, we compare several classical filtering methods [10, 16, 51] in Tab. 4. Our MISC Filter performs favorably against the two representative filters, deformable filter [16] and separable filter [51]. It is because they either cannot adaptively change the weights of different regions or are not conducive to capturing non-local motion. In addition, due to the design of its motion-guided alignment module and collaborative kernel estimation, our method has higher performance than deformable separable filter [10]. Our method allows for handling more complex motions.

Kernel size. We analyze the effect of kernel size n on describing complex motions for the MISC Filter. As shown in Tab. 5, the performance positively correlates with kernel size. It demonstrates the powerful potential of our MISC Filter to handle complex motions. However, the performance gain gradually decreases when the n exceeds 7. Besides, although a larger kernel will allow more pixels to be referenced, it will also increase the number of parameters.

n	3	5	7	9	11
PSNR	32.60	32.76	32.83	32.84	32.87
SSIM	0.957	0.958	0.960	0.960	0.961

Table 5. Results of ablation studies on kernel size n.

Method	RT(s)	#P(M)	PSNR	SSIM
MSUNet [69]	0.08	8.9	37.40	0.9756
DAGF [55]	1.12	1.1	36.49	0.9716
PDCRN [43]	0.08	4.7	37.83	0.9780
ResUNet [64]	0.43	16.5	37.95	0.9790
RDUNet [64]	0.53	47.9	38.13	0.9797
UDC-UNet [37]	0.30	5.7	38.10	0.9796
BNUDC [25]	0.08	4.6	38.22	0.9798
FSI [33]	0.07	3.8	38.60	0.9805
MISC Filter (Ours)	0.05	5.6	38.42	0.9800

Table 6. Results of hardware-induced blur removal.

4.5. Evaluation on hardware-induced blurring

Aside from motion-induced blurring, in the under-display camera (UDC) systems, the pixel array of light-emitting diodes used for display diffracts and attenuates the incident light, resulting in blurring [25]. Therefore, we construct experiments on widely-used UDC benchmark T-OLED [69] in order to explore the ability of our method to remove hardware-induced blurring. As shown in Tab. 6, our method achieves comparable performance compared to most models specifically designed for this task. However, our method still falls short since it cannot handle the low-light problem caused by the attenuation of incident light in this task.

5. Conclusion

In this paper, we introduce a new perspective to handle motion blur in image space instead of features and propose a novel motion-adaptive separable collaborative (MISC) filter. In the MISC Filter, we design a motion-guided alignment module and a separable collaborative filtering module to eliminate motion-induced blur. To achieve this, we introduce an additional motion estimation network to predict the filter parameters and further analyze the coupling strategies between it and the conventional residual reconstruction network. Such a design eliminates blurring induced by complex motions while avoiding cumbersome filter parameter settings. Experimental results show its advantages and the potential to remove blur from other scenarios.

Acknowledgement. This work was supported in part by the NSFC under Grant 62272380 and 62103317, the Fundamental Research Funds for the Central Universities, China (xzy022023051), the Innovative Leading Talents Scholarship of Xi'an Jiaotong University, and MEGVII.

References

[1] Suyash P Awate and Ross T Whitaker. Unsupervised, information-theoretic, adaptive image filtering for image restoration. *IEEE TPAMI*, 28(3):364–376, 2006. 2

[2] Yuval Bahat, Netalee Efrat, and Michal Irani. Non-uniform blind deblurring by reblurring. In *ICCV*, pages 3286–3294, 2017. 1, 2

[3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *CVPR*, pages 3703–3712, 2019. 4

[4] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *ECCV*, pages 221–235. Springer, 2016. 1, 2

[5] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*, pages 4947–4956, 2021. 4

[6] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *CVPRW*, pages 182–192, 2021. 2

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33. Springer, 2022. 1, 2, 5, 6

[8] Zheng Chen, Yulun Zhang, Ding Liu, Jinjin Gu, Linghe Kong, Xin Yuan, et al. Hierarchical integration diffusion model for realistic image deblurring. *NeurIPS*, 36, 2024. 1, 2, 5

[9] Xianhang Cheng and Zhenzhong Chen. Video frame interpolation via deformable separable convolution. In *AAAI*, pages 10607–10614, 2020. 3, 4

[10] Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE TPAMI*, 44(10):7029–7045, 2021. 2, 3, 8

[11] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *CVPR*, pages 9137–9146, 2021. 2

[12] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, pages 4641–4650, 2021. 1, 2, 5, 6

[13] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *ECCV*, pages 53–71. Springer, 2022. 5, 6

[14] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Focal network for image restoration. In *ICCV*, pages 13001–13011, 2023. 2

[15] Yuning Cui, Yi Tao, Wenqi Ren, and Alois Knoll. Dual-domain attention for image deblurring. In *AAAI*, pages 479–487, 2023. 2

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2, 3, 8

[17] Aram Danielyan, Vladimir Katkovnik, and Karen Egiazarian. Bm3d frames and variational image deblurring. *IEEE TIP*, 21(4):1715–1728, 2011. 2

[18] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, pages 2758–2766, 2015. 4

[19] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *CVPR*, pages 18105–18114, 2023. 1, 2, 3, 5, 6

[20] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *CVPR*, pages 9935–9946, 2023. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2

[22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, pages 2462–2470, 2017. 4

[23] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 28, 2015. 4

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. BNUDC: A two-branched deep neural network for restoring images from under-display cameras. In *CVPR*, pages 1950–1959, 2022. 8

[26] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *CVPR*, pages 5886–5895, 2023. 1, 2, 5, 6

[27] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, pages 8183–8192, 2018. 1, 2

[28] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, pages 8878–8887, 2019. 1, 2, 5, 6

[29] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. Adacof: Adaptive collaboration of flows for video frame interpolation. In *CVPR*, pages 5316–5325, 2020. 2, 3

[30] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *ECCV*, pages 736–753. Springer, 2022. 5, 6

[31] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *CVPR*, pages 18278–18289, 2023. 2

[32] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, pages 5687–5696, 2022. 4

[33] Chengxu Liu, Xuan Wang, Shuai Li, Yuzhi Wang, and Xueming Qian. FSI: Frequency and spatial interactive learning for image restoration in under-display cameras. In *ICCV*, pages 12537–12546, 2023. 1, 8

[34] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. 4D LUT: learnable context-aware 4d lookup table for image enhancement. *IEEE TIP*, 32:4742–4756, 2023. 2

[35] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. TTVFI: Learning trajectory-aware transformer for video frame interpolation. *IEEE TIP*, 32:4728–4741, 2023. 4

[36] Chengxu Liu, Xuan Wang, Yuanting Fan Fan, Shuai Li, and Xueming Qian. Decoupling degradations with recurrent network for video restoration in under-display camera. In *AAAI*, 2024. 2

[37] Xina Liu, Jinfan Hu, Xiangyu Chen, and Chao Dong. Udc-unet: Under-display camera image restoration via u-shape dynamic network. In *ECCVW*, pages 113–129. Springer, 2022. 8

[38] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[39] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Zhiyong Gao, and Ming-Ting Sun. Deep kalman filtering network for video compression artifact reduction. In *ECCV*, pages 568–584, 2018. 2, 3

[40] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *AAAI*, pages 1905–1913, 2023. 1, 2, 5, 6

[41] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 3883–3891, 2017. 5, 6, 7

[42] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, pages 261–270, 2017. 4

[43] Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Melvin Kuriakose, and Jiji Charangatt Victor. Transform domain pyramidal dilated convolution networks for restoration of under display camera images. In *ECCVW*, pages 364–378. Springer, 2020. 8

[44] 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 *ECCV*, pages 327–343. Springer, 2020. 2

[45] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *CVPR*, pages 5815–5824. IEEE, 2023. 2, 3

[46] Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *ICCV*, pages 2309–2319, 2021. 2

[47] Roberto Rigamonti, Amos Sironi, Vincent Lepetit, and Pascal Fua. Learning separable filters. In *CVPR*, pages 2754–2761, 2013. 4

[48] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201. Springer, 2020. 3, 5, 6, 7

[49] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *IEEE TPAMI*, 38 (7):1439–1451, 2015. 1, 2

[50] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, pages 5572–5581, 2019. 5, 6, 7

[51] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, pages 11166–11175, 2019. 2, 3, 8

[52] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, pages 3606–3615, 2020. 6

[53] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 4

[54] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, pages 769–777, 2015. 1, 2

[55] Varun Sundar, Sumanth Hegde, Divya Kothandaraman, and Kaushik Mitra. Deep atrous guided filter for image restoration in under display cameras. In *ECCVW*, pages 379–397. Springer, 2020. 8

[56] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 5, 6

[57] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, pages 146–162. Springer, 2022. 1, 2, 5, 6

[58] Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Yen-Yu Lin, and Chia-Wen Lin. BANet: a blur-aware attention network for dynamic scene deblurring. *IEEE TIP*, 31:6789–6799, 2022. 5

[59] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MAXIM: Multi-axis mlp for image processing. In *CVPR*, pages 5769–5780, 2022. 1, 2, 5, 6

[60] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, pages 0–0, 2019. 2

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[62] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 1, 2, 5, 6

[63] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *ICCV*, pages 13095–13105, 2023. 2

[64] Qirui Yang, Yihao Liu, Jigang Tang, and Tao Ku. Residual and dense unet for under-display camera restoration. In *ECCVW*, pages 398–408. Springer, 2021. 8

[65] 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*, pages 14821–14831, 2021. 1, 2, 5, 6

[66] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 1, 2, 5, 6

[67] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 2

[68] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2737–2746, 2020. 2

[69] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In *CVPR*, pages 9179–9188, 2021. 8

[70] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, pages 9308–9316, 2019. 2