

High-Pass Kernel Prediction for Efficient Video Deblurring

Bo Ji Angela Yao
 National University of Singapore
 {jibo, ayao}@comp.nus.edu.sg

Abstract

State-of-the-art video deblurring methods use deep network architectures to recover sharpened video frames. Blurring especially degrades high-frequency (HF) information, yet this aspect is often overlooked by recent models that focus more on enhancing architectural design. Recovering these fine details is challenging, partly due to the spectral bias of neural networks, which are inclined towards learning low-frequency functions. To address this, we enforce explicit network structures to capture the fine details and edges. We dynamically predict adaptive high-pass kernels from a linear combination of high-pass basis kernels to extract high-frequency features. This strategy is highly efficient, resulting in low-memory footprints for training and fast run times for inference, all while achieving state-of-the-art when compared to low-budget models. The code is available at <https://github.com/jibo27/AHFNet>.

1. Introduction

Video deblurring sharpens video frames from blurry input sequences. The blurring process suppresses the high frequencies in the original video, so deblurring can be viewed as restoring the lost frequencies. Classical sharpening techniques enhance the edges and detailing using unsharp masking and inverse filtering [7, 50]. Others [17, 28, 48] try to estimate the underlying blur kernels to reverse the effects. In the case of multi-image deblurring [2, 51], the missing information can be approximated by observations over multiple frames. Classical methods are often interpretable and based on well-understood mathematical principles. However, these classical methods are usually based on degradation models that are too simple for real-world scenes.

Recent advances in video deblurring rely on deep neural networks trained end-to-end, achieving state-of-the-art performance through local alignment [22, 27, 44] and deformable convolutions [15, 43]. While effective, these techniques are computationally expensive and non-ideal for optimized hardware. Specifically, the memory footprint of

such models has exploded, requiring over 200GB of GPU memory [18, 20, 22, 30] to process frames of size 256×256 in a batch. This surge in the training budget not only incurs high financial costs but also leads to significant environmental impact [1]. A larger training budget often indicates a complicated model, which requires more inference time. Apart from the efficiency issues, deep learning-based models [18, 20] have made mostly black-box gains compared to classical deblurring methods. They are relatively agnostic to the underlying deblurring problem, even though deblurring is a well-studied problem in image processing [8, 50]. The interpretability of the functionality of the network modules remains obscure.

Moreover, standard neural network architecture components are counterproductive for deblurring as they tend to suppress high-frequency features, which are crucial for video deblurring. For example, self-attention in transformers acts as a low-pass filter [31, 42]; convolutional networks also tend to weaken the high-frequency components [40]. Finally, neural networks are prone to spectral bias - a preference for learning lower-frequency components [33].

To address these issues, we propose an Adaptive High-Frequency extraction network (AHFNet) for video deblurring. Our method maintains interpretability by explicitly leveraging high frequencies extracted from blurry videos for deblurring. To ensure effective extraction, we define a set of high-pass basis kernels. A high-pass kernel is a kernel that emphasizes features like edges and local details. Then, we dynamically predict mixing coefficients to perform linear combinations on the basis kernels, resulting a new high-pass kernel. The new generated kernels are further convolved with the blurry features to produce crucial details required for deblurring. Our experiments demonstrate that incorporating even the simplest high-pass kernels, such as Sobel filters, can significantly improve deblurring performance. For instance, integrating spatial and temporal gradients into a neural network for video deblurring results in a notable improvement of 0.33 dB (see Section 4.4).

Moreover, our strategy is lightweight and efficient in both training and inference, as our building blocks require only a few convolutional layers. The memory foot-

print for training our model is only one-sixth of that required for state-of-the-art models (see Fig. 3). Compared to other models with similar low-budget training requirements, our method achieves competitive performance on standard deblurring benchmarks [25, 38]. For instance, our model achieves an inference speedup of 35x compared to ERDN [15] while achieving superior PSNR and SSIM scores. (See Fig. 4 and Table 1). Our contributions can be summarized as follows:

- We highlight the critical role of high frequencies in video deblurring and reformulate the deblurring task based on explicitly extracting and applying high-frequency information.
- We introduce a new high-frequency extraction module to adaptively recover spatial and temporal detailing for video deblurring. This module forms our proposed AHFNet for efficient yet effective video deblurring.
- Our proposed AHFNet delivers top accuracy performance among models under similar memory footprints while being more efficient in inference time and GMACs.

2. Related Work

Classic deblurring. Classic deblurring algorithms often leverage image structures or priors, *e.g.*, sparse gradient priors [17, 48], intensity priors [28], and edges [6, 39, 47–49]. Unsharp masking [7, 32, 50] counteracts blurring in an image by reintroducing a scaled version of the image’s high-frequency components. This emphasizes the fine details or edges and enhances the image’s overall sharpness. Other methods iteratively estimate the blur kernel and sharpened outputs [6, 29, 36, 53]. Multi-image blind deconvolution algorithms restore the target image from multiple blurred or relevant images [2, 10, 35, 51, 57]. Classic deblurring algorithms have difficulty generalizing well to diverse and unseen types of blur. Our work explicitly utilizes high-frequency information with the substantial capacity of neural networks to enhance video deblurring.

Deep learning methods. Deep learning methods restore sharp videos with deep neural networks [13, 27, 38, 43, 55]. XYDeblur [13] introduces an one-encoder-two-decoder by rotating decoders to predict complementary deblurring information. Though both XYDeblur and our approach use rotation, XYDeblur rotates decoder parameters for image deblurring; we rotate fixed kernels to enhance frequency for video deblurring. Recent works have focused on information retrieval from neighboring frames with optical flow [22, 44], deformable convolution [15, 43], and feature matching [12, 19]. RVRT [20] combines optical flow, deformable convolution, and attention mechanisms to achieve high restoration quality. ShiftNet [18] uses spatial-temporal

shifts to enhance video aggregation. Yet these operations are computationally expensive and not well-suited for standard hardware. For example, deformable convolution requires increased memory access, irregular data structuring, and dynamic computational graphs, all adding to the complexity and resource demands [9]. Rather than focus on the architecture design, our work takes a distinct direction by placing emphasis on the significance of high frequencies to improve deblurring outcomes.

Kernel prediction networks. Kernel prediction networks determine convolution weights dynamically during inference [5, 14, 24, 56]. The kernel prediction network [46, 54] synthesizes kernel bases and coefficients to derive adaptive filters. The optimization of kernel prediction modules is challenging and requires extensive training data and extended training time to accurately predict a kernel. Our method incorporates specific basis kernels designed to emphasize the extraction of high-frequency components. We solely predict coefficients for their linear combination, which simplifies the computational load.

3. Approach

3.1. Preliminaries

The video blurring process can be viewed as a low-pass filtering on consecutive latent sharp frames [25, 38]. We follow the video blur accumulation model from [25], which forms a blurry frame, x_t , by accumulating sampled sharp frames $s_{t,i}$:

$$x_t \approx g \left(\frac{1}{B} \sum_{i=1}^B s_{t,i} \right), \quad (1)$$

where g is the nonlinear camera response function, B is the number of accumulated frames and $s_{t,i}$ is the original captured frame related to x_t . In the simplest setting where the camera response function is assumed to be an identity function, Eq 1 simplifies to a convolution operation of consecutive images with an averaging kernel. It is well-established that averaging has the effect of low-pass filtering [11]. Consequently, a blurry frame can be regarded as the low-pass filtered outcome of the latent sharp frames. Therefore, our main goal is to recover the lost high frequencies.

3.2. Formulation

Consider a local temporal window of span l around the t -th blurry frame x_t in a video, denoted as $\{x_i\}_{i=t-l}^{t+l} = \{x_{t-l}, \dots, x_t, \dots, x_{t+l}\}$. For a residual network, the deblurred frame $\hat{y}_t$ can be expressed as:

$$\hat{y}_t = x_t + \mathcal{L}(x_t, \mathcal{H}(\{x_i\}_{i=t-l}^{t+l})), \quad (2)$$

i.e. the sum between the original input x_t and some residual component extracted by $\mathcal{L}$. Specifically, the residual component depends on both the current input x_t as well

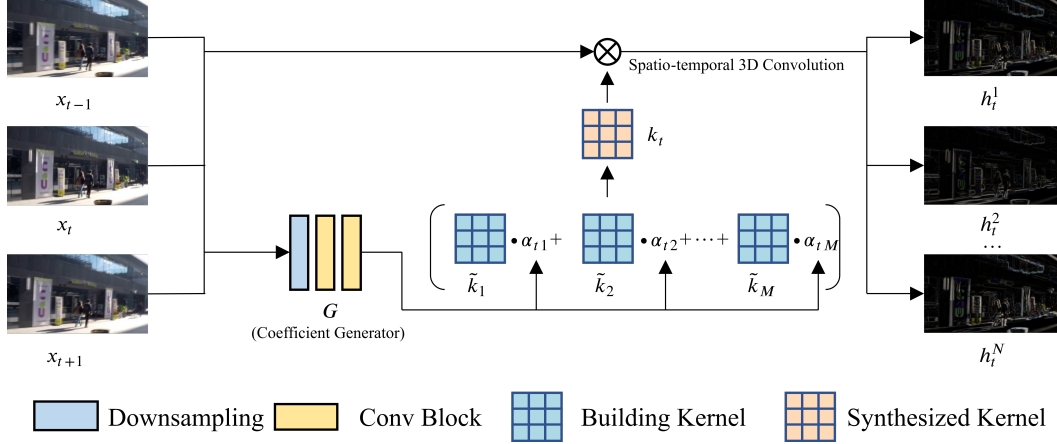


Figure 1. Adaptive high-frequency extraction module ($\mathcal{H}$). We generate the dynamic high-pass kernel k_t for future convolution by performing a linear combination of high-pass basis kernels $\{\tilde{k}_j\}_{j=1}^M$ and the coefficients $\{\alpha_{t,j}\}_{j=1}^M$.

as an operation $\mathcal{H}$ applied to a temporal window of data $\{x_i\}_{i=t-l}^{t+l}$, since blurring often involves inter-frame effects (see Eq. 1).

In the simplest setting when $\mathcal{H}$ is set as a simple high-pass filter and $\mathcal{L}$ operates as scalar multiplication independent of the input, Eq 2 reduces to the classic unsharp masking used for image sharpening:

$$\hat{y}_t = x_t + \lambda(k * x), \quad (3)$$

where $*$ denotes a 2D convolution, λ is a scaling factor and k can be any high-pass filter. Eq 3 shows that unsharp masking explicitly applies high-pass filtering to blurry inputs and adds back the filtered result to get the deblurred output.

Our general formulation in Eq 2 is more flexible and expressive than the standard unsharp masking because we incorporate inter-frame analysis and use a convolutional network to predict the features that play a similar role to $(k * x)$ and λ . Based on our formulation, we break down the deblurring process into two primary phases: 1) extracting high frequencies from both the current and adjacent frames with $\mathcal{H}$; and 2) converting these features obtained from blurry frames back with $\mathcal{L}$, effectively creating a feature map for reconstructing the latent sharp frame.

3.3. Adaptive high-frequency extraction module

Both transformers and convolutional networks have challenges in extracting high frequencies. The self-attention mechanisms in transformers act as low-pass filters and attenuate high-frequency details in the input features [31, 42]. Specifically, the low-pass nature of self-attention comes from the softmax operation that applies a non-negative, weighted averaging over the inputs. As such, previous works [31] have found that transformers have stronger low-pass effects than convolutional operations. Yet from a statistical perspective, feed-forward convolutional networks are also likely to weaken high-frequency components [40], as

the Fourier coefficients mostly locate in low frequencies after the backward propagation for the convolutional weights initialized from a Gaussian distribution.

To mitigate this, we use convolutional networks while constraining the learning space of the convolutional kernels. Our goal is to ensure that the generated features represent the high-frequency components of the input. The strategy is to make the operations similar to high-pass filtering. Specifically, we propose a module $\mathcal{H}$, consisting of kernels which function only as high-pass filters (see Figure 1). Before delving into the details of $\mathcal{H}$, we establish the following proposition on the linear combination of high-pass filters:

Proposition 1 Consider M spatial high-pass filters with impulse responses $h_i(x)$ and corresponding frequency responses $H_i(f)$ with cutoff frequencies f_{ci} , sorted such that $f_{c1} \leq f_{c2} \leq \dots \leq f_{cM}$. Any linear combination of these filters with non-negative coefficients α_i in the spatial domain, represented as

$$h(x) = \sum_{i=1}^M \alpha_i h_i(x), \quad (4)$$

will also act as a high-pass filter in the spatial domain, with a corresponding frequency response $H(f)$ and a cutoff frequency not greater than f_{c1} .

Proposition 1 states that any non-negative linear combination of high-pass filters remains a high-pass filter. The proof is given in the supplementary material. This proposition motivates us to train a kernel prediction module in which the kernel weights are formed by a linear combination of predetermined high-pass kernels and dynamically predicted coefficients. Without imposing this high-pass constraint, arbitrarily learned convolutional kernels may not consistently optimize for high-pass filtering. We verify this empirically in Section 4.3.

Now consider a set of 3D high-pass basis kernels $\{\tilde{k}_j\}_{j=1}^M = \{\tilde{k}_1, \dots, \tilde{k}_M\}$. Each kernel is of size $\mathbb{R}^{T_k \times H_k \times W_k}$ where T_k , H_k , and W_k represent the temporal length, height, and width of the kernel, respectively. The input has T_k consecutive frames of size $C \times T_k \times H \times W$, where C is the input dimension. First, we generate a dynamic kernel k_t from mixing coefficients α_t :

$$k_t = \sum_{j=1}^M \alpha_{t,j} \tilde{k}_j, \quad \text{where } \{\alpha_{t,j}\}_{j=1}^M = \mathcal{G}(\{x_i\}_{i=t-l}^{t+l}), \quad (5)$$

where $\mathcal{G}$ denotes the coefficient generator and $\alpha_{t,j}$ denotes the coefficient for the j -th kernel of x_t . The kernel k_t has a size of $1 \times T_k \times H_k \times W_k$. Proposition 1 assures that k_t remains a high-pass kernel.

Convolving image x_t in space-time with k_t leads to the high-frequency feature h_t . To cover orthogonal directions, we rotate the kernel set $\{\tilde{k}_i\}_{i=1}^M$ by 90 degrees anticlockwise, forming a new set of orthogonal basis kernels, which we denote as $\{\tilde{k}'_i\}_{i=1}^M$. Next, we apply the coefficients $\{\alpha_{t,j}\}_{j=1}^M$ to the set of rotated basis kernel, akin to the method in Eq 5, resulting in another feature, denoted as h'_t . This process can be summarized as

$$h_t = k_t * \{x_i\}_{i=t-l}^{t+l}; \quad h'_t = k'_t * \{x_i\}_{i=t-l}^{t+l}. \quad (6)$$

For clarity, $\tilde{k}_j$ and $\tilde{k}'_j$ represent the original and rotated basis kernels, respectively. k_t and k'_t denote the dynamically predicted corresponding kernels that are further convolved with the input feature x_t .

While we can learn distinct coefficients for the rotated basis kernel $\{\tilde{k}'_i\}_{i=1}^M$, applying the same coefficients achieves comparable performance. Finally, we combine the features extracted in orthogonal directions by summing the absolute values of the feature maps h_t and h'_t . The output can be expressed as:

$$\mathcal{H}(\{x_t\}_{i=t-l}^{t+l}) = \{h_t, h'_t, |h_t| + |h'_t|\}. \quad (7)$$

For brevity, we use $\mathcal{H}(x_t)$ to represent $\mathcal{H}(\{x_t\}_{i=t-l}^{t+l})$. Incorporating rotated kernels enhances diversity in the extracted frequencies while introducing the sum of absolute values facilitates a more comprehensive information aggregation from the two directions.

Predicting the mixing coefficients α leads to the dynamic generation of convolutional weights tailored to match input characteristics. An alternative approach is dynamic convolution [14], which predicts the final high-pass kernel k_t directly instead of the mixing coefficients α . Compared to this alternative, our method reduces the number of learnable weights from $T_k H_k W_k$ to M , simplifying the optimization process. Consequently, our approach is more efficient in both training and inference. Moreover, our method ensures

that the generated kernel is high-pass (see Proposition 1) and has proven effective (see Section 4.3).

In our implementation, $\mathcal{H}$ adopts $M = 4$ basis kernels. Two are normalized 3×3 Sobel kernels in the horizontal and vertical direction. The other two are temporal differencing kernels $\begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -1/2 \\ -1/2 \end{bmatrix}$, denoting $-x_t + x_{t+1}$ for the first and $x_{t-1} - 0.5x_t - 0.5x_{t+1}$ for the second. The second temporal kernel is derived from $\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$ using Gram-Schmidt process to ensure orthogonality to the first. We opt for these four kernels as bases owing to their simplicity and comprehensiveness. This combination covers all directions: horizontal and vertical in the spatial domain, and forward and backward in the temporal domain. Although other alternatives exist, no significant differences have been observed.

3.4. AHFNet

We present the overview of our entire model in Figure 2. The overall architecture of our network can be viewed as the general transformation $\mathcal{L}$ in Eq 2. There are $N + 1$ paths. The first path, $\mathcal{H}^0$, is implemented as an identity operation; while not strictly a high-frequency extraction, we abuse the notation $\mathcal{H}$ for simplicity. The subsequent N extraction modules use the adaptive high-frequency extraction module, $\{\mathcal{H}^n\}_{n=1}^N$, each paired with preprocessing modules $\{\mathcal{P}^n\}_{n=1}^N$. While these paths employ the same architectural design, they operate with unique parameters, resulting in diverse generated kernels k_t . Each extraction module takes a blurry video, X , and converts it into a unique feature set, $\{\mathcal{H}^n(x_t)\}_{n=1}^N$ for the t -th frame x_t . This output is then passed to a respective preprocessing module, $\mathcal{P}^n$, to be combined with other features for further analysis. The result of the n -th path for the t -th frame is:

$$p_t^n = \mathcal{P}^n(h_t) = \mathcal{P}^n(\mathcal{H}^n(x_t)). \quad (8)$$

The recurrent extraction module $\mathcal{R}$ takes the feature set $\{p_t^n\}_{n=0}^N$ and the feature from the previous frame, denoted as h_{prev} , as input. We fuse them using simple concatenation; preliminary experiments showed that more sophisticated fusion structures did not improve performance. The warping module $\mathcal{W}$ warps h_{prev} with the optical flow computed between x_{prev} and x for feature alignment [3]. As a result, the process of the deblurring module can be formulated as:

$$r_t = \mathcal{R}(\{p_t^n\}_{n=0}^N, \mathcal{W}(r_{\text{prev}}, x_{\text{prev}}, x_t)), \quad (9)$$

Our model is bidirectional. r_{prev} denotes r_{t-1} and r_{t+1} during forward and backward pass, respectively. In Figure 2, we make the process for other frames more transparent to highlight the main processing. Finally, the upsample module $\mathcal{U}$ takes the r_t from the forward (r_t^f) and backward passes (r_t^b) as input to reconstruct the final output:

$$\hat{y}_t = x_t + \mathcal{U}(r_t^f, r_t^b) \quad (10)$$

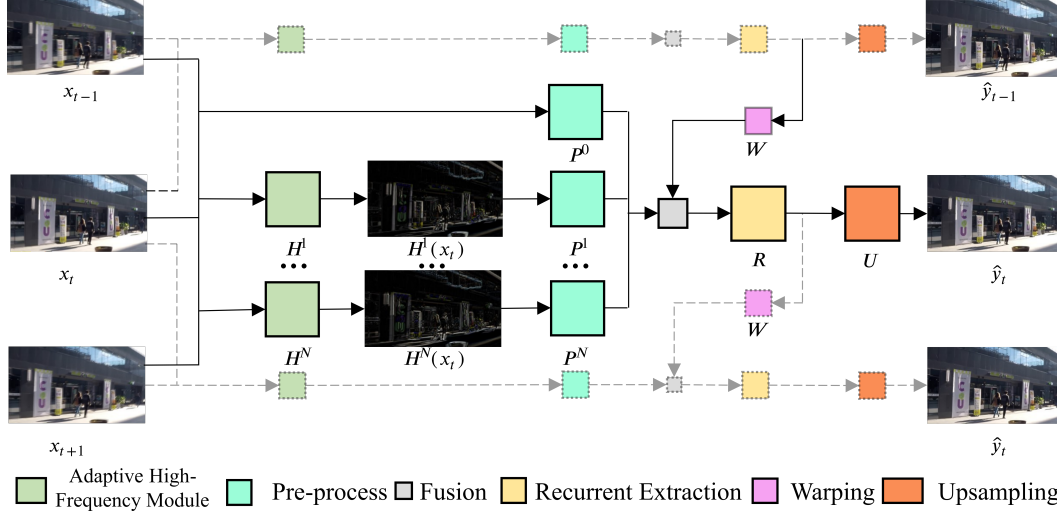


Figure 2. Overview of AHFNet. We extract high frequencies using $\mathcal{H}$, which are explicitly utilized for deblurring.

The predicted residual $\mathcal{U}(r_t^f, r_t^b)$ is equivalent to $\mathcal{L}(x_t, \mathcal{H}(\{x_i\}_{i=t-l}^{t+l}))$ in Eq 2.

The high-frequency extraction modules are integrated into the head stage of the video deblurring model, with the recurrent extraction module $\mathcal{R}$ serving as the primary deblurring component. We provide a detailed architecture in Section 4.1.

Our model strikes a balance between performance and computational budgets in both training and inference. Unlike other works using complicated modules such as transformers [22] and deformable convolutions [15, 43], we mainly use a limited number of standard convolutions to build our network. This ensures the efficiency. Despite its simplicity, this design has been proven effective in improving performance (see Section 4.3 and 4.4).

4. Experiments

4.1. Experimental Setting

Datasets. We used the DVD [38] and GOPRO datasets [25]. The DVD dataset comprises 61 training and 10 testing videos with 6708 blurry-sharp image pairs. The GOPRO dataset has 22 training and 11 testing sequences, with 3214 image pairs. Following [27], we used the version without gamma correction. Each frame is of size 1280×720 .

Evaluation metrics. The standard metrics for deblurring are the peak signal-to-noise (PSNR) and SSIM [45]. The memory footprint for training each model is measured as the GPU memory required to process a batch of eight 256×256 frames, including both forward and backward propagation. The computational expense of each model is indicated by run-time and number of giga multiply-accumulate operations (GMACs) for the original resolution input. Lower run-times and GMACs denote greater efficiency.

Training Details. For training, we used a Charbonnier

loss [4] and the ADAM optimizer [16] with default hyperparameters, *i.e.* $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Charbonnier loss preserves edges better than L2 by penalizing outliers less aggressively and avoids artifacts from L1’s non-differentiability at zero. The initial learning rate was set to 2×10^{-4} and decayed with a cosine restart strategy [23] with a minimum learning rate of 1×10^{-7} ; the number of epochs between two warm starts was set to 300k. Our model is trained for 600k iterations. We applied random rotations and flipping for data augmentation, using a batch size of 8, and 10-frame training sequences with training patches of 256×256 . The training used 2 RTX A5000 GPUs.

Architecture Details. The preprocessing module $\mathcal{P}$ is a one-layer convolution followed by two residual dense blocks that downsample the input by a scale factor of 0.25. The extraction module $\mathcal{R}$ consists of a single convolution layer followed by 30 residual blocks without batch normalization [21]. The output of the residual blocks is upsampled using pixel shuffle modules [37] to obtain the residual in $\mathcal{U}$. SPyNet [34] is used as the warping module $\mathcal{W}$.

4.2. Comparison With State-of-the-Art

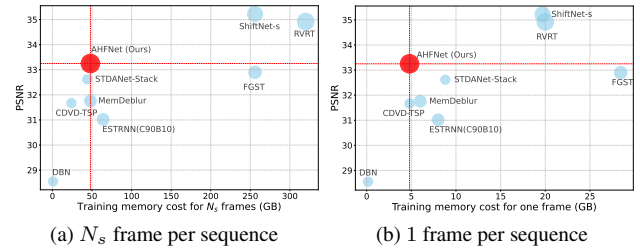


Figure 3. Trade-off between PSNR on GOPRO and the training memory cost. The cost is measured using a 256×256 patch size and a batch size of 8 for processing N_s frames (a) and a single frame (b). The dot size represents N_s , which is the number of frames per training sequence used in the official implementation.

We first present a comparison with all state-of-the-art methods based on the PSNR and training memory cost. Figure 3 shows that our model achieves an optimal trade-off between training memory cost and PSNR. Using only about 50GB of memory for training one batch, we achieve a PSNR of 33.25, outperforming many models, including FGST [22], which requires more than 250GB, as shown in Figure 3a. While models like ShiftNet [18] and RVRT [20] achieve higher PSNRs, they need at least 5 times the GPU memory for training. The left image shows the full GPU memory their official implementations require, which is about 250GB (8 GPUs). We further provide Figure 3b to show the memory cost when processing a single frame during training, as different models use different number of frames per training sequence. For example, our method uses 10 frames per training sequence, while RVRT uses 16. When fixing this variable factor to be the same, our model still locates at the top-left position in Figure 3b, indicating high performance with fewer training resources. One reason for the high training cost of other models is their extensive use of complex operations, such as self-attention, to continually refine specific features.

To ensure a fair comparison and thorough comparison, we evaluate our model against state-of-the-art models trained with similar training budgets on the GOPRO [25] and DVD [38] datasets, as presented in Tables 1 and 2, respectively. Hence, ShiftNet, RVRT and FGST are not included. In our presentation of results, we detail the complexity metrics only in Table 1 and omit these details from Table 2. Our model achieves state-of-the-art with relatively fewer GMACs and less run-time than models with similar training budgets. Compared with MMP-RNN, our PSNR is 0.61dB higher with 70% of the run-time despite larger GMACs. MMP-RNN’s main building block is the residual dense block. The dense connectivity in MMP-RNN leads to more memory operations and is less optimized than normal one, which decreases its runtime. Compared with ERDN and STDANet, our model is superior in performance and efficiency. Figure 4 illustrates the relationship between PSNR and inference time, with the radius of each model denoting GMACs. This visually demonstrates the efficiency and effectiveness of our proposal. Qualitatively, Fig. 5 shows that our model excels in producing the sharpest curves of the car, which is originally severely blurred.

4.3. Predefined versus Learned Kernels

In this section, we evaluate the dynamic approach to generating kernels. There are two primary methods for dynamically generating convolutional kernels k_t : (1) directly predicting k_t [14], and (2) predicting the coefficients $\{\alpha_{t,j}\}_{j=1}^M$ and using basis kernels $\{\tilde{k}_j\}_{j=1}^M$ to construct k_t [46]. As mentioned in Section 3, our method adopts the second generation method because we aim to constrain the generated

Table 1. Performance comparison to models with a similar training memory footprint on the GOPRO dataset

Model	PSNR	SSIM	GMACs	Params (M)	Time(s)
DBN [38]	28.55	0.8595	784.75	15.31	0.063
IFIRNN (c_2h_3) [26]	29.80	0.8900	217.89	1.64	0.053
SRN [41]	29.94	0.8953	1527.01	10.25	0.244
ESTRNN ($C_{90}B_{10}$) [55]	31.02	0.9109	215.26	2.47	0.124
CDVD-TSP [27]	31.67	0.9279	5122.25	16.19	1.763
MemDeblur [12]	31.76	0.9230	344.49	6.99	0.152
ERDN [15]	32.48	0.9329	29944.57	45.68	5.094
STDANet-Stack [52]	32.62	0.9375	6000.00	13.84	2.827
MMP-RNN [44]	32.64	0.9359	264.52	4.05	0.206
AHFNet	33.25	0.9439	461.35	6.75	0.144

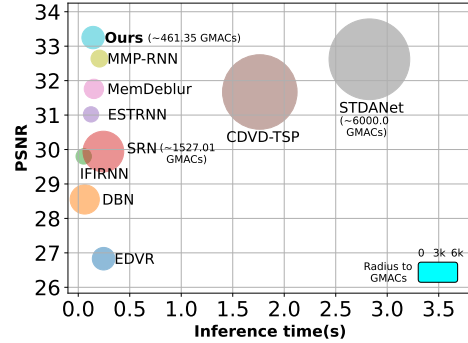


Figure 4. Inference time for a single 1280×720 frame vs. PSNR on the GOPRO dataset [25]. The dot size represents GMACs. Compared models share a similar training memory footprint. Our method achieves the best PSNR with minimal inference time.

Table 2. Performance comparison to models with a similar training memory footprint on the DVD dataset. As ARVo [19] did not provide details such as MACs and there is no official code available, these values are marked as “-”.

Model	PSNR	SSIM	GMACs	Params (M)	Time(s)
SRN	30.53	0.8940	1527.01	10.25	0.244
IFIRNN (c_2h_3)	30.80	0.8991	217.89	1.64	0.053
EDVR	31.82	0.9160	468.25	20.04	0.246
CDVD-TSP	32.13	0.9268	5122.25	16.19	1.763
ARVo [19]	32.80	0.9352	-	-	-
STDANet-Stack	33.05	0.9374	6000.00	13.84	2.827
AHFNet	33.19	0.9400	461.35	6.75	0.144

kernel k_t to be high-pass, which the first generation method does not ensure. For each generation strategy, we create a variant for comparison. Specifically, the first strategy yields the variant denoted as “3D Conv,” and the second yields “Naive Basis Kernels”, where $\{\tilde{k}_j\}_{j=1}^M$ consists of standard basis with patterns like $[1 \ 0 \ 0 \ \dots \ 0]$, $[0 \ 1 \ 0 \ \dots \ 0]$, ..., $[0 \ 0 \ 0 \ \dots \ 1]$. We opt for these specific kernels because they represent standard bases in the vector space. It is obvious that substituting these basis kernels with high-pass kernels yields our method. Both the “3D Conv” and “Naive Basis Kernels” variants have similar complexity to our approach but are not specifically designed to extract high frequencies. Each variant was trained for 300k iterations.

Table 3 shows that our proposed module achieves the



Figure 5. Qualitative comparisons to models with a similar training memory footprint.

Table 3. Validation of the HF extraction module on GOPRO.

	PSNR	SSIM	GMACs	Params (M)
3DConv	32.82	0.9384	462.71	6.75
Naive Kernels	32.80	0.9378	461.35	6.75
AHFNet	32.89	0.9398	461.35	6.75

highest PSNR and SSIM, demonstrating improvement without adding complexity. The PSNR difference between the two generation strategies is 0.02 dB, but the direct generation method requires predicting more learnable parameters since the entire kernel is of size $T_k \times H_k \times W_k$. In contrast, the coefficient generation strategy is more efficient and deals with only M learnable parameters.

Moreover, to study the effects of learning in different frequency subbands, we perform the DFT on the output images, normalize their frequencies to the range $[0, 2\pi]$, and divide the Fourier spectra of the outputs from each variant into 10 equal subbands. Each subband covers an interval of $2\pi/10$, from $[0, 2\pi/10]$ to $(18\pi/10, 2\pi]$. We then calculate the relative average mean squared error (MSE) for each subband on the GOPRO dataset. The MSE values are relative to the variant without the dynamic module, which can result in negative values.

Figure 6 presents the result, where a lower curve indicates better performance. Our method’s curve, depicted in green, consistently shows the lowest values across most subbands, indicating superior performance. Despite slight reductions in performance at the highest frequency areas, likely due to training variance, our approach remains notably effective across most of the subbands.

Additionally, we assess scenarios with varying numbers of paths for extraction, *i.e.* N . The result is in Table 4. Increasing the number of paths can further enhance performance, but this comes at the expense of increased computational resources.

Table 4. Ablation on HF extraction module count N on GOPRO.

N	0	2	4	6
PSNR	32.56	32.68	32.71	32.89
SSIM	0.9366	0.9372	0.9371	0.9398
GMACs	410.73	427.66	444.51	461.35

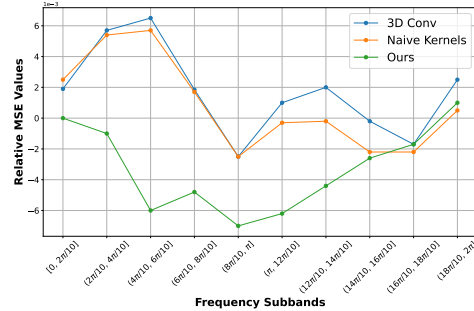


Figure 6. Comparison of learning in various frequency subbands. Lower MSE values indicate better performance.

Table 5. Experiments on various high-pass filters on GOPRO.

Added Representation	PSNR	SSIM	GMACs
+ RGB $\times$ 1	32.47	0.9363	400.82
+ $\nabla^2 x$	32.37	0.9335	402.26
+ $\nabla_t x$	32.41	0.9340	400.82
+ ∇x (Kirsch)	32.51	0.9356	403.76
+ ∇x (Sobel)	32.53	0.9363	403.76
+ RGB $\times$ 2	32.45	0.9356	414.39
+ ∇x (Sobel) + $\nabla_t x$	32.78	0.9394	417.28
+ Random	31.98	0.9312	444.51

4.4. Choice of basis kernels

We evaluate our choice of basis kernels $\{\tilde{k}_j\}_{j=1}^M$ (see Sec. 3.3) by comparing with two other options. First, we

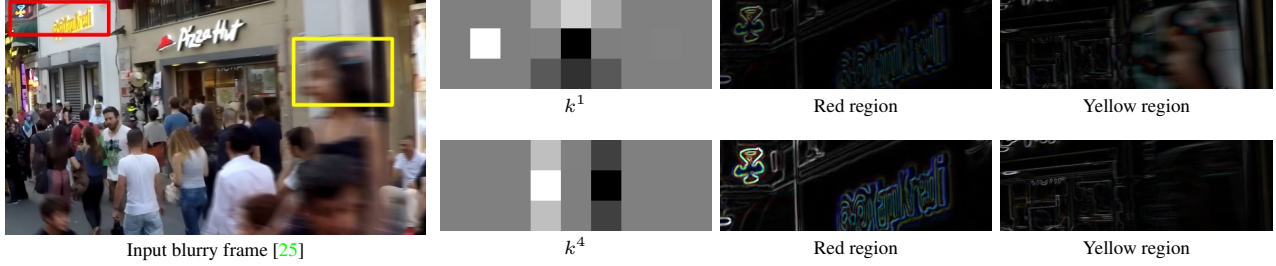


Figure 7. Examples of learned kernels and features on GOPRO dataset (please zoom in for a better view).

use standard high-pass filters, including first- (∇x , e.g. Sobel and Kirsch) and second-order kernels ($\nabla^2 x$, e.g. Laplacian), and a temporal gradient filter ($\nabla_t x$) given by $[[1/2], [-1], [1/2]]$. Secondly, we follow [42] to randomly generate high-pass filters as our basis kernels. Specifically, given an arbitrary matrix $\mathbf{A}$, the corresponding high-pass filter is $(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\mathbf{A}$, where $\mathbf{I}$ is an identity matrix, n is the number of elements in $\mathbf{A}$ and $\mathbf{1}$ is a vector filled with ones. This variant is referred to as “+ Random”, with the basis kernels randomly generated for each training run. We keep these kernels relatively simple, as the subsequent pre-processing module $\mathcal{P}$ (see Fig. 2) use non-linearity to facilitates hierarchical feature extraction, evolving low-level features into complex representations. We used the same architecture in Section 3.4 and replaced these choices with the high-frequency extraction module.

Comparison results are reported in Table 5. As a control study to match the complexity, we also show variants “+ RGB $\times k$ ”, which repeats the early module ($\mathcal{P} \circ \mathcal{H}$) k times. Adding first-order filters improves PSNR by up to 0.06dB across variants, while second-order and temporal gradients do not boost performance. Second-order spatial gradients are more noise-sensitive. Using only the temporal gradient results in inaccurate motion estimations because consecutive frames can have different levels of blur. Integrating first-order and temporal filters (“+ ∇x (Sobel) + $\nabla_t x$ ”) significantly enhances image quality, increasing PSNR by 0.33dB and SSIM by 0.0038 compared to the baseline (“+ RGB $\times 2$ ”). The first-order feature often positively correlates with blur characteristics and helps refine temporal feature. Therefore, their combination achieves the best results. Moreover, compared to the “+ Random” variant, our method elevates PSNR by 0.8dB and SSIM by 0.0082, with fewer GMACs. The “+ Random” variant has higher computational costs, as it requires more efforts to generate high-pass kernels. This indicates that predefined high-pass filters, especially the combination of Sobel and temporal gradients, outperform randomly generated ones. These findings guide our choice of kernels for the AHFNet architecture.

4.5. Visualizations

Figure 7 visualizes the learned kernels, $\{k_t^n\}_{n=1}^N$, and the high-frequency features extracted by our learned kernels,

$\{h_t^n\}_{n=1}^N$. We select two out of $N = 6$ representations for visualization. Kernels are normalized to $[0, 1]$, where white, gray, and black signify values of 1, 0.5, and 0, respectively.

The first row combines both spatial and temporal details. The second row focuses exclusively on the current frame, resulting in a spatial gradient map, where shop signs are clearly visible, but human faces are barely distinguishable. These two features highlight different aspects of the same frame, showing the effectiveness of our model in extracting varied features, which enhances the deblurring process. This diversity arises because each high-pass kernel basis captures high frequencies from different directions, with coefficients predicted by modules initialized with varied parameters. Due to the non-convex nature of neural networks, even slight differences in initialization lead to distinct parameter updates, further diversifying the features. Our choice of basis kernel is currently simple, resulting in similarly simple generated kernels. However, more complex basis kernels may be advantageous.

5. Conclusion

We emphasize the crucial role of high frequencies in video deblurring and formulate our algorithm to explicitly extract and utilize these frequencies. Experimental results underscore the significance of manual high-frequency extraction. Notably, the combination of first-order and temporal gradients substantially enhances performance.

Progressing further, we introduce adaptability in high-frequency extraction by generating the coefficients of several high-pass filter kernels. Initially, we show that the linear combination of positive coefficients and high-pass filter kernels continues to function as a high-pass filter. This assurance confirms our generated kernel’s capacity to extract high frequencies from the input feature efficiently. By integrating this module into our model, we attain state-of-the-art results in video deblurring datasets regarding performance and complexity among low-budget models. Visual demonstrations further highlight the effectiveness of our proposed approach. Our approach shows enhanced performance across various frequency subbands and improved filtered results in the spatial domain.

References

- [1] Brian R Bartoldson, Bhavya Kailkhura, and Davis Blalock. Compute-efficient deep learning: Algorithmic trends and opportunities. *Journal of Machine Learning Research*, 24(122):1–77, 2023. [1](#)
- [2] Jian-Feng Cai, Hui Ji, Chaoqiang Liu, and Zuowei Shen. Blind motion deblurring using multiple images. *Journal of computational physics*, 228(14):5057–5071, 2009. [1](#), [2](#)
- [3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. [4](#)
- [4] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, pages 168–172. IEEE, 1994. [5](#)
- [5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11030–11039, 2020. [2](#)
- [6] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–8. 2009. [2](#)
- [7] Guang Deng. A generalized unsharp masking algorithm. *IEEE transactions on Image Processing*, 20(5):1249–1261, 2010. [1](#), [2](#)
- [8] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. [1](#)
- [9] Juntao Guan, Rui Lai, Yang Lu, Yangang Li, Huanan Li, Lichen Feng, Yintang Yang, and Lin Gu. Memory-efficient deformable convolution based joint denoising and demosaicing for uhd images. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7346–7358, 2022. [2](#)
- [10] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. [2](#)
- [11] Bernd Jähne. *Digital image processing*. Springer Science & Business Media, 2005. [2](#)
- [12] Bo Ji and Angela Yao. Multi-scale memory-based video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1919–1928, 2022. [2](#), [6](#)
- [13] Seo-Won Ji, Jeongmin Lee, Seung-Wook Kim, Jun-Pyo Hong, Seung-Jin Baek, Seung-Won Jung, and Sung-Jea Ko. Xydeblur: divide and conquer for single image deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17421–17430, 2022. [2](#)
- [14] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. [2](#), [4](#), [6](#)
- [15] Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 663–678. Springer, 2022. [1](#), [2](#), [5](#), [6](#)
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#)
- [17] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011. [1](#), [2](#)
- [18] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9822–9832, 2023. [1](#), [2](#), [6](#)
- [19] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7721–7731, 2021. [2](#), [6](#)
- [20] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jie Zhang Cao, Kai Zhang, Radu Timofte, and Luc V Gool. Recurrent video restoration transformer with guided deformable attention. *Advances in Neural Information Processing Systems*, 35:378–393, 2022. [1](#), [2](#), [6](#)
- [21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017. [5](#)
- [22] Jing Lin, Yuanhao Cai, Xiaowan Hu, Haoqian Wang, Youliang Yan, Xueyi Zou, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Flow-guided sparse transformer for video deblurring. *arXiv preprint arXiv:2201.01893*, 2022. [1](#), [2](#), [5](#), [6](#)
- [23] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. [5](#)
- [24] Ningning Ma, Xiangyu Zhang, Jiawei Huang, and Jian Sun. Weightnet: Revisiting the design space of weight networks. In *European Conference on Computer Vision*, pages 776–792. Springer, 2020. [2](#)
- [25] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. [2](#), [5](#), [6](#), [7](#), [8](#)
- [26] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8102–8111, 2019. [6](#)
- [27] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3043–3051, 2020. [1](#), [2](#), [5](#), [6](#), [7](#)

- [28] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. L_0 -regularized intensity and gradient prior for deblurring text images and beyond. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):342–355, 2016. 1, 2
- [29] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1628–1636, 2016. 2
- [30] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Feb 2023. 1
- [31] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 1, 3
- [32] Andrea Polesel, Giovanni Ramponi, and V John Mathews. Image enhancement via adaptive unsharp masking. *IEEE transactions on image processing*, 9(3):505–510, 2000. 2
- [33] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 1
- [34] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 5
- [35] Alex Rav-Acha and Shmuel Peleg. Two motion-blurred images are better than one. *Pattern recognition letters*, 26(3):311–317, 2005. 2
- [36] Qi Shan, Jiaya Jia, and Aseem Agarwala. High-quality motion deblurring from a single image. *Acm transactions on graphics (tog)*, 27(3):1–10, 2008. 2
- [37] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5
- [38] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017. 2, 5, 6, 7
- [39] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Edge-based blur kernel estimation using patch priors. In *IEEE international conference on computational photography (ICCP)*, pages 1–8. IEEE, 2013. 2
- [40] Ling Tang, Wen Shen, Zhanpeng Zhou, Yuefeng Chen, and Quanshi Zhang. Defects of convolutional decoder networks in frequency representation. *arXiv preprint arXiv:2210.09020*, 2022. 1, 3
- [41] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 6
- [42] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*, 2022. 1, 3, 8
- [43] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 5
- [44] Yusheng Wang, Yunfan Lu, Ye Gao, Lin Wang, Zhihang Zhong, Yinqiang Zheng, and Atsushi Yamashita. Efficient video deblurring guided by motion magnitude. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 413–429. Springer, 2022. 1, 2, 6, 7
- [45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [46] Zhihao Xia, Federico Perazzi, Michaël Gharbi, Kalyan Sunkavalli, and Ayan Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11844–11853, 2020. 2, 6
- [47] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part I 11*, pages 157–170. Springer, 2010. 2
- [48] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1107–1114, 2013. 1, 2
- [49] Liuge Yang and Hui Ji. A variational em framework with adaptive edge selection for blind motion deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10167–10176, 2019. 2
- [50] Wei Ye and Kai-Kuang Ma. Blurriness-guided unsharp masking. *IEEE Transactions on Image Processing*, 27(9):4465–4477, 2018. 1, 2
- [51] Haichao Zhang, David Wipf, and Yanning Zhang. Multi-image blind deblurring using a coupled adaptive sparse prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1051–1058, 2013. 1, 2
- [52] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 581–596. Springer, 2022. 6, 7
- [53] Meina Zhang, Yingying Fang, Guoxi Ni, and Tieyong Zeng. Pixel screening based intermediate correction for blind deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5892–5900, 2022. 2
- [54] Yi Zhang, Dasong Li, Xiaoyu Shi, Dailan He, Kangning Song, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Kbnnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881*, 2023. 2

- [55] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 191–207. Springer, 2020. 2, 6, 7
- [56] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2482–2491, 2019. 2
- [57] Xiang Zhu, Filip Šroubek, and Peyman Milanfar. Deconvolving psfs for a better motion deblurring using multiple images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12*, pages 636–647. Springer, 2012. 2