# Exploring Temporal Frequency Spectrum in Deep Video Deblurring

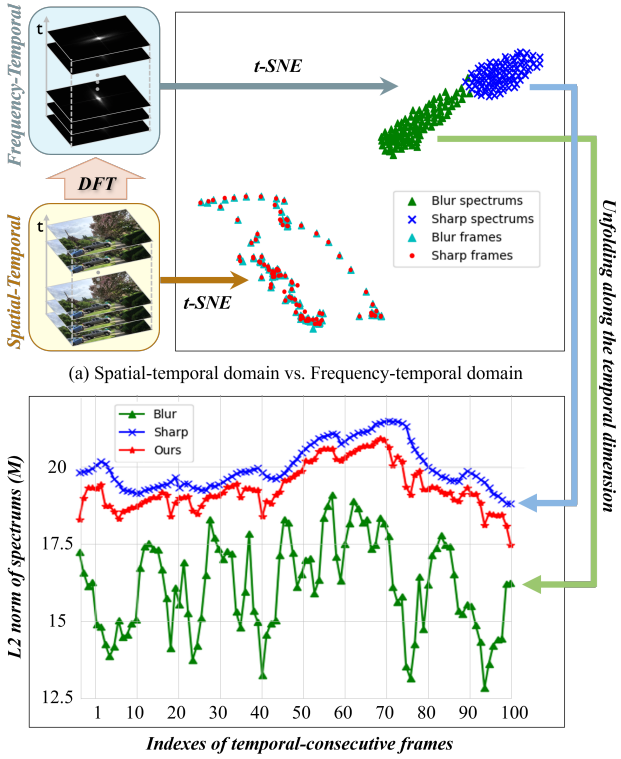
Qi Zhu [1]*, Man Zhou [1]*, Naishan Zheng [1], Chongyi Li [2], Jie Huang [1], Feng Zhao [1]†
[1]University of Science and Technology of China, [2]Nankai University
{zqcrafts,manman,nszheng,hj0117}@mail.ustc.edu.cn,lichongyi25@gmail.com,
fzhao956@ustc.edu.cn

## Abstract

*Video deblurring aims to restore the latent video frames from their blurred counterparts. Despite the remarkable progress, most promising video deblurring methods only investigate the temporal priors in the spatial domain and rarely explore their its potential in the frequency domain. In this paper, we revisit the blurred sequence in the Fourier space and figure out some intrinsic frequency-temporal priors that imply the temporal blur degradation can be accessibly decoupled in the potential frequency domain. Based on these priors, we propose a novel Fourier-based frequency-temporal video deblurring solution, where the core design accommodates the temporal spectrum to a popular video deblurring pipeline of feature extraction, alignment, aggregation, and optimization. Specifically, we design a Spectrum Prior-guided Alignment module by leveraging enlarged blur information in the potential spectrum to mitigate the blur effects on the alignment. Then, Temporal Energy prior-driven Aggregation is implemented to replenish the original local features by estimating the temporal spectrum energy as the global sharpness guidance. In addition, the customized frequency loss is devised to optimize the proposed method for decent spectral distribution. Extensive experiments demonstrate that our model performs favorably against other state-of-the-art methods, thus confirming the effectiveness of frequency-temporal prior modeling.*

## 1. Introduction

Video deblurring, as a fundamental vision task, aims to recover the latent frame from the blurred sequence by leveraging intrinsic temporal information. Therefore, many research efforts have been advocated to explore the potential of hidden temporal information in blurred sequences, which can be categorized into two groups: traditional optimization and deep learning-based methods.




(a) Spatial-temporal domain vs. Frequency-temporal domain

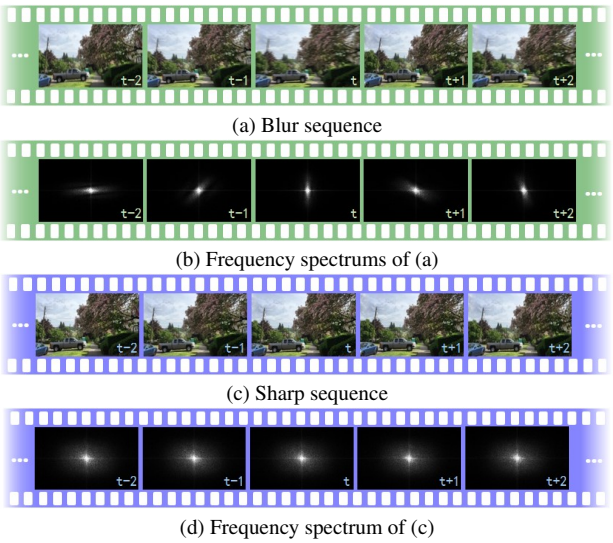
(b) Temporal distribution in the frequency domain

Figure 1. (a) Clustering blurry and sharp frames in the spatial domain and frequency domain via t-SNE, respectively. The blurry and sharp frames in the frequency domain are separated while they are tangly in the spatial domain. (b) Unfolding the sets of the frequency domain in (a) along the temporal dimension. The blurred video appears to have greater temporal fluctuation than the sharp one in terms of the L2 norm of spectrums.

Traditional optimization methods often highlight the assumptions over the blur degradation process and apply some hand-crafted temporal priors to alleviate the video deblurring, *e.g.*, temporal sharpness prior [2], motion-blurred prior [1], and temporal coherence prior [6]. However, these priors are difficult to design and the methods are also difficult to optimize, limiting their practical usage.

In recent years, we have witnessed explosive deep

*Both authors contributed equally to this research.
†Corresponding author.




(a) Blur sequence

(b) Frequency spectrums of (a)

(c) Sharp sequence

(d) Frequency spectrum of (c)

Figure 2. Visualization comparison of several consecutive blurry frames (green) and sharp frames (blue). (a) and (c) are in the spatial domain, while (b) and (d) are in the frequency domain.

learning-based video deblurring approaches for addressing the above challenges. Most of them follow the common pipeline: **feature extraction, alignment, fusion, and optimization**. Specifically, the pioneering EDVR [26] formulates an implicit alignment via the redesigned deformable convolution [42]. Instead, RTA [37] employs a gradual refinement scheme to execute the motion compensation for more accurate temporal modeling. Despite the remarkable progress, the above strategies only investigate the temporal information from the perspective of spatial domain and have not fully explored its potential. **We thus wonder "Can we provide a new solution to effectively model the temporal prior information?".**

To answer this question, we first revisit the differences between blurry-sharp pairs in the spatial and frequency domains respectively (see Figure 1(a)), and then unfold the frequency information of the blurry-sharp sets along the time dimension (see Figure 1(b)). In addition, we crop a video clip and perform the Discrete Fourier Transform (DFT) to visually illustrate our motivation (see Figure 2). On the basis of the above analyses, we can infer the following rules to encourage us to effectively explore and exploit the temporal prior information:

- **(1)** In terms of Figure 1(a), the feature distributions of the blurry frames and their sharp counterparts are intertwined in the spatial domain. Instead, their frequency feature distributions are distinguished by employing DFT. Therefore, the blurred degradation can be better modeled in the frequency domain.

- **(2)** In terms of Figure 1(b), due to the addition of motion blur, the spectral energy calculated by the L2 norm in the blurry video fluctuates more intensely than in the sharp ones temporally. Furthermore, for qualitative comparison, Figure 2(b) and (d) shows that the frequency spectrum of the blurry video has more obvious temporal differences than the spectrum of the sharp video. In short, the frequency spectrum can enlarge the unpredictable change of temporal motion blur.

- **(3)** In the vertical comparison of Figure 1(b), the spectrum norm of the sharp video is larger than the blur one. Furthermore, the high-frequency information is lost due to blur degradation, as demonstrated in Figure 2(b) and (d). This phenomenon implies that the blur degradation may decay the energy spectrum of the sharp videos.

Based on the above observations, we propose a novel Fourier-based frequency-temporal solution for video deblurring. The key insight is to transform the blurred feature sequences into the frequency domain applying DFT, then explore and exploit the temporal sharp cues over the above pipeline of video deblurring as follows:

- **Feature extraction.** Based on observation (1), the blur degradation can be effectively modeled in the frequency domain. Therefore, we design the Spatial-frequency Feature Extraction (SFE) block by employing the Fourier transform and pure convolution unit.

- **Alignment.** Temporal alignment aims to reduce the context difference between adjacent frames. However, it often becomes not so effective over the unpredictable blur case where high-frequency details are lost severely. Based on observation (2), we propose the Spectrum Prior-guided Alignment (SPA) module by excavating the enlarged blur degradation information in the frequency spectrum to relieve the negative impact caused by motion blur in the alignment.

- **Fusion.** Temporal fusion is responsible for aggregating clear patches from multiple frames. However, most existing methods only focus on local spatial features, which are limited by the preceding alignment results. To address this issue, based on observations (2) and (3), we propose the Temporal Energy Attention (TEA) module, which equips the temporal spectrum energy as the global sharpness guidance to achieve complementary effects with previous local spatial manners.

- **Optimization.** Frequency spectrum can be regarded as an indicator of global blurriness. Based on observations (1) and (3), we devise the frequency spectrum loss and energy loss functions to better optimize the proposed solution in the frequency domain.

Extensive experiments are performed over multiple video deblurring tasks and validate the superiority of our proposed method. Specifically, in Figure 1(b), our solution is capable of restoring a closer spectrum distribution with ground-truth frames.

## 2. Related Work

**Image Deblurring.** With the advances in vision benchmarks, CNN models have excelled in various image enhancement tasks [5, 10, 11, 14, 29, 35, 36, 38, 39, 41], by innovative architectures and specialized modules. Image deblurring seeks to produce sharp images from blurred ones. Traditional efforts to refine deblurring performance hinge on various priors for natural images and kernels, such as the sparse kernel prior [4], $l_0$ gradient prior [30], normalized sparsity prior [8], and dark channels [19]. Yet, these approaches often fall short when addressing spatially variant blur. The advent of deep learning shifts focus towards advanced non-uniform deblurring techniques [16, 24, 25, 32, 33, 34]. For instance, Nah *et al.* [17] employ a multi-scale loss function for a fine-tuned approach. DeepRFT [16], on the other hand, leverages the spectral difference between a sharp image and its blurry one, addressing limitations in the spatial domain.

**Video Deblurring.** While single image deblurring focuses solely on one frame, video deblurring leverages temporal information to yield visually compelling outcomes. Many existing approaches employ CNN-based structures. In this domain, temporal alignment is designed to harness sharp patches from adjacent frames. Several methods use optical flow [23, 31] and deformable convolution [26] to estimate the motions and align them with adjacent frames explicitly or implicitly. In addition to alignment, the rational use of multiple frames is also significant. Li *et al.* [13] effectively exploited the depth map as guidance through Spatial Feature Transform (SFT) [27] to better extract the blurred frames' features. The authors in [18] developed a temporal sharpness prior to achieve the decent latent frame restoration. Lai *et al.* [12] crafted a correlation-based aggregation module to efficiently process neighboring sharp patches, while RTA [37] brought an iterative alignment process, allowing for incremental motion compensation enhancements. However, the above methods less consider the frequency spectrum, which may limit the exploration of temporal information in video deblurring.

## 3. Motivation

As stated in the introduction, we derive some temporal prior information from Figures 1 and 2 to guide the video deblurring. In this section, some mathematical and visual analyses are given to better motivate our work.

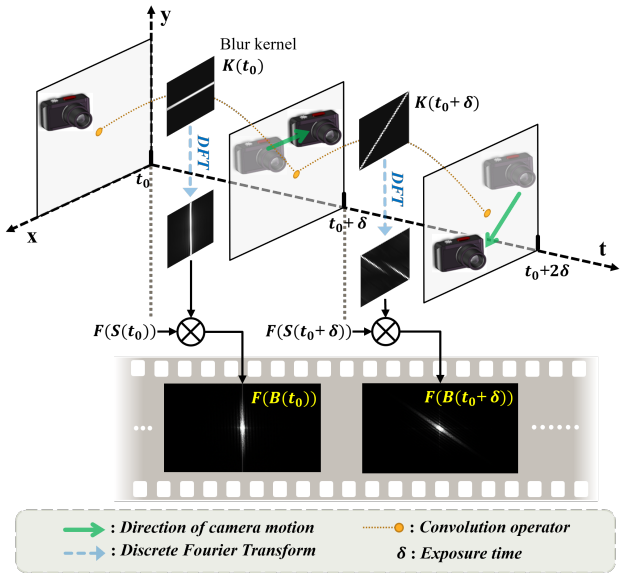



Figure 3. Different oriented blur kernels are used to simulate the direction of the camera motion during the exposure time. The spectrum of the captured image is shown in the bottom sequence. We can observe that the spectrum implies a directional feature that is relevant to the camera motion.

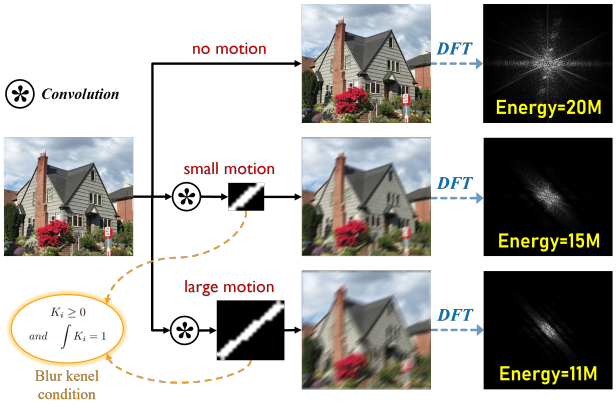



Figure 4. Different sizes of motion blur kernels convolve the input to produce frames with different blur degrees. From top to bottom, as the blur degree increases, the spectrum energies of these frames decrease gradually.

### 3.1. Approximate Motion Blur Modeling

In this work, for the sake of clarity and brevity, we leave out the effect of depth variation. The camera shake degradation model can be simplified mathematically:

$$B_t = S_t \star K_t + n_t, \tag{1}$$

where $K_t$ is an unknown blur kernel and $n_t$ is additive white noise. Note that although kernel-based modeling is efficient, the existing datasets employ more advanced simulation methods, such as temporal averaging and bi-camera

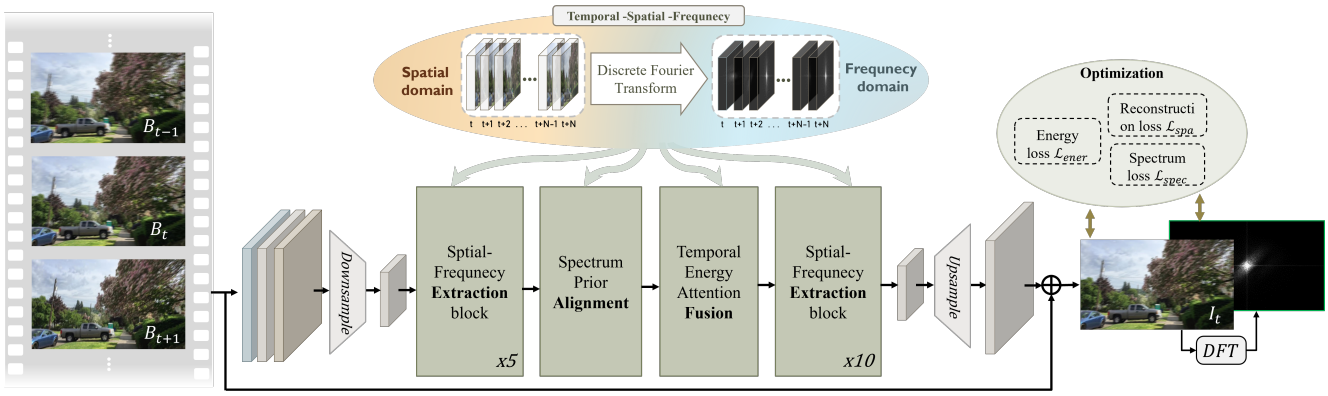



Figure 5. The framework of the proposed network. To exploit the potential frequency domain, we novelly adopt the frequency-temporal information to accommodate the popular video deblurring pipeline, including feature extraction, alignment, fusion, and optimization.

shots. We will analyze the rationality of the following priors on more realistic data in the supplementary material.

### 3.2. Blur Model on Temporal Fourier Spectrum

**Prior 1.** ***The blurred degradation can be better modeled in the frequency domain.***

It is well known that an image can be represented by real gray values in the spatial domain, whereas by complex frequency values in the frequency domain. From the image representation perspective, because both motion and frequency are directional, while the gray value is not, the camera motion is much easier to be observed in the frequency domain than in the spatial domain. Thus, the sharp and blurred images are distinguished decently in the frequency domain, as shown in Figure 1(a).

**Prior 2.** ***The frequency spectrum can enlarge the unpredictable change of temporal motion blur.***

As mentioned above, the Fourier transform is widely utilized to assess the frequency characteristics of images. For images that contain multiple color channels, the Fourier transform is computed separately for each channel. Given an image $S$, the Fourier transform F transfers it to Fourier space as the complex component, which is expressed as,

$$\mathcal{F}(S)(u,v) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} S(h,w) e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}. \tag{2}$$

As shown in Figure 3, we give some motion blur kernels with $\theta$ direction at $t$-th time, $K(t)$, to simulate the camera motion. These kernels have faint frequency responses along the $\theta$ direction instead of extensive frequency responses along the $\theta + \pi/2$ direction. According to Eq. (2), we perform the DFT for these kernels and observe that their frequency spectrums have large magnitudes along the $\theta + \pi/2$ direction, as shown in Figure 3. Furthermore, according to the convolution theorem, the convolution operation in the space domain is equal to the product in the Fourier frequency domain, which can be expressed as:

$$F(B_t) = F(S_t \star K_t) = F(S_t) \cdot F(K_t), \tag{3}$$

where $F(.)$ stands for the Fourier transform function. According to Eq. (3), the special pattern of strips in the spectrums of motion kernels will bring into the spectrums of blurred frames. Thus, the spectrums of blurred frames obtain a directional pattern roughly perpendicular to the direction of the camera motion, as depicted in the bottom row of Figure 3. Moreover, in terms of blur sizes, they can also be represented by the spectrum, which is stated in Prior 3 below. In summary, the frequency spectrum can amplify the unpredictable fluctuations in temporal motion blur. Inspired by this observation, we develop the SPA alignment module to relieve the negative impact caused by motion blur in the alignment.

**Prior 3.** ***The blur degradation may decay the energy spectrum of the sharp videos.***

Without compromising generality, the blur kernel $K_t$ can be normalized [3]. Because the integration of incoherent light is always non-negative, the blur kernel should be non-negative. In this case, we have the constraint condition to the kernel as follows:

$$K_t \geq 0 \quad and \quad \int K_t = 1. \tag{4}$$

Based on this, we infer that motion blur does not amplify the Fourier spectrum, which is proved in the following:

$$\begin{aligned} |F(K_t(x))| &= \left| \int K_t(x) e^{-j\omega x} dx \right| \\ &\leq \int |K_t(x)| dx = \int K_t(x) dx = 1, \end{aligned} \tag{5}$$

where $|.|$ denotes modulus value and $x$ is the spatial location of the kernel $K_t$. Combining Eqs. (3) and (5), the more

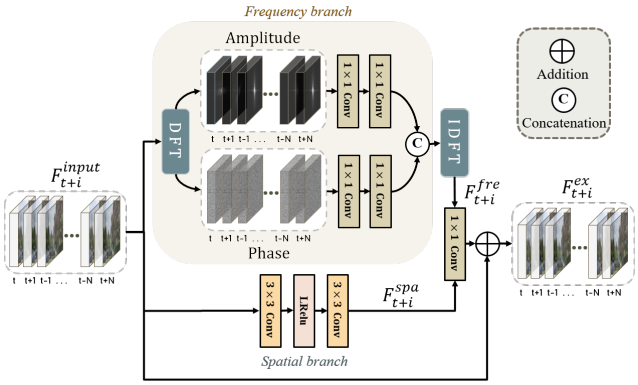



Figure 6. The detail of the Spatial-Frequency Extraction (SFE) block. The upper branch extracts features in the frequency domain, while the bottom branch extracts information from the spatial domain.

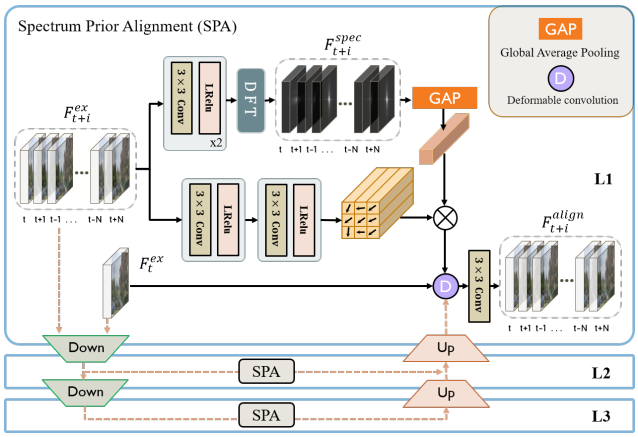



Figure 7. The detail of Spectrum Prior-guided Alignment (SPA). Global direction information in the spectrum modulates the offsets to align the input sequence effectively. Additionally, the pyramid strategy is performed to adapt to various sizes of motions.

significant blur may mean more energy loss. We further prove this conclusion by a toy experiment, as shown in the right column of Figure 4. Motivated by this finding, we design the temporal energy attention to take advantage of the frames where their energies are not attenuated by blur degradation in the temporal sequence.

# 4. Method

The overall framework of the proposed method is shown in Figure 5, which aims to restore the latent frame $I_t$ given $2N+1$ consecutive blurred frames $B_{t+i}$, where $i = 0, \pm1, \pm2, ..., \pm N$, under the supervision of the ground truth $S_t$. We innovatively incorporate the frequency-temporal information into the widely-used video deblurring pipeline, which includes feature extraction, alignment, fusion, and optimization. The specific details will be described in the following subsections.

## 4.1. Spatial-Frequency Extraction Block

In this work, we find that the blur degradation can be efficiently modeled in the frequency domain, as shown in Prior 1 of Section 3.2. Motivated by this, we develop the Spatial-Frequency Extraction (SFE) block, which consists of a spatial branch and a frequency branch, as shown in Figure 6. For the spatial branch, we adopt several convolutions to capture the spatial content and details. Meanwhile, the frequency branch is responsible for the blur degradation modeling, which divides the input features into the amplitude spectrum and phase spectrum by DFT. Then, to effectively modulate the frequency information, these spectrums are processed by $1 \times 1$ convolution kernel and inverse them to the spatial domain by IDFT, which is expressed as:

$$F^{fre}_{t+i} = F^{-1}(([c_2(Amp(F^{input}_{t+i})), c_1(Pha(F^{input}_{t+i}))])), \tag{6}$$

where $Amp$ and $Pha$ separately denote the amplitude spectrum and the phase spectrum, and $F^{-1}$ stands for the IDFT transform. Finally, the extracted feature from the two branches of the SFE block is expressed as:

$$F^{ex}_{t+i} = c_3([F^{fre}_{t+i}, F^{spa}_{t+i}]) + F^{input}_{t+i}, \tag{7}$$

where $c_1$, $c_2$, and $c_3$ represent different convolution layers with $1 \times 1$ kernels.

## 4.2. Spectrum Prior-guided Alignment

Existing temporal alignment methods usually focus on spatial features only to reduce content differences. However, due to the temporal unpredictable motion blur, most of them achieve sub-optimal results. To solve this problem, we deepen it by revisiting the blur degradation in the Fourier domain and observe that the frequency spectrum can enlarge the unpredictable change of temporal motion blur, as shown in Prior 2 of Section 3.2. Motivated by this prior, we propose the Spectrum Prior-guided Alignment (SPA) module by exploiting the enlarged blur information from the frequency spectrum to relieve the negative impact of the unpredictable blur in the alignment. Specifically, as shown in Figure 7, we first apply the Fourier transform to obtain the spectrum of the spatial features. To efficiently align adjacent frames, we adopt deformable convolution to learn the warping function from the $(t+i)$-th features to the $t$-th features implicitly. For the deformable convolution, the learned offset for each location usually is obtained by each 2M channel, where M denotes the size of the convolution kernel. To improve the alignment in large blurred cases, we perform the global average pooling for the spectrum feature to exploit the global motion information to modulate the offset of each location. So, given extracted features $F^{ex}_{t+i}$ and their spectrum features $F^{spec}_{t+i}$, the offset $\Delta x_{t+i}$ for $(t+i)$-th time is learned by:

$$\Delta x_{t+i} = c_4([F^{ex}_{t}, F^{ex}_{t+i}]) \cdot g([F^{spec}_{t}, F^{spec}_{t+i}]), \tag{8}$$

Table 1. Quantitative comparison in terms of PSNR, SSIM and model size on the video deblurring dataset [22]. The best results are in **bold**.

| Method | Kim *el al.* | Tao *et al.* | DGN | STFAN | EDVR | SFE | TSP | PVD | PVD-small | RTA | Ours | Ours-small |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSNR (dB) | 26.94 | 29.98 | 30.19 | 31.15 | 31.91 | 31.68 | 32.13 | 32.31 | 31.25 | 32.92 | **33.25** | 32.48 |
| SSIM | 0.8158 | 0.8842 | 0.9194 | 0.9049 | 0.9211 | 0.9157 | 0.9268 | 0.9260 | 0.9080 | 0.9480 | **0.9491** | 0.9291 |
| Model size (M) | - | **3.80** | 11.36 | 5.37 | 23.60 | 16.25 | 16.19 | 10.50 | 6.10 | 16.70 | 14.76 | 4.04 |

(a) Input (b) GT (c) Tao et al. [25] (d) SFE [28]

(e) STFAN [40] (f) TSP [18] (g) PVD [21] (h) Ours

Figure 8. Deblurred results on the test dataset [22]. The proposed method generates much clearer frames.

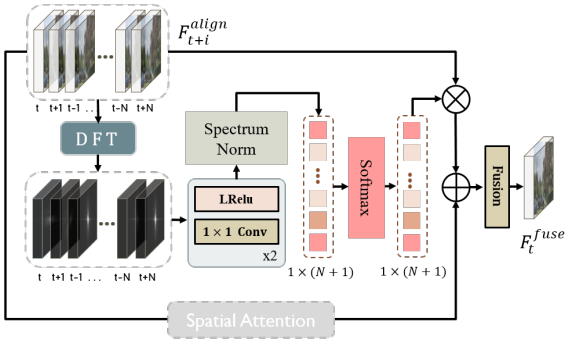



Figure 9. The detail of Temporal Energy Attention (TEA) fusion. Multiple frames are fused decently based on the difference of spectral norms in the temporal dimension.

where $c_4(.)$ and $g(.)$ denote several standard convolution and global average pooling functions. With the learned offsets, we warp the adjacent frame to the reference by the deformable convolution. In addition, the degree of motion is sensitive to change due to the depth variation. To tackle the problem, we use a feature pyramid strategy [15] to ensure the adaptation for different scales of blurs.

### 4.3. Temporal Energy Attention Fusion

In video deblurring, the key challenge is exploiting sharper patches in the adjacent frames and aggregating them into the latent frame. Most existing methods [12, 18, 26] only focus on local spatial features, which are limited by the preceding alignment results. To mitigate the problem, we explore the global guidance in the aggregation via the Fourier transform and find that the sharper region usually contains more considerable spectrum energies, as shown in Prior 3 of Section 3.2. Therefore, we equip the temporal spectrum energy into the aggregation to estimate the sharpness degree of inter-frames for more effective temporal fusion. Specifically, we develop the Temporal Energy Attention (TEA) fusion shown in Figure 9, which combines the advantage of the spatial and frequency domains. For the frequency domain, we take the energy spectrum of multiple frames as a classifier to judge the importance of these frames. The frequency attention map can be computed as:

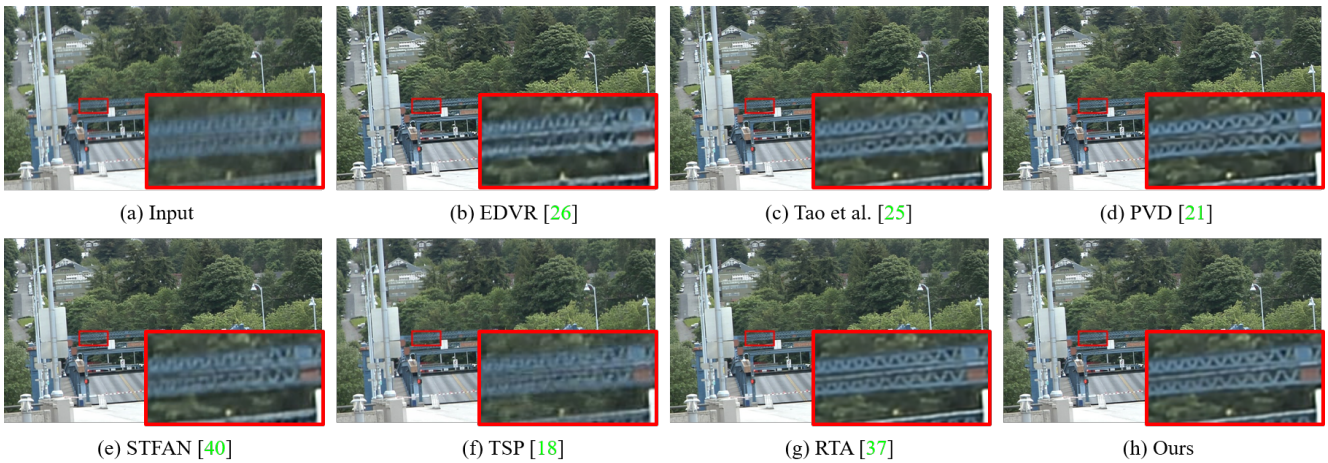
$$A_{t+i}^{spec} = \frac{e^{\|f(F_{t+i}^{align})\|}}{\sum_{i=-N}^{N} e^{\|f(F_{t+i}^{align})\|}}, \tag{9}$$

where $F_{t+i}^{align}$ denote the aligned feature in the $(t+i)$-th time. Then, to aggregate more local textures and details, the spatial attention as [26] is performed along with the temporal energy attention. Finally, the aligned feature is aggregated under the guidance of the spatial and frequency information, and the output $F_t^{fuse}$ is obtained.

### 4.4. Optimization

For video deblurring, we consider two loss functions to measure the difference between the restored image $I_{t+i}$ and sharp image $S_{t+i}$ from two perspectives, i.e., spatial domain and frequency domain. In the spatial domain, we adopt the Charbonnier penalty function [9] as the spatial loss to focus on the pixel-wise details for restoration, which is defined as:

$$L_{spa} = \sqrt{(I_t - S_t)^2 + \varepsilon^2}, \tag{10}$$



(a) Input (b) EDVR [26] (c) Tao et al. [25] (d) PVD [21]

(e) STFAN [40] (f) TSP [18] (g) RTA [37] (h) Ours

Figure 10. Qualitative comparison on the real blurred videos [22]. The deblurred results in (b)-(g) are still not clear. The proposed method removes the blur and generates better details.

where $\varepsilon$ is set to $1\times10^{-3}$ and $||.||$ denotes the L2 norm. Motivated by Figures 1 and 2, we find that the spectrum can not only contain high and low-frequency signals, but also pull apart the difference in terms of blur degrees. Therefore, to further recover high-frequency details, the frequency loss consisting of $L_{spec}$ and $L_{ener}$ is developed for the supervision from the ground-truth amplitude and energy spectrums during the training stage, which is defined as:

$$L_{spec} = \sqrt{(F(I_i) - F(S_i))^2}, \tag{11}$$

$$L_{ener} = Norm[F(I_i)] - Norm[F(S_i)] \tag{12}$$

where $F$ and $Norm$ denote the Fourier transform and L2-norm operator. The overall loss function for deblurring is:

$$L_{all} = L_{spa} + \lambda(L_{spec} + L_{ener}), \tag{13}$$

where $\lambda$ is the weight factor and is set to 0.1 empirically.

# 5. Experiments

## 5.1. Implementation Details

We randomly crop a $256 \times 256$ patch from each image and set the mini-batch size for each GPU as 4. The channel size in each residual block is set to 128. The network takes five consecutive frames (*i.e.*, $N = 2$) as inputs unless otherwise specified. We augment the training data with random horizontal flips and $90^\circ$ rotations. We train our model with Adam optimizer [7] by setting $\beta_1$ =0.9 and $\beta_2$ =0.999. The learning rate is initialized as $1 \times 10^{-4}$ and decayed by 50% every 200K iterations. The proposed network converges after 900K iterations. We implement our model with the PyTorch framework and train it using 4 NVIDIA 3090 GPUs. The proposed network is trained on the DVD dataset [22], which contains 71 videos (6,708 blurred-sharp pairs). The 61 videos (5,708 pairs) for training and 10 videos (1,000 pairs) for testing.

## 5.2. Comparison with State-of-the-Arts

We evaluate our network against several state-of-the-art methods, including one conventional model [6] and some CNN-based techniques [18, 20, 21, 25, 26, 28, 37, 40]. On the DVD dataset [22], our network excels the others in terms of PSNR and SSIM (see Table 1). A streamlined version of our model, with feature channels reduced to 64, matches the leading methods but is more size-efficient. Figure 8 highlights our deblurring outcomes. The method by Tao et al. [25] often faces challenges in rendering textures and details accurately. Similarly, techniques proposed by Xiang et al. [28] and Pan et al. [18] have their distinct limitations. In contrast, our method not only proficiently removes blur but also preserves fine details with aplomb. To further demonstrate our method's capabilities, we tested it on real-world blurry videos from Su et al. [22]. The results, evident in Figure 10, show our approach's superior performance, especially in addressing significant blurs and in restoring intricate details.

# 6. Ablation Studies and Discussion

In this section, we briefly discuss the effectiveness of the main proposed methods for lack of space. More details and discussions are presented in supplementary materials to confirm the effectiveness of frequency-temporal modeling.

## 6.1. Effectiveness of the TEA Fusion Module

We develop the TEA fusion module to mine the sharp patches from neighboring frames. To better understand this module, we visualize the learned frequency attention map, as shown in Figure 11. The sharp frame (top left) is paid more attention, and the significantly blurry frame (top right) has been less exploited. To further prove the effectiveness of the TEA fusion module, we perform ablation experiments,

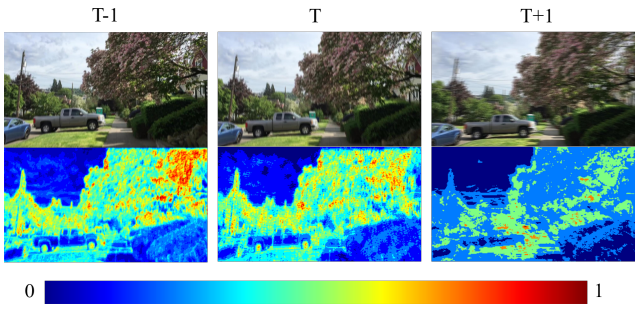



Figure 11. Visualization in the TEA. The top row is a video piece, while the bottom is the corresponding attention map. It is shown that the sharper frame (left) obtains more attention and vice versa.

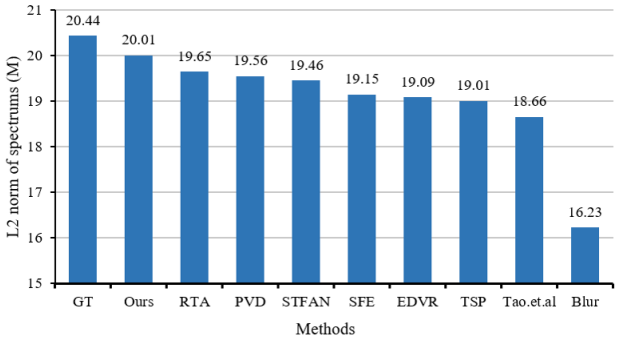



Figure 12. The spectrum norm comparison with existing methods, where M denotes million. Our method achieves the closest results to the ground truth among all methods.

as depicted in Table 2. The last two columns indicate our proposed fusion strategy can assist the model in leveraging the information from the adjacent frames better than direct concatenating and using single frequency and single spatial attentions.

### 6.2. Effectiveness of the SPA Module

To validate the effectiveness of the SPA module, we perform the ablations, as shown in Table 3. Due to the introduction of the frequency spectrum and the pyramid strategy, we achieve better performance in terms of PSNR and SSIM.

### 6.3. Loss Investigation

We present the qualitative results of the proposed method by diversely combining the frequency loss functions in Table 4. The results without the frequency loss achieve poor performance due to the lack of frequency-aware constraint. Meanwhile, the large weight for frequency loss may not bring better results. The frequency loss for the global structure and the spatial loss for the local texture are mutually reinforcing to supervise the deblurring jointly. As shown in Figure 12, our deblurring results have closer frequency distribution to the ground truth than the other methods.

Table 2. Comparison of different types of attentions on the blurry videos of DVD. $A$, $A_s$, $A_f$, and $A_{fs}$ represent "without attention", "with spatial attention", "with frequency attention", and "with frequency-spatial attention", respectively.

| Method | wo/$A$ | w/$A_s$ | w/$A_f$ | w/$A_{fs}$ |
|---|---|---|---|---|
| PSNR↑ | 32.74 | 33.05 | 33.11 | **33.25** |
| SSIM↑ | 0.9314 | 0.9426 | 0.9452 | **0.9491** |

Table 3. The effectiveness of the spectrum prior alignment. "Ms" and "Fs" mean the multi-scale and frequency spectrum modulating operations, respectively.

| Method | wo/SPA | wo/Ms | wo/Fs | w/SPA |
|---|---|---|---|---|
| PSNR↑ | 32.51 | 32.91 | 33.03 | **33.25** |
| SSIM↑ | 0.9296 | 0.9341 | 0.9415 | **0.9491** |

Table 4. Ablation studies about the loss weight in terms of PSNR and SSIM.

| $\lambda$ | 1 | 0.5 | 0.1 | 0.05 | 0 |
|---|---|---|---|---|---|
| PSNR↑ | 33.13 | 33.17 | **33.25** | 33.14 | 33.09 |
| SSIM↑ | 0.9484 | 0.9489 | **0.9491** | 0.9485 | 0.9478 |

## 7. Conclusion

Motion blur caused by hand-held devices is a major problem in video deblurring. In this work, based on the motion blur modeling, we probe into the temporal spectrum and find that the temporal frequency is beneficial to model the temporal information in video deblurring. To this end, we propose a Fourier-based Frequency-Temporal Network for video deblurring. Specifically, we devise the Spectrum Prior-guided Alignment for the different-range adjacent frame in a global-to-detail strategy. Then, the temporal energy attention is developed to effectively aggregate sharper scene patches from neighboring frames. Besides, the frequency losses are applied to reconstruct the latent frame with decent spectral distribution. Extensive experiments illustrate that our proposed model performs favorably against previous state-of-the-art methods, confirming its contribution to frequency-temporal modeling for video deblurring.

## Acknowledgments

This work was supported by the JKW Research Funds under Grant 20-163-14-LZ-001-004-01, and the Anhui Provincial Natural Science Foundation under Grant 2108085UD12. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

## References


[1] Leah Bar, Benjamin Berkels, Martin Rumpf, and Guillermo Sapiro. A variational framework for simultaneous motion estimation and restoration of motion-blurred video. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007. 1

[2] Sunghyun Cho, Jue Wang, and Seungyong Lee. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Transactions on Graphics*, 31(4):1–9, 2012. 1

[3] Mauricio Delbracio and Guillermo Sapiro. Burst deblurring: Removing camera shake through Fourier burst accumulation. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2385–2393, 2015. 4

[4] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *ACM SIGGRAPH*, page 787–794. 3

[5] Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, and Zhiwei Xiong. Learning sample relationship for exposure correction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9904–9913, June 2023. 3

[6] Tae Hyun Kim and Kyoung Mu Lee. Generalized video deblurring for dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5426–5434. 1, 7

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 7

[8] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 233–240, 2011. 3

[9] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5835–5843, 2017. 6

[10] Chongyi Li, Saeed Anwar, and Fatih Porikli. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition [Pattern Recognition Best Paper Honourable Mention]*, 98:107038, 2020. 3

[11] Chongyi Li, Jichang Guo, and Chunle Guo. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal Processing Letters*, 25(3):323–327, 2018. 3

[12] 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. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7717–7727, 2021. 3, 6

[13] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Dynamic scene deblurring by depth guided model. *IEEE Transactions on Image Processing*, 29:5273–5288, 2020. 3

[14] Zheng Liang, Weidong Zhang, Rui Ruan, Peixian Zhuang, and Chongyi Li. Gifm: An image restoration method with generalized image formation model for poor visible conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 3

[15] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 936–944, 2017. 6

[16] Xintian Mao, Yiming Liu, Wei lei Shen, Qingli Li, and Yan Wang. Deep residual Fourier transformation for single image deblurring. *ArXiv*, abs/2111.11745, 2021. 3

[17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–265, 2017. 3

[18] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3040–3048, 2020. 3, 6, 7

[19] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016. 3

[20] 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. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 7

[21] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM Transactions on Graphics*, 40(5):1–18, 2021. 7

[22] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 237–246, 2017. 6, 7

[23] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 3

[24] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 769–777, 2015. 3

[25] Xin Tao, Hongyun Gao, Yi Wang, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8174–8182, 2018. 3, 7

[26] Xintao Wang, Kelvin C. K. Chan, K. Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. *Workshop on Computer Vision and Pattern Recognition*, pages 1954–1963, 2019. 2, 3, 6, 7

[27] Xintao Wang, K. Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 606–615, 2018. 3

[28] Xinguang Xiang, Hao Wei, and Jinshan Pan. Deep video deblurring using sharpness features from exemplars. *IEEE Transactions on Image Processing*, 29:8976–8987, 2020. 7

[29] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2113–2122, 2021. 3

[30] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1107–1114, 2013. 3

[31] Tianfan Xue, Baian Chen, Jiajun Wu, D. Wei, and William T. Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2018. 3

[32] Jiawei Zhang, Jinshan Pan, Jimmy S. J. Ren, Yibing Song, Linchao Bao, Rynson W. H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2521–2529, 2018. 3

[33] Kaihao Zhang, Wenhan Luo, Björn Stenger, Wenqi Ren, Lin Ma, and Hongdong Li. Every moment matters: Detail-aware networks to bring a blurry image alive. *ACM International Conference on Multimedia*, pages 384–392, 2020. 3

[34] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Björn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2734–2743, 2020. 3

[35] Naishan Zheng, Jie Huang, Man Zhou, Zizheng Yang, Qi Zhu, and Feng Zhao. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3678–3686, 2023. 3

[36] Naishan Zheng, Jie Huang, Qi Zhu, Man Zhou, Feng Zhao, and Zheng-Jun Zha. Enhancement by your aesthetic: An intelligible unsupervised personalized enhancer for low-light images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6521–6529, 2022. 3

[37] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting temporal alignment for video restoration. *Computer Vision and Pattern Recognition*, pages 6043–6052, 2022. 2, 3, 7

[38] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*, pages 42589–42601, 2023. 3

[39] man zhou, Hu Yu, Jie Huang, Feng Zhao, Jinwei Gu, Chen Change Loy, Deyu Meng, and Chongyi Li. Deep Fourier up-sampling. In *Advances in Neural Information Processing Systems*, volume 35, pages 22995–23008, 2022. 3

[40] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. Spatio-temporal filter adaptive network for video deblurring. In *IEEE Conference on International Conference on Computer Vision*, pages 2482–2491, 2019. 7

[41] Qi Zhu, Zeyu Xiao, Jie Huang, and Feng Zhao. Dast-net: Depth-aware spatio-temporal network for video deblurring. In *2022 IEEE International Conference on Multimedia and Expo*, pages 1–6, 2022. 3

[42] Xizhou Zhu, Han Hu, Stephen Ching-Feng Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9300–9308, 2019. 2