

SelfHVD: Self-Supervised Handheld Video Deblurring

Honglei Xu, Zhilu Zhang*, Junjie Fan, Xiaohe Wu, Wangmeng Zuo
 Harbin Institute of Technology

cshongleixu@gmail.com, cszlzhang@outlook.com, {zhou93108, csxhwu, cswmzuo}@gmail.com

Abstract

Shooting video with handheld shooting devices often results in blurry frames due to shaking hands and other instability factors. Although previous video deblurring methods have achieved impressive progress, they still struggle to perform satisfactorily on real-world handheld video due to the blur domain gap between training and testing data. To address the issue, we propose a self-supervised method for handheld video deblurring, which is driven by sharp clues in the video. First, to train the deblurring model, we extract the sharp clues from the video and take them as misalignment labels of neighboring blurry frames. Second, to improve the deblurring ability of the model, we propose a novel Self-Enhanced Video Deblurring (SEVD) method to create higher-quality paired video data. Third, we propose a Self-Constrained Spatial Consistency Maintenance (SCSCM) method to regularize the model, preventing position shifts between the output and input frames. Moreover, we construct synthetic and real-world handheld video datasets for handheld video deblurring. Extensive experiments on these and other common real-world datasets demonstrate that our method significantly outperforms existing self-supervised ones. The code and datasets are publicly available at <https://cshonglei.github.io/SelfHVD>.

1. Introduction

Videos captured by handheld shooting devices generally suffer from significant blur due to shaking caused by hands, vibration caused by walking, and other instability factors. Modern shooting devices, such as smartphones, are often equipped with image stabilization technologies to alleviate this problem. Take commonly used Optical Image Stabilization (OIS) as an example, it has become a standard feature integrated by smartphone manufacturers, *e.g.*, Huawei, Xiaomi and Apple, which uses a Micro-Electro-Mechanical System (MEMS) gyroscope and gravity sensor to detect

* Corresponding author.

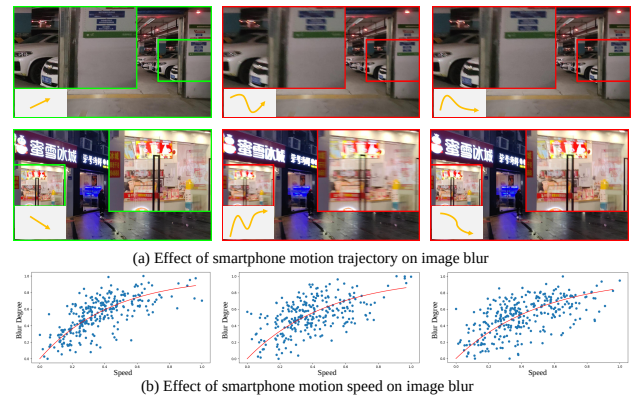


Figure 1. (a) Effect of the shooting device (with OIS) motion trajectory on image blur. The bottom left corner of the image shows a rough trajectory representation during the exposure time. The more complex the trajectory, the higher the probability of blurry frames (red box). Sharp frames (green box) can be captured when the trajectory is simple and even straight. (b) Effect of the shooting device (with OIS) motion speed on image blur. We randomly select 3 videos and count the correlation between the blur degree and the motion speed. The faster the speed, the higher the probability of blurring.

movement and adjust the imaging system. For example, if the shooting device is moved slightly to the left, the OIS would perceive this, and then move the lens module or imaging sensor slightly to the right accordingly. However, when the shooting device motion trajectory is complex or its motion speed is fast, the OIS may fail due to its own ability limitations and untimely response. In this case, blur may still appear in the video (see Fig. 1).

A mechanical solution is to use an additional gimbal stabilizer to clamp the shooting device to shoot. But it requires extra money costs and carrying space. Another solution is to post-process the blurry video, *i.e.*, video deblurring. It is relatively cheap and can be integrated into the shooting device system. Recently, learning-based image [8, 16, 24, 47, 54, 59, 60] and video deblurring methods [3, 6, 18, 21–23, 33, 36, 56, 57, 61] have made great progress, especially in network design. Nevertheless, their

pre-trained models are usually effective only on blurry data that is similar to the training samples, and their generalization ability is worrying. Specifically on the handheld video, the image blur is not only affected by camera shake, but also by OIS correction, so its blur distribution is significantly different from that in the existing training datasets (e.g., GoPro [31], BSD [62]). It leads the existing models to perform poorly on handheld video deblurring.

To address the issue, a straightforward idea is to collect paired handheld video deblurring datasets to train existing networks. But it is with high cost and the process of capturing and pre-processing datasets can be quite complex and cumbersome. Fortunately, when the shooting device motion trajectory is simple (e.g., a straight line) and the speed is slow, OIS may work well, so that sharp frames can be obtained, as shown in Fig. 1. Sharp frames have the potential to provide deblurring clues and supervision for neighboring blurry frames. Thus, it may be feasible to learn a video deblurring model in a self-supervised manner, which can circumvent the need for paired data.

Specifically, in this work, we propose a self-supervised handheld video deblurring method, named SelfHVD. First, we divide the video into multiple segments and select at least one sharpest frame from each segment. Then we align the frame with the neighboring blurry ones, thus the aligned sharp frame can be taken as supervision of the video deblurring model. Second, using the above strategy is not enough, as the upper limit of the deblurring model trained in this way is only the selected sharp frame that may be suboptimally clear or not cover sufficient blurry areas. To further improve the deblurring model, we propose a novel Self-Enhanced Video Deblurring (SEVD) method, which utilizes deblurring ability of the previously trained model to construct higher-quality paired training data. In the paired data, SEVD takes the blurry video with some sharp clues removed as the input, and uses the deblurring results (generated from the blurry video with sharp clues) or the selected sharp frame as supervision. In this way, the model can learn to produce results that surpass the sharpest frame in the input segment. Third, as the number of training iterations increases, the misalignment between supervision and input leads to the spatial position shift between output and input. To avoid this issue, we propose a Self-Constrained Spatial Consistency Maintenance (SCSCM) method. SCSCM constrains the current deblurring results from being spatially consistent with the results generated by the earlier deblurring model, as we observe that early deblurring models generally do not yet suffer from the position shift issue. Moreover, SCSCM can help optimize the model more steadily as training progresses.

For method validation and evaluation, we construct a synthetic dataset (GoProShake) and a real-world dataset (HVD) by collecting hundreds of real-world handheld

videos with HUAWEI P40. Extensive experiments on these two and other common real-world datasets. We note some self-supervised deblurring methods [11, 52] designed for general blurry video. They usually train the deblurring model by constructing paired data from blurry video, where the sharp information from the video is taken as supervision and is blurred as input. However, their synthetic blurry videos still differ from real-world blurry ones, which prevents their models from performing satisfactorily. From the experimental results, our SelfHVD achieves a significant improvement over them.

The contributions are summarized as follows.

- Based on the observation that sharp clues exist in handheld blurry video, we explore a self-supervised method for handheld video deblurring.
- We propose a novel Self-Enhanced Video Deblurring (SEVD) method to improve the deblurring ability of the model and a Self-Constrained Spatial Consistency Maintenance (SCSCM) method to regularize the model to prevent position shift between the output and input.
- We construct a synthetic dataset (GoProShake) and a real-world dataset (HVD) for handheld video deblurring. Extensive experiments on these two and other common real-world datasets demonstrate that our method significantly outperforms existing self-supervised ones.

2. Related Work

2.1. Supervised Image Deblurring

Traditional image deblurring techniques often utilize variational optimization [9, 15, 17, 30], which depend on prior assumptions about blur kernels and images to tackle the ill-posed nature of the inverse problem. With the rise of deep learning, substantial progress has been achieved [5, 20, 31, 38, 55]. Nah *et al.* [31] proposed a CNN-based model to deblur without blur kernel estimation. Chen *et al.* [5] introduced a simple but effective baseline network for image deblurring. Restormer [55] utilized a transformer-based architecture to restore images. Furthermore, SwinIR [20] designed networks based on the Swin Transformer [27]. For processing unknown blur, Blur2Blur [38] proposed to transform a blurry image into another image with known blur, thus being more amenable for deblurring.

2.2. Supervised Video Deblurring

Several synthetic [31, 32, 44] and real-world [39, 62] datasets have been used to train supervised video deblurring models. Compared to image deblurring, video deblurring can leverage spatio-temporal information within videos to enhance model performance. On the one hand, several methods [3, 18, 23, 36, 56, 61] have employed RNN-based models to leverage spatio-temporal information in videos. IFIRNN[61] iteratively updates the hidden

state via reusing RNN cell parameters. ESTRNN [61] employs a GSA module to catch spatially and temporally varying blurs. BasicVSR++ [3] adopts aggressive bidirectional propagation. STDAN [56] and FGST [23] utilize flow-guided attention to align and fuse information from adjacent frames. DSTNet [36] develops a wavelet-based feature propagation technique to transfer features in the frequency domain. ShiftNet [18] proposes a grouped spatio-temporal shift operation to aggregate spatio-temporal features efficiently. On the other hand, several studies [21, 22, 57] have explored Transformer-based architectures for video deblurring. VRT [21] utilizes a spatio-temporal self-attention mechanism to integrate information across video frames. RVRT [22] proposed a recurrent video restoration transformer with guided deformable attention. BSSTNet [57] converts the originally dense attention into a sparse form, enabling a more extensive utilization of information throughout the entire video sequence. In addition, some methods [41, 42, 45] utilize sharp frames in input videos to improve video deblurring. Although these supervised video deblurring methods have impressive results on corresponding datasets, they still perform poorly on real-world videos with unseen blur, *e.g.*, handheld blurry ones.

2.3. Self-Supervised Deblurring

Paired real-world data is difficult to obtain, thus some methods [7, 11, 25, 34, 51, 52] have proposed to learn video deblurring models in a self-supervision manner. Chi *et al.* [7] build a self-supervised auxiliary reconstruction task that shares a portion of the network with the primary deblurring task. Motivated that an ideal deblurring result should contain zero-magnitude motion blur that is hard to be amplified, Nah *et al.* [34] proposed a novel reblurring loss to make the result sharper. Ren *et al.* [52] suggested blurring sharp frames in the video using randomly generated blur kernels to obtain paired data for training the model. Furthermore, Liu *et al.* [25] utilized GAN [10] to optimize a blurring model in an unpaired training manner, and DaDeblur [11] used a diffusion-based blurring model [50] to blur sharp images for fine-tuning deblurring model. However, artificially blurry images still differ from real blurry ones.

3. Proposed Method

3.1. Selecting Sharp Frames as Supervision

Characteristic of Handheld Video. Image stabilization technologies, such as Electronic Image Stabilization (EIS) and Optical Image Stabilization (OIS) have been commonly used in modern shooting devices. They first obtain the motion information from the Inertial Measurement Unit (IMU) (including gyroscope and gravity sensor), then perform attitude calculation [2, 19, 28, 40] on the IMU data. During the calculation, Kalman filter [12], Mahony filter [29], and

others are used to smooth the current data according to historical data. Finally, motion compensation is carried out according to acquired motion information to achieve image stabilization. Such stabilization technologies usually work well when the shooting device motion trajectory is simple and the speed is slow, but they may fail when the trajectory is complex or the speed is fast. Therefore, blurry frames and sharp frames often coexist in handheld videos, which gives us an opportunity to explore self-supervised methods for handheld video deblurring, with further discussion in Sec. A of the supplementary material.

Selecting Sharp Frames. A straightforward idea is to select these sharp frames as supervision of the video deblurring model. As the sharp frame detection approach suggested by Ren *et al.* [52], the variance of the image Laplacian can be considered as a measurement of sharpness degree. Given an image \mathbf{I} , the variance of its Laplacian is:

$$v_l(\mathbf{I}) = \mathbb{E}[(\Delta\mathbf{I} - \overline{\Delta\mathbf{I}})^2], \quad (1)$$

where $\Delta\mathbf{I}$ is the image Laplacian obtained by convolving \mathbf{I} with the Laplacian mask, and $\overline{\Delta\mathbf{I}}$ is the mean value of $\Delta\mathbf{I}$. Then, we construct a histogram of the sharpness degrees, and expect the histogram to exhibit a bimodal distribution, with the two peaks corresponding to sharp and blurry frames, respectively. To define the threshold that separates the two classes, we use an automatic image thresholding technique, *i.e.*, Otsu’s method [35], which minimizes the intra-class variance based on the histogram of sharpness levels. Frames with sharpness values below the threshold are classified as blurry ones, while those above the threshold are classified as sharp frames.

When applying the above selection method to the whole video, the selected sharp frames may be unevenly distributed, resulting in failure to cover most of the video scenes. Thus, we further split the video into segments, and for a video segment without the sharp frames, we regard the frame with the highest v_l as its sharp frame. Finally, we define these sharp frames determined by the above global and local selection steps as \mathbf{S} . In our implementation, we divide the video into segments of 20 frames, and the accuracy is computed by comparing our selected sharp frames with manually labeled ones, reaching 96.77% on GoProShake and 91.88% on HVD.

Taking Sharp Frames as Supervision. Given a handheld blurry video \mathbf{B} consisting of N frames $\mathbf{B}_{n=1}^N$, video deblurring aims to restore the corresponding sharp components $\mathbf{R}_{n=1}^N$. When taking the selected sharp frames \mathbf{S} as supervision, the optimal parameters $\Theta_{\mathcal{D}}$ of deblurring model \mathcal{D} can be formulated as,

$$\Theta_{\mathcal{D}}^* = \arg \min_{\Theta_{\mathcal{D}}} \mathcal{L}(\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}}), \mathbf{S}), \quad (2)$$

where \mathcal{L} denotes the loss function.

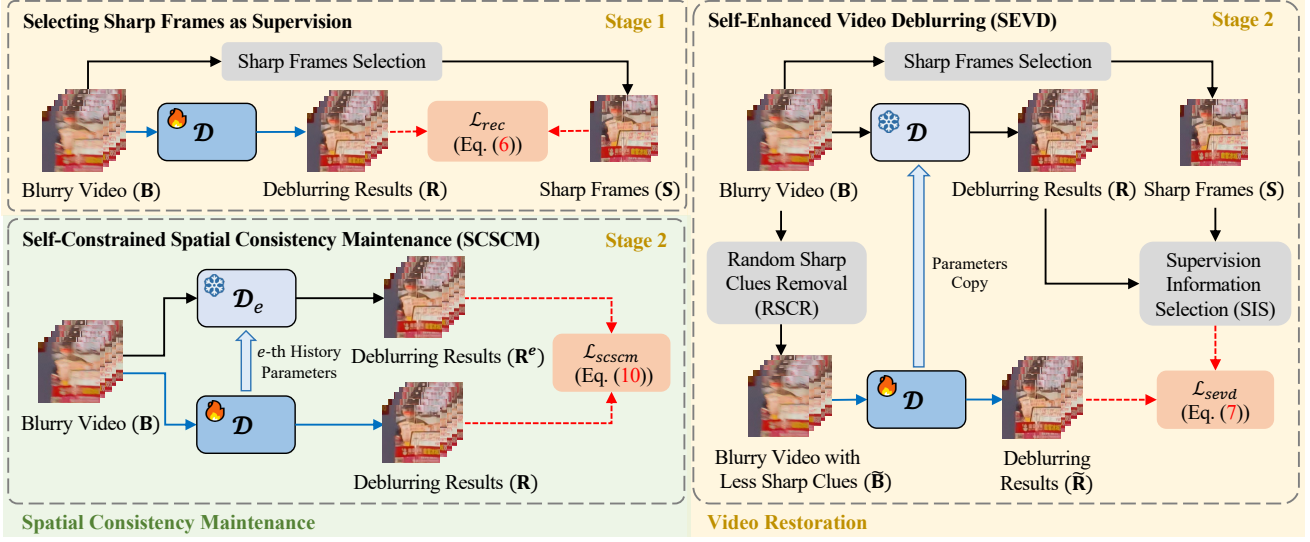


Figure 2. Overview of our SelfHVD. Given a blurry video captured by a handheld shooting device, we first select the sharp frames and take them as misalignment labels. Then, Self-Enhanced Video Deblurring (SEVD) constructs higher-quality paired training data to further improve the model performance. Self-Constrained Spatial Consistency Maintenance (SCSCM) is proposed to prevent position shifts between output and input frames.

Specifically, for the input frame \mathbf{B}_i , we identify the temporally closest sharp frame \mathbf{S}_j as its misalignment label, where $j = \mathcal{J}(i)$ and \mathcal{J} denotes the function that determines the temporally closest frame index. We use optical flow model (*i.e.*, SEA-RAFT [48]) to align output \mathbf{R}_i and label \mathbf{S}_j . The optical flow from \mathbf{R}_i to \mathbf{S}_j can be denoted as $\Phi_{i \rightarrow j}$. Then, the sharp frame \mathbf{S}_j can be backward warped via bilinear resampling, *i.e.*,

$$\mathbf{S}_{j \rightarrow i} = \mathcal{W}(\mathbf{S}_j, \Phi_{i \rightarrow j}), \quad (3)$$

where \mathcal{W} is the warping operation. Moreover, we design two masks to exclude incorrect alignment and occluded regions, respectively. The former mask can be estimated by the uncertainty map of the optical flow, and SEA-RAFT [48] can directly output it. Thus, the mask can be written as,

$$\mathbf{M}_{uncer}^i = \mathbf{U}_{j \rightarrow i} \odot \mathcal{W}(\mathbf{U}_{i \rightarrow j}, \Phi_{i \rightarrow j}), \quad (4)$$

where \odot is the pixel-wise multiplication operation. $\mathbf{U}_{j \rightarrow i}$ and $\mathbf{U}_{i \rightarrow j}$ represents the uncertainty map of $\Phi_{j \rightarrow i}$ and $\Phi_{i \rightarrow j}$, respectively. The latter mask can be estimated using the forward-backward flow consistency [1], and it can be written as,

$$\mathbf{M}_{occ}^i = \min(s \|\mathcal{W}(\mathcal{W}(\mathbf{G}; \Phi_{j \rightarrow i}); \Phi_{i \rightarrow j}) - \mathbf{G}\|_2, \mathbf{1}), \quad (5)$$

where \mathbf{G} is an image coordinate map. The scaling factor s controls the strength of the occlusion map. Finally, the reconstruction loss \mathcal{L}_{rec} can be formulated as,

$$\mathcal{L}_{rec} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{M}_i \odot (\mathbf{R}_i - \mathbf{S}_{j \rightarrow i})\|_1, \quad (6)$$

where $\mathbf{R}_i = \mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_i$ and $\mathbf{M}_i = \mathbf{M}_{uncer}^i \odot \mathbf{M}_{occ}^i$, as validated through ablation studies in Sec. D.1 and further shown in Fig. J and Fig. K in the supplementary material.

3.2. Self-Enhanced Video Deblurring

As previously described, we select the relatively sharp frames from each video segment as supervision of the deblurring model. However, it may not be enough, as the upper limit of the deblurring model trained in this way is only the selected sharp frame, while the sharp frame may be suboptimally clear, and the aligned sharp frame may not cover sufficient blurry areas. To address this issue, we propose the Self-Enhanced Video Deblurring (SEVD) method to utilize the existing deblurring ability of the model to construct higher-quality paired training data. This not only improves the overall deblurring performance but also enables the model to handle object motion blur.

First, we use a Random Sharp Clues Removal (RSCR) strategy to randomly remove the sharp clues from the input video and replace them with adjacent blurry frames. Specifically, we use the method proposed in **Selecting Sharp Frames** (Sec. 3.1) to distinguish between blurry and sharp frames in every video segment. Define the number of sharp frames as L , then we randomly replace l ($0 < l \leq L$) sharp frames with the temporally closest blurry frames to obtain the video $\tilde{\mathbf{B}}$, which has less sharp clues. Thus, the sharp frame \mathbf{S} is with higher quality than the clearest frame in $\tilde{\mathbf{B}}$. Taking \mathbf{S} as supervision of input video $\tilde{\mathbf{B}}$ can help model to break through its own input (*i.e.*, $\tilde{\mathbf{B}}$) to learn better results.

Second, since the selected sharp frame may not cover sufficient blurry areas and some aligned areas will be ex-



Figure 3. From left to right: sharp-clues-less blurry video, deblurring result of sharp-clues-less blurry video, warped sharp frame, deblurring result of original input video, occlusion map. From top to bottom, the unmasked region ratio in \mathbf{M}_{occ}^i is 0.90, 0.64, and 0.16. The smiley denotes the final supervision for $\mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$.

cluded (see Eq. (6)) due to inaccurate optical flow and occlusion, only using sharp frames as supervision may be not effective enough. In fact, the result $\mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$ (*i.e.*, deblurring the original input video $\tilde{\mathbf{B}}$) may be not less clear than \mathbf{S} . More importantly, $\mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$ is perfectly aligned with the corresponding frame in $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$. Thus, we can further take $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$ as the target of $\mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$.

In practice, in order to use higher-quality images as supervision of input $\tilde{\mathbf{B}}$, we suggest a Supervision Information Selection (SIS) strategy to select a better one from \mathbf{S} and $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$. Specifically, if the aligned sharp frame $\mathbf{S}_{j \rightarrow i}$ does not exhibit excessive distortion due to content difference between \mathbf{S}_j and $\mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$, and is sharper than the corresponding $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$, as shown in the top row of Fig. 3, we take $\mathbf{S}_{j \rightarrow i}$ as the supervision. It is noted that we consider distortion to have occurred when the proportion of unmasked regions in the occlusion map \mathbf{M}_{occ} is below a distortion threshold. Otherwise, as shown in the middle and bottom row of Fig. 3, $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$ is better than the aligned sharp frame, thus we choose $\mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$ to serve as supervision. Finally, \mathcal{L}_{sevd} can be defined as

$$\mathcal{L}_{sevd} = \frac{1}{N} \sum_{i=1}^N \begin{cases} \left\| \mathbf{M}_i \odot (\tilde{\mathbf{R}}_i - \mathbf{S}_{j \rightarrow i}) \right\|_1 & \text{if } c \text{ is True,} \\ \left\| \tilde{\mathbf{R}}_i - sg(\mathbf{R}_k) \right\|_1 & \text{if } c \text{ is False,} \end{cases} \quad (7)$$

$$c = \text{mean}(\mathbf{M}_{occ}^i) > \tau \text{ and } v_l(\mathbf{S}_{j \rightarrow i}) > v_l(\mathbf{R}_k), \quad (8)$$

where $\tilde{\mathbf{R}}_i = \mathcal{D}(\tilde{\mathbf{B}}; \Theta_{\mathcal{D}})_i$, $\mathbf{R}_k = \mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}})_k$, k is the frame index of \mathbf{R} corresponding to i -th frame in $\tilde{\mathbf{R}}$, $\text{mean}(\mathbf{M}_{occ}^i)$ is the unmasked regions proportion, τ is a threshold, $v_l(\cdot)$ represents sharpness (see Eq. (1)), and $sg(\cdot)$ is the stop gradient operation.

SEVD effectively enhances the performance in removing camera motion blur. Beyond that, it also enables the

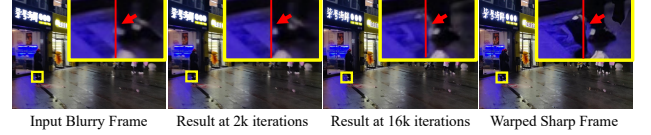


Figure 4. Vanilla self-supervised training performs well in early stages, but fails to maintain spatial consistency between input and output as training progresses.

restoration of object motion blur. Specifically, object motion is typically non-uniform, so relatively sharp content is retained when the object is still or moves slowly. Video deblurring models can aggregate information across multiple frames, allowing these sharp contents to provide crucial clues for dealing with blur when the object moves fast. Compared to deblurring results from the blurry video without sharp clues, those from the original blurry video perform better on object motion blur, as illustrated in the middle row of Fig. 3 and Fig. E in the supplementary material. As a result, the high-quality training pairs constructed by SEVD also offer more reliable supervision for object motion deblurring.

3.3. Self-Constrained Spatial Consistency Maintenance

Although we have carefully designed the self-supervised deblurring method, the spatial inconsistency issue between output and input is prone to occur as training progresses, as shown in Fig. 4 (c). In fact, it is hard to make sure of perfect alignment between the aligned sharp frames and the input frames, even using the most advanced optical flow network. Once alignment errors occur, even small ones, the model can gradually learn to shift the position of the input, bringing adverse effects on subsequent model training.

Fortunately, information bottleneck theory [46] has indicated that mutual information between network characteristics and the input increases first and then decreases during training [26]. As illustrated in Fig. 4, in the early stage, spatial consistency can be maintained; in the later stage, the inconsistency issue starts to occur. Inspired by this, we propose Self-Constrained Spatial Consistency Maintenance (SCSCM), using the historical result as auxiliary supervision to help the deblurring model maintain spatial consistency between the input and output.

Specifically, we denote the historical model parameters at e -th training iterations as $\Theta_{\mathcal{D}_e}$. The output \mathbf{R}^e from $\Theta_{\mathcal{D}_e}$ is both sharper than the input \mathbf{B} and aligned with \mathbf{B} , and it can be written as,

$$\mathbf{R}^e = \mathcal{D}(\mathbf{B}; \Theta_{\mathcal{D}_e}). \quad (9)$$

To ensure that the model parameters are updated under the supervision of sharper frames while remaining close to previously learned ones that implicitly preserve spatial align-

ment with the input, we incorporate SCSCM into the learning objective after e training iterations, defined as:

$$\mathcal{L}_{scscm} = \frac{1}{N} \sum_{i=1}^N \|\tilde{\mathbf{R}}_i - sg(\mathbf{R}_k^e)\|_1, \quad (10)$$

where e would be updated with the latest training #iterations if the output from the latest model is still aligned with the input and is sharper than the previous output.

3.4. Learning Objectives

To further improve the visual quality, we adopt the perceptual loss during training, which can be written as:

$$\mathcal{L}_{vgg1} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{M}_i \odot (\phi(\mathbf{R}_i) - \phi(\mathbf{S}_{j \rightarrow i}))\|_1, \quad (11)$$

$$\mathcal{L}_{vgg2} = \frac{1}{N} \sum_{i=1}^N \begin{cases} \|\mathbf{M}_i \odot (\phi(\tilde{\mathbf{R}}_i) - \phi(\mathbf{S}_{j \rightarrow i}))\|_1 & \text{if } c \text{ is True,} \\ \|\phi(\tilde{\mathbf{R}}_i) - sg(\phi(\mathbf{R}_k))\|_1 & \text{if } c \text{ is False,} \end{cases} \quad (12)$$

where ϕ is the pre-trained VGG [43] and c is the select condition (see Eq. (8)).

The overall training of our self-supervised handheld video deblurring model consists of two stages, and the learning objective in the first stage can be written as,

$$\mathcal{L}_{s1} = \mathcal{L}_{rec} + \beta \mathcal{L}_{vgg1}, \quad (13)$$

where β is the weight of \mathcal{L}_{vgg} and is set to 1. In the second stage, the model has had a certain deblurring ability, we introduce Self-Enhanced Video Deblurring (SEVD) to improve the quality of training data and deploy Self-Constrained Spatial Consistency Maintenance (SCSCM) to ensure spatial consistency between input and output. The learning objective can be written as,

$$\mathcal{L}_{s2} = \mathcal{L}_{sevd} + \lambda \mathcal{L}_{scscm} + \beta \mathcal{L}_{vgg2}, \quad (14)$$

where λ is the weight of \mathcal{L}_{scscm} and is set to 1.

4. Experiments

4.1. Implementation Details

Datasets. To verify the effectiveness of our method, we propose a synthetic handheld video deblurring dataset GoProShake, and a real-world dataset HVD. GoProShake simulates handheld motion with OIS using camera trajectories from MonST3R [58], while HVD is collected using HUAWEI P40 (HVD-Huawei), Xiaomi 15 (HVD-Xiaomi) and iPhone 16 (HVD-iPhone). Please refer to Sec. B in the supplementary material for detailed descriptions.

Framework Details. Note that this work does not focus on

the design of network architectures, and we employ existing ones directly. We adopt CNN-based (*i.e.* IFIRNN [33], ESTRNN [61], BasicVSR++ [3]) and Transformer-based (*i.e.* RVRT [22]) video deblurring models as the reconstruction network. We use the pre-trained SEA-RAFT [48] to estimate optical flow.

Training Details. During the training phase, the input frames are randomly cropped into patches with resolutions of 256×256, along with the application of random flipping and rotation. During the testing phase, the resolution of frames remains unchanged. The video deblurring model is optimized using Adam optimizer [14], where $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is $1e^{-4}$, gradually decayed to $1e^{-7}$ by the cosine annealing strategy. We set random seed to 0 and the distortion threshold τ to 0.5. All experiments are conducted with PyTorch [37] on a single Nvidia GeForce RTX A6000 GPU.

Evaluation Configurations. For the synthetic dataset GoProShake, we use PSNR, SSIM [49] as evaluation metrics. For the real-world dataset HVD, because there is no ground truth, we use no-reference image quality assessment MUSIQ [13] and MANIQA [53] as evaluation metrics. Additional evaluation details are provided in Sec. E of the supplementary material.

4.2. Comparison with State-of-the-Arts

Quantitative Analysis. Table 1 shows the quantitative comparison on the synthetic dataset GoProShake and the real-world dataset HVD. The results show that SelfHVD outperforms the previous self-supervised methods Ren *et al.* [52] and DaDeblur [11]. Moreover, SelfHVD achieves results comparable to the corresponding supervised methods on GoProShake. For results of supervised pre-training on the other real-world dataset along with their self-supervised fine-tuning performance on HVD, please refer to Sec. D.5 in the supplementary material. To further verify the effectiveness of our method, we conduct additional comparisons with DaDeblur [11] on publicly available real-world datasets, including BSD [61], RBVD [4], and RealBlur [39]. As shown in Table 2, under the same test-time training setting on ESTRNN [61] as DaDeblur [11], our method consistently outperforms both the baseline (*i.e.*, ESTRNN [61] trained on GoPro [31]) and DaDeblur [11], with less training time (including data preprocessing). On BSD [61], our method consistently surpasses both the baseline and DaDeblur [11] under various exposure settings, demonstrating strong robustness. And on RealBlur [39] and RBVD [4], our approach also achieves the highest PSNR and SSIM. These experimental results show that our method can achieve effective handheld video deblurring through a self-supervised approach without ground truth.

Qualitative Analysis. The visual comparisons on GoProShake and HVD datasets are shown in Fig. 5. As il-

Table 1. Quantitative comparison on the synthetic GoProShake and real-world HVD datasets. ‘*Network*’ in ‘SelfHVD_{*Network*}’ denotes the deblurring network we use, where ESTRNN [61] is also adopted by Ren *et al.* [52] and DaDeblur [11].

Methods	GoProShake	HVD-Huawei	HVD-Xiaomi	HVD-iPhone	
	PSNR↑ / SSIM↑	MUSIQ↑ / MANIQA↑	MUSIQ↑ / MANIQA↑	MUSIQ↑ / MANIQA↑	
Fully-Supervised	IFIRNN [33]	34.66 / 0.9448	24.1043 / 0.1916	29.7281 / 0.2212	22.3710 / 0.2535
	ESTRNN [61]	34.19 / 0.9369	24.0383 / 0.1917	29.4117 / 0.2193	21.6487 / 0.2506
	RVRT [22]	37.02 / 0.9473	24.9269 / 0.1923	30.2976 / 0.2215	22.4535 / 0.2546
	BasicVSR++ [3]	37.99 / 0.9683	25.2499 / 0.2006	30.0775 / 0.2235	22.6709 / 0.2564
Self-Supervised	Ren <i>et al.</i> [52]	25.05 / 0.7428	22.5433 / 0.1771	22.7757 / 0.2193	20.2299 / 0.2653
	DaDeblur [11]	29.54 / 0.8772	26.8422 / 0.2025	32.3833 / 0.2322	25.3244 / 0.2684
	SelfHVD _{IFIRNN}	34.32 / 0.9302	27.6922 / 0.2137	32.7765 / 0.2530	25.7711 / 0.2756
	SelfHVD _{ESTRNN}	33.60 / 0.9216	27.6873 / 0.2126	32.4002 / 0.2574	25.7556 / 0.2842
	SelfHVD _{RVRT}	36.31 / 0.9300	27.8345 / 0.2088	32.4416 / 0.2303	25.8437 / 0.2711
	SelfHVD _{BasicVSR++}	37.44 / 0.9359	28.0040 / 0.2175	32.8564 / 0.2236	25.7022 / 0.2686

Table 2. Quantitative results on BSD [61], RealBlur [39] and RBVD [4] via test-time fine-tuning. Each group reports PSNR and SSIM. All methods use ESTRNN [61] as the deblurring network, with its GoPro-trained version serving as the baseline.

Methods	BSD-1ms8ms [61]	BSD-2ms16ms [61]	BSD-3ms24ms [61]	RealBlur [39]	RBVD [4]
	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑	PSNR↑ / SSIM↑
Baseline	25.57 / 0.747	24.64 / 0.726	26.01 / 0.748	25.87 / 0.773	24.47 / 0.725
+Blur2Blur [38]	25.64 / 0.750	24.72 / 0.728	26.01 / 0.749	26.16 / 0.800	24.45 / 0.725
+DaDeblur [11]	29.44 / 0.843	28.36 / 0.820	28.23 / 0.808	27.41 / 0.819	27.02 / 0.771
+Ours	31.01 / 0.873	29.00 / 0.832	29.31 / 0.825	28.76 / 0.848	27.69 / 0.785

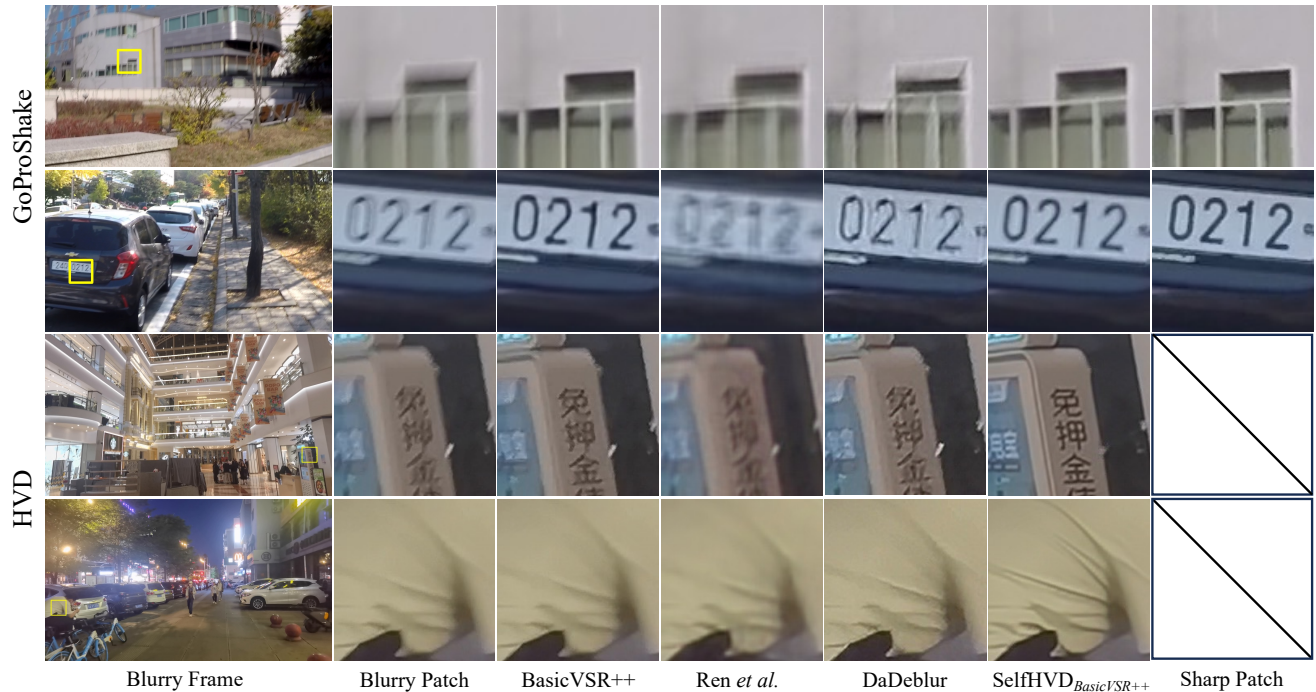


Figure 5. Qualitative comparison on synthetic GoProShake and real-world HVD datasets.

illustrated, SelfHVD_{BasicVSR++} achieves better deblurring results compared to previous self-supervised methods, even for object motion blur (Sec. C in the supplementary material). Moreover, the results of SelfHVD_{BasicVSR++} are visually comparable to BasicVSR++ [3] on GoProShake.

And under the same test-time training setting as DaDeblur [11], our method also achieves better visual results than DaDeblur [11] on BSD [61], RealBlur [39] and RBVD [4], as shown in Fig. 6. And additional visual results are in Sec. E of the supplementary material.

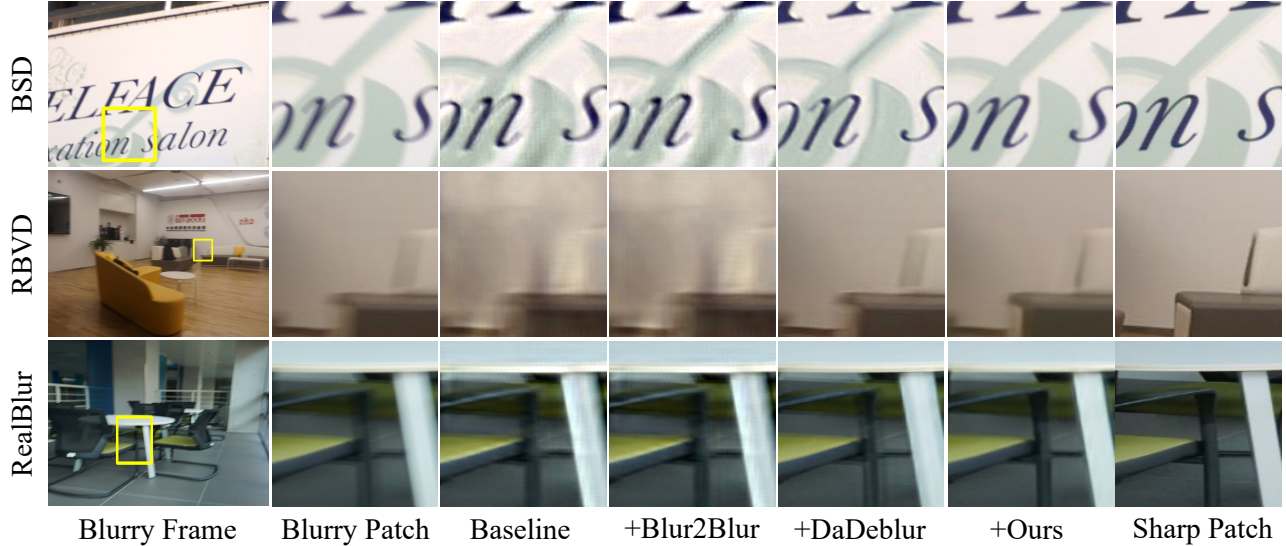


Figure 6. Qualitative results on BSD [61], RealBlur [39], and RBVD [4] datasets.

Table 3. Effect of SEVD and SCSCM.

SEVD	SCSCM	PSNR \uparrow / SSIM \uparrow
\times	\times	35.61 / 0.9263
\times	\checkmark	37.09 / 0.9342
\checkmark	\times	36.67 / 0.9290
\checkmark	\checkmark	37.44 / 0.9359

Table 4. Effect of RSCR and SIS in SEVD.

RSCR	SIS	PSNR \uparrow / SSIM \uparrow
\times	\times	35.04 / 0.9165
\times	\checkmark	35.11 / 0.9165
\checkmark	\times	36.27 / 0.9302
\checkmark	\checkmark	37.44 / 0.9359

Table 5. Effect of #update of D_e parameters in SCSCM.

#Update	PSNR \uparrow / SSIM \uparrow
1	36.00 / 0.9284
2	37.44 / 0.9359
3	37.57 / 0.9348

5. Ablation studies

To verify the effectiveness of each component, we conducted ablation studies with SelfHVD_{BasicVSR++}. Additional ablation studies are provided in Sec. D of the supplementary material, including analyses on mask design, SEVD and SCSCM module, optical flow models, sharp frame selection intervals, and supervised pre-training.

Effect of SEVD and SCSCM. To evaluate the individual and combined contributions of Self-Enhanced Video Deblurring (SEVD) and Self-Constrained Spatial Consistency Maintenance (SCSCM), we conduct an ablation study with four configurations, as shown in Table 3. The results demonstrate that introducing either SEVD or SCSCM independently leads to noticeable improvements in both PSNR and SSIM. Notably, the integration of both modules yields the highest performance, indicating their complementary effects in enhancing video deblurring quality.

Effect of RSCR and SIS in SEVD. We conduct an ablation study to evaluate the effectiveness of Residual Sharp Clue Removal (RSCR) and Supervision Information Selection (SIS), as shown in Table 4. RSCR brings improvement compared to randomly removing frames, confirming its targeted design is effective. SIS further improves performance by selecting the better supervision between aligned sharp frames and restored frames.

Effect of #Update of D_e Parameters. To further analyze the impact of #update of D_e parameters in SCSCM, we conducted additional experiments with varying #update, as shown in Table 5. The results indicate that increasing the #update from 1 to 2 brings a substantial improvement in both PSNR and SSIM. However, further increasing #update to 3 yields only marginal gains. We finally set the #update of D_e parameters to be 2.

6. Conclusion

Based on the observation that sharp clues frequently appear in blurry videos captured by handheld smartphones, we propose a self-supervised handheld video deblurring method, called SelfHVD. First, we extract sharp clues as supervision. Subsequently, Self-Enhanced Video Deblurring (SEVD) leverages the model’s existing deblurring ability to construct higher-quality and diverse paired training data, while Self-Constrained Spatial Consistency Maintenance (SCSCM) ensures the spatial alignment between input and output. Finally, we construct a synthetic dataset GoProShake and a real-world dataset HVD collected with a smartphone. Extensive experiments on these two datasets and the common real-world datasets demonstrate that SelfHVD significantly outperforms existing self-supervised ones.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62476067, the China Postdoctoral Science Foundation under Grant No. 2025M784371, and the OPPO Research Fund.

References

- [1] Luis Alvarez, Rachid Deriche, Théo Papadopoulo, and Javier Sánchez. Symmetrical dense optical flow estimation with occlusions detection. *IJCV*, 2007. 4
- [2] Steven Bell, Alejandro Troccoli, and Kari Pulli. A non-linear filter for gyroscope-based video stabilization. In *ECCV*, 2014. 3
- [3] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, 2020. 1, 2, 3, 6, 7
- [4] Zhu Chao, Dong Hang, Pan Jinshan, Liang Boyang, Huang Yuhao, Fu Lean, and Wang Fei. Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring. In *AAAI*, 2022. 6, 7, 8
- [5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 2
- [6] Lufei Chen, Xiangpeng Tian, Shuhua Xiong, Yinjie Lei, and Chao Ren. Unsupervised blind image deblurring based on self-enhancement. In *CVPR*, 2024. 1
- [7] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *CVPR*, 2021. 3
- [8] Jiangxin Dong, Jinshan Pan, Zhongbao Yang, and Jinhui Tang. Multi-scale residual low-pass filter network for image deblurring. In *ICCV*, 2023. 1
- [9] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *Acm Siggraph 2006 Papers*, 2006. 2
- [10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 3
- [11] Jin-Ting He, Fu-Jen Tsai, Jia-Hao Wu, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, and Yen-Yu Lin. Domain-adaptive-video-deblurring-via-test-time-blurring. In *ECCV*, 2024. 2, 3, 6, 7
- [12] Re Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 1960. 3
- [13] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 6
- [14] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv*, 2014. 6
- [15] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, 2011. 2
- [16] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 1
- [17] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Efficient marginal likelihood optimization in blind deconvolution. In *CVPR*, 2011. 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 *CVPR*, 2023. 1, 2, 3
- [19] Tzoo-Hseng S Li and Ching-Chang Chen. Nonlinear complementary filters on the special orthogonal group. *IEEE Transactions on Consumer Electronics*, 2013. 3
- [20] Jingyun Liang, Jiezhong Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 2
- [21] Jingyun Liang, Jiezhong Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv*, 2022. 1, 3
- [22] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhong Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. *arXiv*, 2022. 3, 6, 7
- [23] 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. In *ICML*, 2022. 1, 2, 3
- [24] Jingbo Lin, Zhilu Zhang, Wenbo Li, Renjing Pei, Hang Xu, Hongzhi Zhang, and Wangmeng Zuo. Unirestor: Universal image restoration via adaptively estimating image degradation at proper granularity. In *ICLR*, 2026. 1
- [25] Po-Sheng Liu, Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, and Yen-Yu Lin. Meta transferring for deblurring. In *BMVC*, 2022. 3
- [26] Xiaohui Liu, Zhilu Zhang, Xiaohe Wu, Chaoyu Feng, Xiaotao Wang, Lei Lei, and Wangmeng Zuo. Learning real-world image de-weathering with imperfect supervision. In *AAAI*, 2024. 5
- [27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2
- [28] Sebastian Madgwick et al. An efficient orientation filter for inertial and inertial/magnetic sensor arrays. *Report x-io and University of Bristol (UK)*, 2010. 3
- [29] Robert Mahony, Tarek Hamel, and Jean-Michel Pflimlin. Nonlinear complementary filters on the special orthogonal group. *IEEE Transactions on automatic control*, 2008. 3
- [30] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, 2014. 2
- [31] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 6
- [32] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPRW*, 2019. 2
- [33] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, 2019. 1, 6, 7

- [34] Seungjun Nah, Sanghyun Son, Jaerin Lee, and Kyoung Mu Lee. Clean images are hard to reblur: A new clue for deblurring. In *ICLR*, 2022. 3
- [35] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *Automatica*, 1975. 3
- [36] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *CVPR*, 2023. 1, 2, 3
- [37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 6
- [38] Bang-Dang Pham, Phong Tran, Anh Tran, Cuong Pham, Rang Nguyen, and Minh Hoai. Blur2blur: Blur conversion for unsupervised image deblurring on unknown domains. In *CVPR*, 2024. 2, 7
- [39] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 2, 6, 7, 8
- [40] Simone Sabatelli, Marco Galgani, Luca Fanucci, and Alessandro Rocchi. A double-stage kalman filter for orientation tracking with an integrated processor in 9-d imu. *IEEE Transactions on Instrumentation and Measurement*, 2012. 3
- [41] Wei Shang, Dongwei Ren, Dongqing Zou, Jimmy S Ren, Ping Luo, and Wangmeng Zuo. Bringing events into video deblurring with non-consecutively blurry frames. In *ICCV*, 2021. 3
- [42] Wei Shang, Dongwei Ren, Yi Yang, and Wangmeng Zuo. Aggregating nearest sharp features via hybrid transformers for video deblurring. *Information Sciences*, 2025. 3
- [43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 6
- [44] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 2
- [45] Yang Tian, Fabio Brau, Giulio Rossolini, Giorgio Buttazzo, and Hao Meng. Video deblurring by sharpness prior detection and edge information. *arXiv*, 2025. 3
- [46] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015. 5
- [47] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022. 1
- [48] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, 2024. 4, 6
- [49] 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*, 2004. 6
- [50] Jia-Hao Wu, Fu-Jen Tsai, Yan-Tsung Peng, Chung-Chi Tsai, Chia-Wen Lin, and Yen-Yu Lin. Id-blau: Image deblurring by implicit diffusion-based reblurring augmentation. In *CVPR*, 2024. 3
- [51] Renlong Wu, Zhilu Zhang, Mingyang Chen, Xiaopeng Fan, Zifei Yan, and Wangmeng Zuo. Deblur4dgs: 4d gaussian splatting from blurry monocular video. *arXiv*, 2024. 3
- [52] Qifeng Chen Xuanchi Ren, Zian Qian. Video deblurring by fitting to test data. In *arxiv*, 2020. 2, 3, 6, 7
- [53] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022. 6
- [54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1
- [55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 2
- [56] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *ECCV*, 2022. 1, 2, 3
- [57] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Blur-aware spatio-temporal sparse transformer for video deblurring. In *CVPR*, 2024. 1, 3
- [58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv*, 2024. 6
- [59] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, 2020. 1
- [60] Zhilu Zhang, Rongjian Xu, Ming Liu, Zifei Yan, and Wangmeng Zuo. Self-supervised image restoration with blurry and noisy pairs. In *NeurIPS*, 2022. 1
- [61] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8
- [62] Zhihang Zhong, Ye Gao, Yinqiang Zheng, Bo Zheng, and Imari Sato. Real-world video deblurring: A benchmark dataset and an efficient recurrent neural network. *IJVC*, 2023. 2