A Simple Baseline for Video Restoration with Grouped Spatial-temporal Shift

Dasong Li[1] Xiaoyu Shi[1] Yi Zhang[1] Ka Chun Cheung[2] Simon See[2]
Xiaogang Wang[1,4] Hongwei Qin[3] Hongsheng Li[1,4]

[1]CUHK MMLab [2]NVIDIA AI Technology Center [3]SenseTime Research [4]CPII under InnoHK
{dasongli@link, hsli@ee}.cuhk.edu.hk

Abstract

Video restoration, which aims to restore clear frames from degraded videos, has numerous important applications. The key to video restoration depends on utilizing inter-frame information. However, existing deep learning methods often rely on complicated network architectures, such as optical flow estimation, deformable convolution, and cross-frame self-attention layers, resulting in high computational costs. In this study, we propose a simple yet effective framework for video restoration. Our approach is based on grouped spatial-temporal shift, which is a lightweight and straightforward technique that can implicitly capture inter-frame correspondences for multi-frame aggregation. By introducing grouped spatial shift, we attain expansive effective receptive fields. Combined with basic 2D convolution, this simple framework can effectively aggregate inter-frame information. Extensive experiments demonstrate that our framework outperforms the previous state-of-the-art method, while using less than a quarter of its computational cost, on both video deblurring and video denoising tasks. These results indicate the potential for our approach to significantly reduce computational overhead while maintaining high-quality results. Code is avaliable at https://github.com/dasongli1/Shift-Net.

1. Introduction

The popularity of capturing videos using handheld devices continues to surge. However, these videos often suffer from various types of degradation, including image noise due to low-cost sensors and severe blurs resulting from camera shake or object movement. Consequently, video restoration has garnered significant attention in recent years.

The keys of video restoration methods lie in designing components to realize alignment across frames. While several methods [7, 38, 39, 53, 60] employ convolutional networks for multi-frame fusion without explicit alignment, their performance tends to be suboptimal. Most methods rely on explicit alignment to establish temporal correspondences, using techniques such as optical flow [46, 61] or deformable convolution [11, 69]. However, these approaches often necessitate either complex or computationally expensive network architectures to achieve large receptive fields, and they may fail in scenarios involving large displacements [27], frame noise [8, 63], and blurry regions [7, 48]. Recently, transformer [12, 15, 34] becomes promising alternatives for attaining long-range receptive fields. A video restoration transformer (VRT) [32] is developed to model long-range dependency, but its large number of self-attention layers make it computationally demanding. Inspired by the success of the Swin transformer [34], large kernel convolutions [14, 35] emerge as a direct solution to obtain large effective receptive fields. However, extremely large kernels (e.g. kernel size $> 13\times13$) does not necessarily guarantee improved performance. (shown in 5).

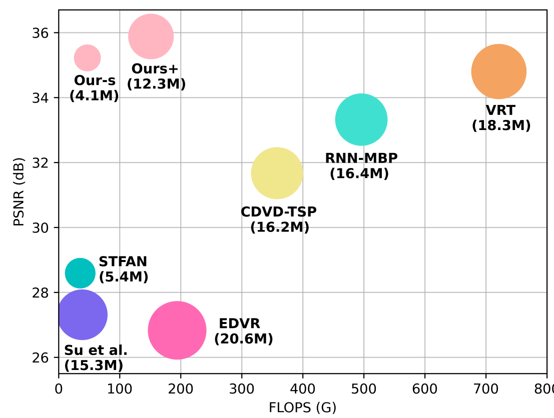

Figure 1. Video deblurring on GoPro dataset [40]. Our models have fewer parameters (disk sizes) and occupy the top-left corner, indicating superior performances (PSNR on y-axis) with less computational cost (FLOPS on x-axis).

In this study, we propose a simple, yet effective spatial-temporal shift block to achieve large effective receptive field for temporal correspondence. We introduce a Group Shift-Net, which incorporates the proposed spatial-temporal shift

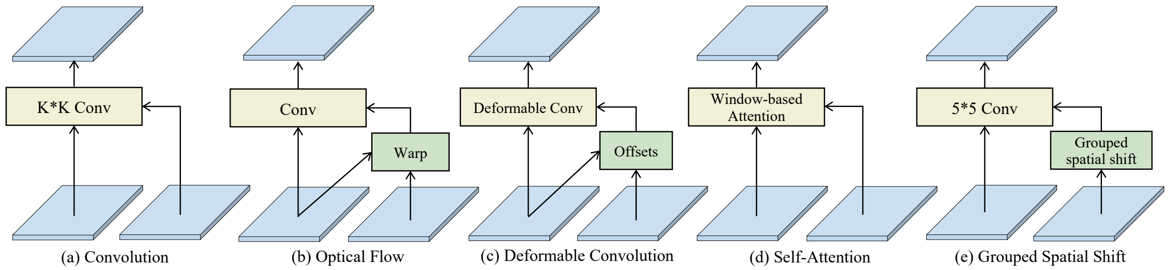

Figure 2. Different modules for multi-frame aggregation: a) convolution [53], b) optical flow [32,42], c) deformable convolution [11,54,57], d) self-attention [32, 34] and e) our grouped spatial shift. Point-wise convolution, shortcut and normalization are omitted for simplicity.

blocks for alignment along with basic 2D U-Nets for frame-wise feature encoding and restoration. The grouped spatial-temporal shift process involves the separate shifting of input clip features in both temporal and spatial dimensions, followed by fusion using 2D convolution blocks. Despite its minimal computational demands, the shift block offers large receptive fields for efficient multi-frame fusion. By stacking multiple spatial-temporal shift blocks, the aggregation of long-term information is achieved. This streamlined framework models long-term dependencies without depending on resource-demanding optical flow estimation [19,47,61], deformable convolution [11, 54, 57], or self-attention [32].

Notably, while temporal shift module (TSM) [33] was originally proposed for video understanding, it is not effective for video restoration. Our method distinguishes itself from TSM in three fundamental ways: a) *Alternative bi-directional temporal shift.* TSM [33] employs bi-directional *channel* shift during training, causing misalignment of channels across three frames, which in turn increases the difficulty of multi-frame aggregation. Conversely, our method utilizes alternative *temporal shifts*, effectively circumventing this issue. b) *Spatial shift.* In addition, our approach also incorporates a spatial shift for multi-frame features. We divide the features into several groups, each with distinct shift lengths and directions in the 2D dimension. This grouped spatial shift offers multiple candidate displacements for matching misaligned features. c) *Feature fusion.* To seamlessly merge various shifted groups, the kernel size of the convolution is set equal to the base shift length. By combining elements b) and c), the spatial-temporal shift achieves large receptive fields (e.g. 23×23).

The contributions of this study are two-fold: 1) We propose a simple, yet effective framework for video restoration, which introduces a grouped spatial-temporal shift for efficient and effective temporal feature aggregation 2) Our framework surpasses state-of-the-art methods with much fever FLOPs on both video deblurring and video denoising tasks, demonstrating its generalization capability.

2. Related Work

A series of methods have been proposed to explore *temporal information* for video restoration.

Temporal alignment. Temporal alignment is a vital step to model temporal correspondences of misaligned frames in videos. Early learning-based methods [2, 24, 29, 48, 52] employ traditional image alignment methods [58] to model the motions. To handle complicated motions, Xue et al. [61] propose task-oriented flow by fine-tuning a pretrained optical flow model [43] on different video restoration tasks. Dynamic filters [23, 68] are also proposed to achieve motion compensation. Tian et al. [54, 57] propose to utilize deformable convolution [11] for feature alignment. Chan et al. [5] leverage the optical flow to guide the deformable alignment for stable training [4], which is also adopt by the latest transformer-based method VRT [32]. Such alignment techniques increase the model complexity and might fail in the case of large displacement [27], noise [8, 63, 67], blurry regions [7, 48]. Zhu et al. [7] demonstates that optical flow or deformable convolution cannot estimate the alignment information well because of the significant influence of the motion blur. A series of methods [7, 38, 53] are proposed to utilize convolution networks to handle motion implicitly. However, the networks with small kernel sizes usually have narrow receptive fields [37], which limits the model capacity to address large displacements.

Long-term information aggregation. To obtain the long-term information from distant frames, learning-based methods can be classified as sliding window-based methods and recurrent methods. Sliding window-based methods [42, 53] usually take several adjacent frames as input and output the center restored frame. The information can only be aggregated within the fixed sliding window. In contrast, several methods [3, 5, 20, 38, 48] utilize the recurrent framework for long-term information aggregation. The faulty prediction and misalignment are accumulated frame by frame, which may deteriorate the long-term dependency modeling [6].

Shift operations. Wu et al. [59] combine shift operation and 1×1 convolution as an efficient alternative to 3×3 convolution. Its variants [10, 21] further propose learnable active shifts. Zhang et al. [66] adopt shift and 1×1 convolution for efficient image super-resolution. Lin et al. [33] propose a temporal shift module (TSM) for video understanding. Rong et al. [44] apply temporal shift on wavelet transforms for burst denoising. Liu et al. [34] perform self-

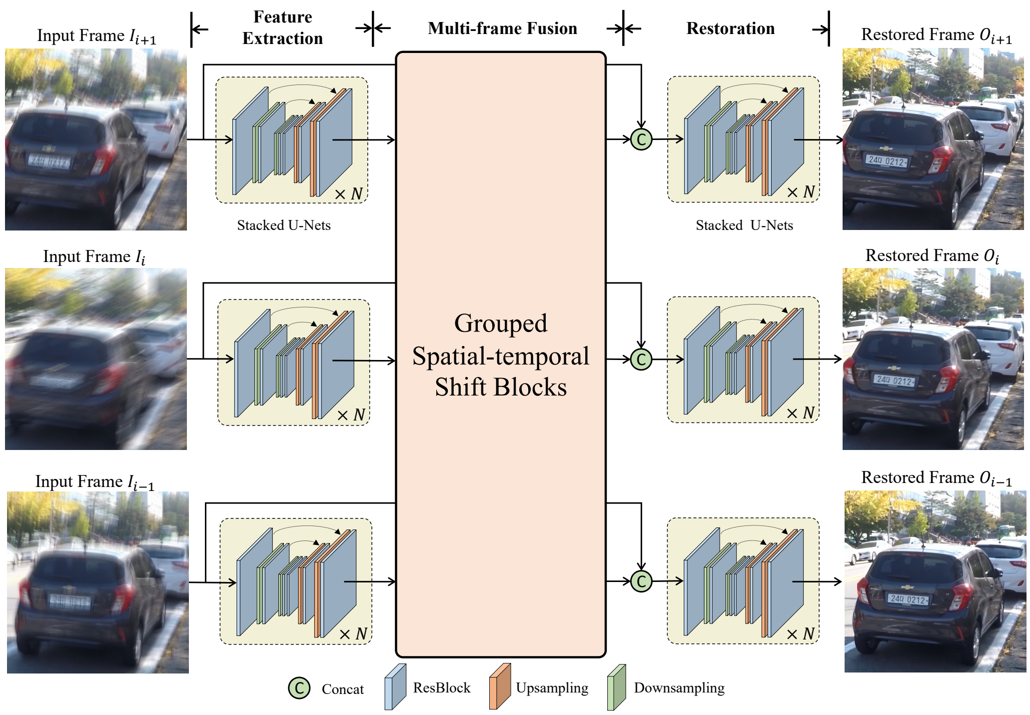

Figure 3. Overview of the Group Shift-Net. It adopts a three-stage design: feature extraction, multi-frame fusion, and final restoration. Grouped spatial-temporal shift blocks are proposed to achieve multi-frame aggregation.

attention with shifted windows to boost the performance of vision transformer [15]. Recently, a series of MLP-based architectures [31,56,62] couple the spatial shifts with multi-layer perceptron to achieve competitive performances in high-level visions tasks. Liang et al. [32] propose a video restoration transformer (VRT), where one video is partitioned into 2-frame clips at each layer and shifted for every other layer to perform temporal self-attention. However, it has a large number of self-attention layers and is computational costly. We extend shift operations to derive a large receptive field with small kernel convolutions.

3. Method

Most previous methods in video restoration adopt complicated architectures, such as optical flow [61], deformable convolution [11], and self-attention layers [32]. We propose a simple, yet effective grouped spatial-temporal shift block to establish temporal correspondences implicitly.

3.1. Overview of Group Shift-Net

Given consecutive degraded frames $\{I_i \in \mathbb{R}^{h\times w\times c_{in}}\}_i^T$, where T denotes the frame number, Group Shift-Net outputs the high-quality frames $\{O_i \in \mathbb{R}^{h\times w\times c_{out}}\}_i^T$. As shown in Fig 3, our framework adopts a three-stage design: 1) feature extraction, 2) multi-frame feature fusion with grouped spatial-temporal shift, and 3) final restoration.

Feature extraction. Each frame I_i usually suffers from different types of degradation (e.g. noise or blur), which affects temporal correspondences modeling. Inspired by [6], 2D U-Net-like structures [45] are adopted to mitigate negative impacts of degradation and extract frame-wise features.

Multi-frame feature fusion. At this stage, we propose a grouped spatial-temporal shift block to shift different features groups of neighboring frames to the reference frame to establish the temporal correspondences implicitly. The key-frame feature $f_i \in \mathbb{R}^{h\times w\times c}$ is fully aggregated with those of the neighboring frames to obtain the corresponding aggregated feature $A_i \in \mathbb{R}^{h\times w\times c}$. Spatial-temporal shifts of different directions and distances are adopted to provide multiple candidate displacements for matching the frames. By stacking multiple grouped spatial-temporal shift blocks, our framework can achieve long-term aggregation.

Final restoration. At last, U-Net-like structures take the low-quality input frames $\{I_i\}_i^T$ and corresponding aggregated features $\{A_i\}_i^T$ as input and produces each frame's final result O_i. The loss function L is formulated as

$$L = \frac{1}{T}\sum_{i=1}^{T} ||H_i - O_i||_1. \tag{1}$$

3.2. Frame-wise Processing

For feature extraction of stage 1 and final restoration of stage 3, we stack N 2D slim U-Nets consecutively to extract features and conduct restoration effectively. Stacking multiple U-Nets [41] was explored before, which leads to a deeper network depth and a larger receptive field than a single U-Net [64] with the same computational cost. At each

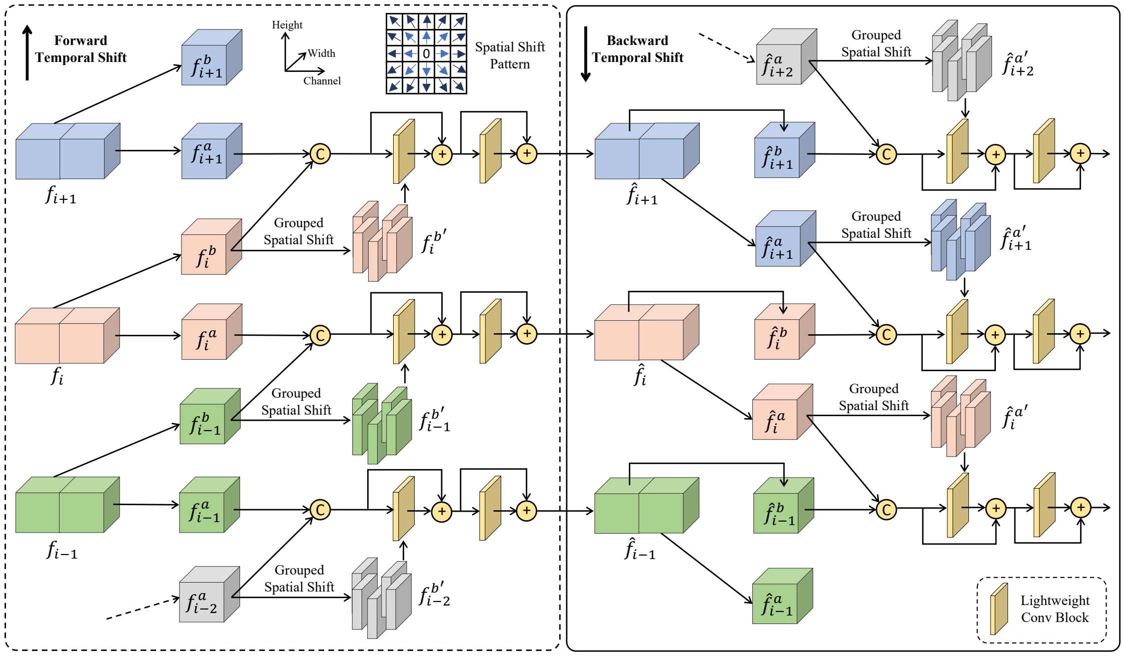

Figure 4. The operations of Grouped Spatial-temporal Shift (GSTS). We stack the forward temporal shift (FTS) blocks (*Left*) and backward temporal shift (BTS) blocks (*Right*) alternatively to achieve bi-directional propagation. Grouped spatial shift provides multiple candidate displacements within large spatial fields and establish temporal correspondences implicitly.

U-Net, we utilize residual blocks [17] to extract features. Average pooling and bilinear upsampling is adopted to adjust feature resolutions. The output features of the previous U-Net are directly passed to the next U-Net as input. The number N and channels of stacked U-Nets are adjusted to meet different requirements of computational cost.

3.3. Grouped Spatial-temporal Shift

In multi-frame fusion, frame-wise feature f_i is aggregated with neighboring features $\{f_{i-t}, \dots, f_{i+t}\}$ to obtain temporally fused features F_i. We adopt a 2D U-Net structure [45] for multi-frame fusion and keep skip connections in the U-Net. We replace several 2D convolution blocks by stacking multiple grouped spatial-temporal shift (GSTS) blocks to effectively establish temporal correspondences and conduct multi-frame fusion. The GSTS blocks are not applied at the finest scale to save the computational cost. A GSTS block consists of three components: 1) a temporal shift, 2) a spatial shift, and 3) a lightweight fusion layer, organized in the way shown in Figure 4.

Grouped temporal shift. It is observed in our experiment (Table 4) that, handling three frames simultaneously [33] would increase the difficulty of multi-frame fusion. To avoid it, our temporal shift processes only two adjacent frames. Grouped temporal shift blocks are either a forward temporal shift (FTS) block fusing $\{f_{i-1}, f_i\}$ (Figure 4 *Left*) or a backward temporal shift (BTS) block fusing $\{f_{i+1}, f_i\}$ (Figure 4 *Right*). To achieve bi-directional aggregation, we stack FTS blocks and BTS blocks alternatively.

In a temporal shift, multi-frame features $f_i \in \mathbb{R}^{h\times w\times c}$ are split (i.e. grouped) equally along the channel dimension to obtain two feature groups: f_i^a and f_i^b, where $f_i^a, f_i^b \in \mathbb{R}^{h\times w\times \frac{c}{2}}$. In the forward shift, f_i^a is not shifted and is aggregated with the forward-shifted feature f_{i-1}^b from time $i-1$. In the backward shift, f_i^a is backward-shifted to be aggregated with f_{i-1}^b for restoring I_{i-1}. In other words, both FTS and BTS blocks keep half of the feature channels (one feature group) for characterizing visual appearance at current time i and shift the other half of channels (the other feature group) for propagating information for inter-frame aggregation. For simplicity, in the following paragraphs, we explain the details of the FTS block (i.e. how f_i^a is aggregated with f_{i-1}^b), and the BTS block is similarly defined.

Grouped spatial shift. Concatenating f_i^a and f_{i-1}^b for restoring frame i does not account for the spatial misalignment between two frames i and i-1. Therefore, we perform additional spatial shift on the propagated feature group $f_{i-1}^b \in \mathbb{R}^{h\times w\times \frac{c}{2}}$ to achieve a large spatial range for spatial misalignment. Specifically, we first equally split (i.e. group) f_{i-1}^b along the channel dimension to obtain M feature slices $f_{i-1,m}^b \in \mathbb{R}^{h\times w\times \frac{c}{2M}}$, where $m = 1, \dots, M$ is the slice index. For each feature slice $f_{i-1,m}^b$, we spatially shift it by $(\Delta x_m, \Delta y_m)$ pixels in the x and y directions to obtain the shifted feature slice $f_{i-1,m}^{b'}$:

$$f_{i-1,m}^{b'} = \text{Shift}(f_{i-1,m}^b, \Delta x_m, \Delta y_m). \tag{2}$$

$|\Delta x_m| = k_x * (s-1) + 1, |\Delta y_m| = k_y * (s-1) + 1$, where k_x, k_y are integers and s is defined as the base length

Method	EDVR	Su et al.	STFAN	TSP	MPRNet	MSDI	NAFNet	RNN-MBP	VRT	Ours-s	Ours	Ours+
PSNR	26.83	27.31	28.59	31.67	32.66	33.28	33.69	33.32	34.81	35.22	35.49	**35.88**
SSIM	0.843	0.826	0.861	0.928	0.959	0.964	0.967	0.963	0.972	0.975	0.976	**0.979**
Params (M)	20.6	15.3	5.37	16.17	20.1	241.3	67.8	16.4	18.3	4.1	10.5	12.3
FLOPS (G)	194.2	38.7	35.4	357.9	760.1	336.4	63.3	496.0	721.3	47.1	146.5	151.3

Table 1. Quantitative comparison on GoPro [40] test set.

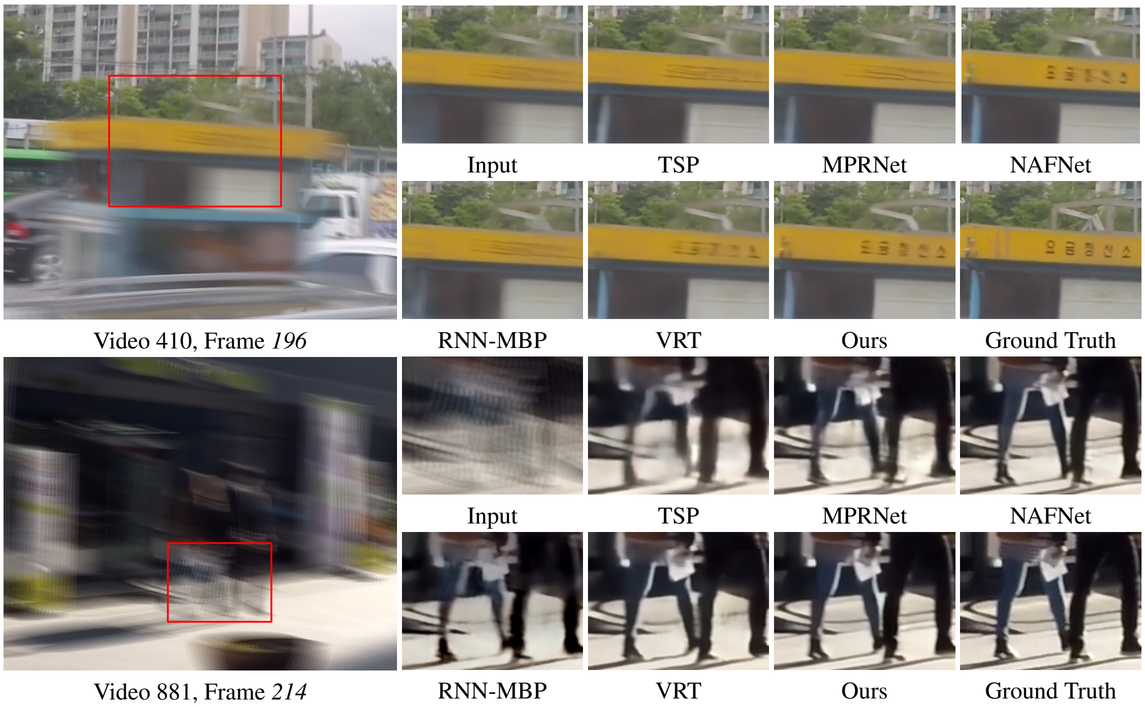

Figure 5. Video deblurring on GoPro [40] test set. Our method recovers more details than other methods.

of spatial shift. When the spatial shift causes void pixels in the border, we set them to zero. For a Δx_m pixels shift, the corresponding feature group is shifted spatially by Δx_m-1 pixels, followed by a depth-wise 3×3 convolution, which handles objects across two shifts and achieve smooth translation between two adjacent shifted feature slices. Then we concatenate all feature groups $f^{b'}_{i-1,m}$ along the channel dimension to obtain the spatially shifted feature $f^{b'}_{i-1}$:

$$f^{b'}_{i-1} = \mathrm{Concat}(f^{b'}_{i-1,1}, \ldots, f^{b'}_{i-1,M}). \tag{3}$$

For example, when $M = 9$ and $\Delta x_m, \Delta y_m \in \{-1, 0, 1\}$, the spatial shift operation creates 9 feature slices and shifts the different slices by the 9 directions. In our implementation, we set $M = 25$ and $\Delta x_m, \Delta y_m \in \{-9, -5, 0, 5, 9\}$ to enlarge the alignment and fusion's receptive fields, so as to handle large displacements across frames.

Fusion layer. We utilize a fusion layer F to aggregate multi-frame features $f^a_i, f^b_{i-1}, f^{b'}_{i-1}$. The fusion layer F contains two lightweight convolution blocks and each block adopts the combination between NAFNet [9] and Super Kernels [51], utilizing point-wise convolutions, depth-wise convolutions and gated layers to avoid heavy computation.

The output fused feature $\hat{f}_i$ of frame i is calculated as

$$\hat{f}_i = \mathrm{Concat}(f^a_i, f^b_{i-1}) + \mathrm{F}(f^a_i, f^b_{i-1}, f^{b'}_{i-1}). \tag{4}$$

The output feature $\hat{f}_i$ is fed to the next GSTS block. To effectively merge shifted features, the kernel size of convolutions is set to be equal to the base shift length s.

3.4. How Grouped Spatial Shift Help Restoration?

We provide visualization and analysis to explore how different shifted features groups help video restoration. We input two neighboring frames into Group Shift-Net. To analyze the feature map of grouped spatial shifts, we sample

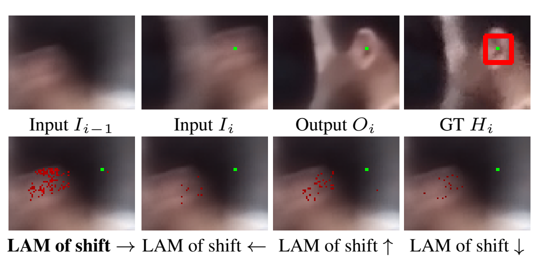

Figure 6. Local attribute visualization [16] of four shift directions. The saturation of red dots represent contribution weights of different areas in restoration of the marked local patch.

Method	EDVR	Su et al.	STFAN	TSP	PVDNet	ARVo	STDAN	ERDN	RNN-MBP	VRT	Ours-s	Ours	Ours+
PSNR	28.51	30.01	31.15	32.13	32.31	32.80	33.05	33.31	33.32	34.27	34.18	34.58	**34.69**
SSIM	0.864	0.887	0.905	0.927	0.926	0.935	0.937	0.940	0.963	0.965	0.965	0.968	**0.969**

Table 2. Quantitative comparison on DVD [50] test set.

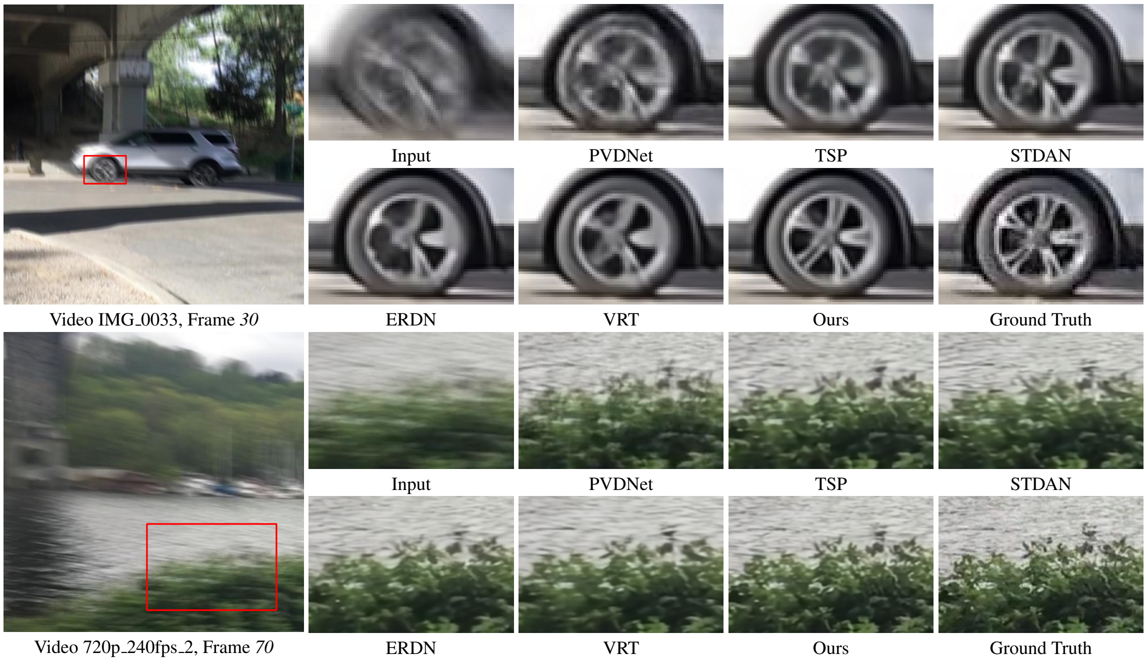

Figure 7. Video deblurring results on DVD [50]. Our method performs better at reconstructing details of leaves and the moving tire.

a 16×16 grid area from the resultant feature map. Local attribution map (LAM) [16] is performed to analyze contribution weights of all shifted features in restoring the 16×16 grid. The contribution weights are visualized as the red dots in Figure 6. When the color of dots is more saturated, the local area is more important in restoration. It is shown that the shifted features are more important in restoring O_i, when a shift direction is similar to the motion between I_{i-1} and I_i. Moreover, our method could obtain expansive effective receptive fields for temporal correspondence establishment.

4. Experiments

We conduct experiments and ablation study on two tasks: video deblurring and video denoising.

Datasets. For video deblurring, we train and evaluate our method on GOPRO [40] and DVD [50] datasets. GOPRO [40] dataset contains 2,103 and 1,111 frames as training and test sets, respectively. DVD [50] includes 5,708 frames for training and 1,000 frames for testing. For video denoising, we follow Huang et al. [18] to train our model with noise level $\sigma \in \mathcal{U}[0, 50]$ on DAVIS [25] dataset and test it on DAVIS [25] of different noise levels.

Model Scaling. We provide small model (denoted as "Ours-s"), base model (denoted as "Ours") to meet different computational requirements. For both models, We replace the convolution blocks by multiple GSTS blocks in Stage-2's UNet. We further observe that merely replacing convolution blocks in decoders of Stage-2's UNet (denoted as "Ours+") could boost the performances further. The details of different models are in Appendix.

Implementation details. Our network is end-to-end trained. The base shift length s is set to 5. The networks are trained with a batch size of 8 for 750 epochs. The reparameterization technique [13] is adopted to optimize convolutions in GSTS. The patch size is set as 256×256. Horizontal and vertical flips are adopted for data augmentation. We use the Adam optimizer [26] and the learning rate is decreased from 4×10^{-4} to 1×10^{-7} according to the cosine annealing strategy [36]. At inference of video deblurring, "Ours-s" processes 100 frames simultaneously. "Ours" and "Ours+" process only 50 frames due to the memory limit.

4.1. Video Deblurring Results

Quantitative comparison. We compare our method with state-of-the-art deblurring methods including EDVR [57], Su et al. [50], STFAN [68], TSP [42], MPRNet [64], MSDI-Net [28], NAFNet [9], RNN-MBP [7], STDAN [65], ERDN [22] and VRT [32]. As shown in Tables 1 and 2, "our+" outperforms VRT [32], the most competitive method, by 1.07 dB and 0.42 dB PSNR on GoPro and DVD, respectively, with only 21% of its flops. For a more intuitive comparison, we provide the PSNR-Params-FLOPS plot in Figure 1. The

Dataset	σ	VLNB	DVDNet	FastDVD	EMVD-L	PaCNet	Huang et al.	FloRNN	Tempformer	VRT	Ours-s	Ours	Ours+
	10	38.85	38.13	38.71	38.57	39.97	39.67	40.16	40.17	40.82	40.55	40.75	**40.85**
	20	35.68	35.70	35.77	35.39	36.82	36.33	37.52	37.36	38.15	37.84	38.19	**38.24**
DAVIS	30	33.73	34.08	34.04	33.89	34.79	34.62	35.89	35.66	36.52	36.25	36.62	**36.68**
	40	32.32	32.86	32.82	32.40	33.34	33.40	34.66	34.42	35.32	35.11	35.47	**35.56**
	50	31.13	31.85	31.86	31.47	32.20	32.41	33.67	33.44	34.36	34.20	34.53	**34.64**
Params (M)		-	-	2.5	9.6	2.87	13.95	11.8	-	18.3	3.7	10.8	12.9
FLOPS (G)		-	-	41.8	69.5	-	48.5	189.7	-	721.3	47.2	146.8	173.2

Table 3. Quantitative comparison on DAVIS [25] test set.

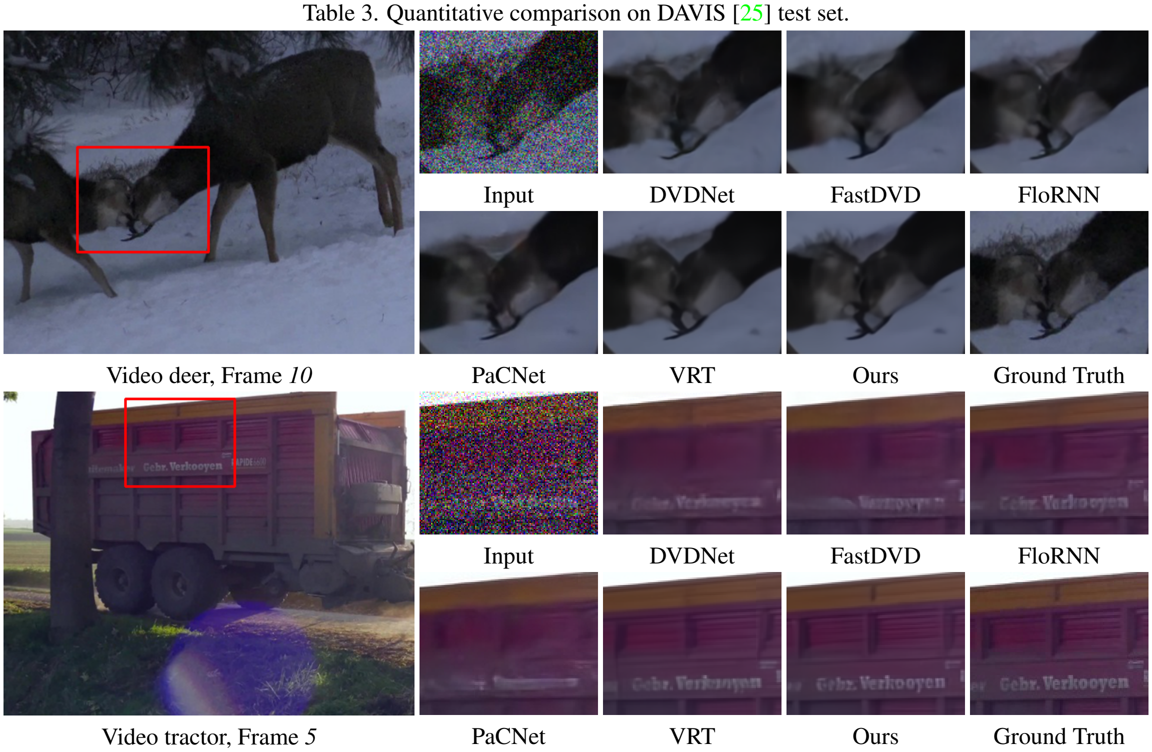

Figure 8. Video denoising results on DAVIS [25] test set. Our method reconstructs more details of textures and texts.

two versions of our model occupy the top-left corner, showing the best performances with less computational cost. Notably, "Ours-s" surpasses STFAN [68] by a significant 6.63 dB PSNR with the fewest parameters.

Qualitative comparison. Figure 5 provides the visualization of two hard deblurring cases. As one can see from the full images, there exist severe blurry regions due to camera shaking and object movement. On the zoomed-in patchs, our model reconstructs much sharper letters, building structures and boundaries of moving legs.

4.2. Video Denoising Results

Quantitative comparison. We compare our method with SOTA video denoising methods VLNB [1], DVDNet [52], FastDVD [53], EMVD-L [38], PaCNet [55], Huang et al. [18], FloRNN [30], Tempformer [49] and VRT [32]. It is shown in Table 3 that we achieve best performances in 5 noise levels on with less computational cost. Moreover, our small model performs better than previous lightweight models, such as FastDVD [53], EMVD-L [38].

Qualitative comparison. Figure 8 visualizes the denoising results of DAVIS [25] . Note the zoomed-in regions in the red boxes. Other models generate over-smooth results, while our model reconstructs more details in grass and texts.

4.3. Ablation Study

We demonstrate the effectiveness of each key component in Group Shift-Net. All compared methods are trained and evaluated with the same training settings of our base model.

Spatial Temporal shift. We evaluate the impact of *grouped spatial shift* and *alternative temporal shift* in Table 4. At first, We remove *grouped spatial shift* and merely apply alternative temporal shift . The kernel size of convolution in the fusion layers is set to be 3×3, which is widely used previously [33, 44]. It suffers a drop of 0.35 dB PSNR. Then we replace alternative temporal shift by bi-directional shift, where the fusion layer would aggregate $\frac{3}{4}$ channels of feature f_i, $\frac{1}{8}$ channels of feature f_{i-1}, and $\frac{1}{8}$ channels of

Alternative Temporal Shift	Spatial Shift	PSNR
✗	✗	34.81
✓	✗	35.14
✓	✓	**35.49**

Table 4. Ablation of grouped spatial-temporal shift.

Receptive Field	5×5	9×9	17×17	25×25	33×33
Ours w/o spatial	35.14	35.20	35.18	35.16	35.17

Receptive Field	13×13	23×23	33×33
Ours	35.39	**35.49**	35.48

Table 5. Receptive field and spatial shift in a fusion layer.

$\Delta x_m, \Delta y_m$	$\{0\}$	$\{0, \pm 1\}$	$\{0, \pm 2, \pm 3\}$	$\{0, \pm 3, \pm 5\}$
PSNR	35.20	35.29	35.37	35.35
$\Delta x_m, \Delta y_m$	$\{0, \pm 4, \pm 7\}$	$\{0, \pm 5, \pm 9\}$	$\{0, \pm 6, \pm 11\}$	$\{0, \pm 7, \pm 13\}$
PSNR	35.44	**35.49**	35.46	35.40

$\Delta x_m, \Delta y_m$	$\{0\}$	$\{0, \pm 5\}$	$\{0, \pm 5, \pm 9\}$	$\{0, \pm 5, \pm 9, \pm 13\}$
PSNR	35.14	35.34	**35.49**	35.47

Table 6. Ablation of $(\Delta x_m, \Delta y_m)$ in grouped spatial shift.

feature f_{i+1}. This operation causes a decrease of 0.33 dB PSNR. The ablation illustrates the importance of *grouped spatial shift* and *alternative temporal shifts*.

Receptive field in fusion layers. We change the base shift length s to be $3, 5, 7$. The corresponding receptive fields of a fusion layer (depth-wise convolutions with kernel size s+1) are $13 \times 13, 23 \times 23, 33 \times 33$. We also remove spatial shift and enlarge the kernel sizes of convolutions to achieve similiar receptive field (denoted as "Ours w/o spatial"). The kernel sizes of the depth-wise convolution are set to be 3, 5, 9, 13, 17 and the corresponding receptive fields are $5 \times 5, 9 \times 9, 17 \times 17, 25 \times 25, 33 \times 33$, respectively. It is shown in Table 5 that larger kernel convolutions cannot achieve better performances. It might be because extremely large kernels are stiill difficult to optimize. It is also observed that our method surpasses optimizing larger kernels by about 0.3 dB PSNR, which demonstrates the superiority of spatial shift.

Grouped spatial shift. We first set $M = 25$ and set the kernel size in fusion layers to be 5. Then we change the base shift length s from 1 to 7 and $\Delta x_m, \Delta y_m$ as shown in Table 6. $\Delta x_m, \Delta y_m \in \{0\}$ means that only temporal shift is applied. It is observed that the model with $\Delta x_m, \Delta y_m \in \{0, \pm 5, \pm 9\}$ achieves the best performance. When the base shift length s increases, the spatial shifts with larger receptive fields achieve better performance. The models suffer degraded performances when the shift length is larger than the kernel size. It is because convolutions would not filter shifted features seamlessly. Then we change the number M of shifts and $\Delta x_m, \Delta y_m$ are changed with the number M. It is shown in Table 6 that the models with $M = 49$ and $M = 25$ achieve the similar performances, which outperform the model with $M = 9$ by about 0.15 dB PSNR.

Temporal consistency. Following Tempformer [49], we add noise with 12 different noise seeds on DAVIS to create a dataset of 12-frame sequences. Mean absolute error between adjacent outputs is taken as the metric. Table 7 shows that our method and VRT achieve similar consistency.

Replacing shift blocks with optical flow, DCN and self-attention. We first replace our fusion layer in shift blocks with DCN (denoted as "GSTS+DCN") and cross-frame (shifted window size=8) self-attention layers (denoted as "GSTS+self-attn"). Table 8 shows that our simple structure achieves better performance than DCN and self-attention. Then we replace shift blocks with optical flow (a pre-trained SPyNet as initialization), DCN layers and cross-frame self-attention layers (shifted window size = 8), separately. It is observed in Table 8 that our method achieves better performance with less computational cost.

Method	$\sigma = 10$	$\sigma = 30$	$\sigma = 50$
VRT	1.5×10^{-3}	1.6×10^{-3}	2.0×10^{-3}
Ours	1.7×10^{-3}	1.7×10^{-3}	1.9×10^{-3}

Table 7. Temporal consistency evaluation of video denoising.

Method	Deblurring	DAVIS denoising			Params	FLOPs
	GoPro	σ=10	σ=30	σ=50		
GSTS + self-attn	34.67	40.54	36.28	34.17	11.1 (M)	168.7 (G)
GSTS + DCN	33.74	40.02	35.65	33.43	17.9 (M)	210.3 (G)
Optical Flow	34.14	40.07	35.68	33.55	14.2 (M)	189.4 (G)
self-attn	34.52	40.58	36.31	34.17	10.6 (M)	153.6 (G)
DCN	33.66	39.91	35.48	33.18	17.2 (M)	203.8 (G)
Ours	**35.49**	**40.75**	**36.62**	**34.53**	10.8 (M)	146.8 (G)

Table 8. Replacing shift blocks with several variants.

s	Deblurring	DAVIS (480×854)			DAVIS (240×427)			DAVIS (960×1708)		
	GoPro	σ=10	σ=30	σ=50	σ=10	σ=30	σ=50	σ=10	σ=30	σ=50
3	35.39	40.65	36.47	34.34	39.07	34.71	32.65	**43.20**	38.92	36.93
5	**35.49**	**40.75**	**36.62**	**34.53**	**39.13**	**34.82**	**32.77**	**43.20**	**39.05**	**37.11**
7	35.48	40.63	36.45	34.32	39.05	34.69	32.62	43.19	38.94	36.98
9	35.33	40.56	36.35	34.20	39.01	34.61	32.54	43.19	38.86	36.84

Table 9. Shift length s on different degradation and resolutions.

Shift length s. We first evaluate shift length s on different types of degradation, such as blur and noise. It is observed in Table 9 that the network with s=5 achieves the best performance on both video deblurring and denoising. We further apply bicubic upsampling and bilinear downsampling on DAVIS (480×854) to obtain a downsampled DAVIS dataset (240×427) and a upsampled DAVIS dataset (960×1708). As shown in Table 9, the network with s=5 achieves the best performance at all resolutions.

5. Conclusion

In this paper, we propose a simple and effective framework for video restoration that does not require complicated architectures like optical flow, deformable convolution, or self-attention. Instead, we introduce a simple spatial temporal shift block for implicit temporal correspondence modeling. Our method outperforms state-of-the-art methods with less computational cost on video deblurring and denoising tasks. We do not foresee any negative social impact resulting from this work.

Acknowledgments. This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund Project 14204021 and Reseasrch Impact Fund Project R5001-18 of Hong Kong RGC. Hongsheng Li and Xiaogang Wang are PIs of CPII under the InnoHK.

References

[1] Pablo Arias and Jean-Michel Morel. Video denoising via empirical bayesian estimation of space-time patches. *J. Math. Imaging Vis.*, 60(1):70–93, jan 2018. 7

[2] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2848–2857, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 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 conference on computer vision and pattern recognition*, 2021. 2

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *AAAI Conference on Artificial Intelligence*, 2021. 2

[5] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. BasicVSR++: Improving video super-resolution with enhanced propagation and alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[6] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3

[7] 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. 1, 2, 6

[8] Chen Chen, Qifeng Chen, Minh N. Do, and Vladlen Koltun. Seeing motion in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2

[9] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022. 5, 6

[10] W. Chen, D. Xie, Y. Zhang, and S. Pu. All you need is a few shifts: Designing efficient convolutional neural networks for image classification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7234–7243, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 2

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2017. 1, 2, 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[13] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742, 2021. 6

[14] Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *arXiv preprint arXiv:2203.06717*, 2022. 1

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 3

[16] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9199–9208, 2021. 5, 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[18] Cong Huang, Jiahao Li, Bin Li, Dong Liu, and Yan Lu. Neural compression-based feature learning for video restoration, 2022. 6, 7

[19] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 2

[20] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. Video super-resolution with recurrent structure-detail network. *CoRR*, abs/2008.00455, 2020. 2

[21] Yunho Jeon and Junmo Kim. Constructing fast network through deconstruction of convolution. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5955–5965, Red Hook, NY, USA, 2018. Curran Associates Inc. 2

[22] Bangrui Jiang, Zhihuai Xie, Zhen Xia, Songnan Li, and Shan Liu. Erdn: Equivalent receptive field deformable network for video deblurring. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 663–678, Cham, 2022. Springer Nature Switzerland. 6

[23] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3224–3232, 2018. 2

[24] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016. 2

[25] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Asian Conference on Computer Vision*, pages 123–141. Springer, 2018. 6, 7

[26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun,

editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[27] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu 0002, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7721–7731. Computer Vision Foundation / IEEE, 2021. 1, 2

[28] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 736–753, Cham, 2022. Springer Nature Switzerland. 6

[29] Dasong Li, Yi Zhang, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Efficient burst raw denoising with variance stabilization and multi-frequency denoising network, 2022. 2

[30] Junyi Li, Xiaohe Wu, Zhenxing Niu, and Wangmeng Zuo. Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones. *arXiv preprint arXiv:2204.05532*, 2022. 7

[31] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision, 2021. 3

[32] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 1, 2, 3, 6, 7

[33] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2, 4, 7

[34] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. 1

[36] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 6

[37] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4905–4913, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[38] M. Maggioni, Y. Huang, C. Li, S. Xiao, Z. Fu, and F. Song. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 7

[39] Ben Mildenhall, Jonathan T. Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, 2018. 1

[40] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 5, 6

[41] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing. 3

[42] Jinshan Pan, Haoran Bai, and Jinhui Tang. Cascaded deep video deblurring using temporal sharpness prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6

[43] Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2720–2729, 2017. 2

[44] Xuejian Rong, Denis Demandolx, Kevin Matzen, Priyam Chatterjee, and Yingli Tian. Burst denoising via temporally shifted wavelet transforms. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 240–256, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 7

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015. 3, 4

[46] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023. 1

[47] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. *arXiv preprint arXiv:2303.01237*, 2023. 2

[48] 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 (TOG)*, 40(5), 2021. 1, 2

[49] Mingyang Song, Yang Zhang, and Tunç O. Aydın. Tempformer: Temporally consistent transformer for video denoising. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, 2022. 7, 8

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

[51] SHANGKUN SUN, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. SKFlow: Learning optical flow with super kernels. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave,

and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 5

[52] Matias Tassano, Julie Delon, and Thomas Veit. dvdnet: a fast network for deep video denoising. In *2019 IEEE International Conference on Image Processing*, Taipei, Taiwan, Sept. 2019. 2, 7

[53] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 7

[54] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[55] G. Vaksman, M. Elad, and P. Milanfar. Patch craft: Video denoising by deep modeling and patch matching. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2137–2146, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 7

[56] Guangting Wang, Yucheng Zhao, Chuanxin Tang, Chong Luo, and Wenjun Zeng. When shift operation meets vision transformer: An extremely simple alternative to attention mechanism, 2022. 3

[57] Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2, 6

[58] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *2013 IEEE International Conference on Computer Vision*, pages 1385–1392, 2013. 2

[59] B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholaminejad, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9127–9135, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. 2

[60] Z. Xia, F. Perazzi, M. Gharbi, K. Sunkavalli, and A. Chakrabarti. Basis prediction networks for effective burst denoising with large kernels. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[61] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8):1106–1125, 2019. 1, 2, 3

[62] T. Yu, X. Li, Y. Cai, M. Sun, and P. Li. S2-mlp: Spatial-shift mlp architecture for vision. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3615–3624, Los Alamitos, CA, USA, jan 2022. IEEE Computer Society. 3

[63] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. Supervised raw video denoising with a benchmark dataset on dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[64] 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. 3, 6

[65] Huicong Zhang, Haozhe Xie, and Hongxun Yao. Spatio-temporal deformable attention network for video deblurring. In *ECCV*, 2022. 6

[66] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *European Conference on Computer Vision*, 2022. 2

[67] Yi Zhang, Dasong Li, Xiaoyu Shi, Dailan He, Kangning Song, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Kbnet: Kernel basis network for image restoration, 2023. 2

[68] 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 International Conference on Computer Vision*, 2019. 2, 6, 7

[69] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 1