EDVR: Video Restoration with Enhanced Deformable Convolutional Networks

Xintao Wang[1] Kelvin C.K. Chan[2] Ke Yu[1] Chao Dong[3] Chen Change Loy[2]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong [2]Nanyang Technological University, Singapore
[3]SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{wx016, yk017}@ie.cuhk.edu.hk chao.dong@siat.ac.cn {chan0899, ccloy}@ntu.edu.sg

Abstract

Video restoration tasks, including super-resolution, deblurring, etc, are drawing increasing attention in the computer vision community. A challenging benchmark named REDS is released in the NTIRE19 Challenge. This new benchmark challenges existing methods from two aspects: (1) how to align multiple frames given large motions, and (2) how to effectively fuse different frames with diverse motion and blur. In this work, we propose a novel Video Restoration framework with Enhanced Deformable convolutions, termed EDVR, to address these challenges. First, to handle large motions, we devise a Pyramid, Cascading and Deformable (PCD) alignment module, in which frame alignment is done at the feature level using deformable convolutions in a coarse-to-fine manner. Second, we propose a Temporal and Spatial Attention (TSA) fusion module, in which attention is applied both temporally and spatially, so as to emphasize important features for subsequent restoration. Thanks to these modules, our EDVR wins the champions and outperforms the second place by a large margin in all four tracks in the NTIRE19 video restoration and enhancement challenges. EDVR also demonstrates superior performance to state-of-the-art published methods on video super-resolution and deblurring. The code is available at https://github.com/xinntao/EDVR.

1. Introduction

In this paper, we describe our winning solution in the NTIRE 2019 challenges on video restoration and enhancement. The challenge releases a valuable benchmark, known as REalistic and Diverse Scenes dataset (REDS) [26], for the aforementioned tasks. In comparison to existing datasets, videos in REDS contain larger and more complex motions, making it more realistic and challenging. The competition enables fair comparisons among different algorithms and promotes the progress of video restoration.

Image restoration tasks such as super-resolution (SR) [5, 20, 41, 18, 45, 51] and deblurring [27, 15, 38] have experienced significant improvements over the last few years thanks to deep learning. The successes encourage the community to further attempt deep learning on the more challenging video restoration problems. Earlier studies [36, 4, 33, 19, 11] treat video restoration as a simple extension of image restoration. The temporal redundancy among neighboring frames is not fully exploited. Recent studies [2, 48, 37, 32] address the aforementioned problem with more elaborated pipelines that typically consist of four components, namely feature extraction, alignment, fusion, and reconstruction. The challenge lies in the design of the alignment and fusion modules when a video contains occlusion, large motion, and severe blurring. To obtain high-quality outputs, one has to (1) align and establish accurate correspondences among multiple frames, and (2) effectively fuse the aligned features for reconstruction.

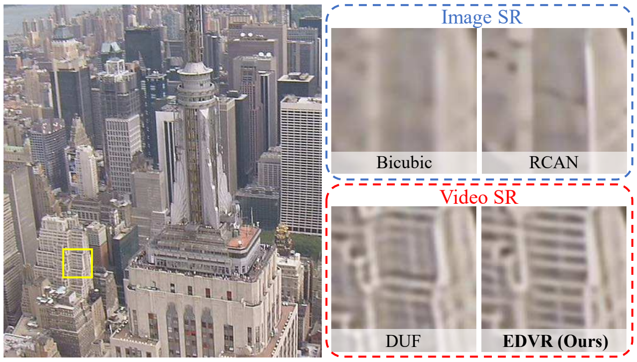

Figure 1. A comparison between image super-resolution and video super-resolution (×4). RCAN [51] and DUF [10] are the state-of-the-art methods of image and video super-resolution, respectively.

Alignment. Most existing approaches perform alignment by explicitly estimating optical flow field between the reference and its neighboring frames [2, 48, 13]. The neighboring frames are warped based on the estimated motion fields. Another branch of studies achieve implicit motion compensation by dynamic filtering [10] or deformable convolution [40]. REDS imposes a great challenge to existing alignment algorithms. In particular, precise flow estimation and accurate warping can be challenging and time-consuming for flow-based methods. In the case of large motions, it is difficult to perform motion compensation either explicitly or implicitly within a single scale of resolution.

Fusion. Fusing features from aligned frames is another critical step in the video restoration task. Most existing methods either use convolutions to perform early fusion on all frames [2] or adopt recurrent networks to gradually fuse multiple frames [32, 6]. Liu *et al.* [22] propose a temporal adaptive network that can dynamically fuse across different temporal scales. None of these existing methods consider the underlying visual informativeness on each frame – different frames and locations are not equally informative or beneficial to the reconstruction, as some frames or regions are affected by imperfect alignment and blurring.

Our Solution. We propose a unified framework, called EDVR, which is extensible to various video restoration tasks, including super-resolution and deblurring. The cores of EDVR are (1) an alignment module known as Pyramid, Cascading and Deformable convolutions (PCD), and (2) a fusion module known as Temporal and Spatial Attention (TSA).

The PCD module is inspired by TDAN [40] in using deformable convolutions to align each neighboring frame to the reference frame at the feature level. Different from TDAN, we perform alignment in a coarse-to-fine manner to handle large and complex motions. Specifically, we use a pyramid structure that first aligns features in lower scales with coarse estimations, and then propagates the offsets and aligned features to higher scales to facilitate precise motion compensation, similar to the notion adopted in optical flow estimation [7, 9]. Moreover, we cascade an additional deformable convolution after the pyramidal alignment operation to further improve the robustness of alignment.

The proposed TSA is a fusion module that helps aggregate information across multiple aligned features. To better consider the visual informativeness on each frame, we introduce temporal attention by computing the element-wise correlation between the features of the reference frame and each neighboring frame. The correlation coefficients then weigh each neighboring feature at each location, indicating how informative it is for reconstructing the reference image. The weighted features from all frames are then convolved and fused together. After the fusion with temporal attention, we further apply spatial attention to assign weights to each location in each channel to exploit cross-channel and spatial information more effectively.

We participated in all the four tracks in the video restoration and enhancement challenges [29, 28], including video super-resolution (clean/blur) and video deblurring (clean/compression artifacts). Thanks to the effective alignment and fusion modules, our EDVR has won the champion in all the four challenging tracks, demonstrating the effectiveness and the generalizability of our method. In addition to the competition results, we also report comparative results on existing benchmarks of video super-resolution and deblurring. Our EDVR shows superior performance to state-of-the-art methods in these video restoration tasks.

2. Related Work

Video Restoration. Since the pioneer work of SRCNN [5], deep learning methods have brought significant improvements in image and video super-resolution [20, 41, 18, 23, 49, 46, 2, 22, 32, 37, 48]. For video super-resolution, temporal alignment plays an important role and has been extensively studied. Several methods [2, 37, 32] use optical flow to estimate the motions between images and perform warping. However, accurate flow is difficult to obtain given occlusion and large motions. TOFlow [48] also reveals that the standard optical flow is not the optimal motion representation for video restoration. DUF [10] and TDAN [40] circumvent the problem by implicit motion compensation and surpass the flow-based methods. Our EDVR also enjoys the merits of implicit alignment, with a pyramid and cascading architecture to handle large motions.

Video deblurring also benefits from the development of learning-based methods [12, 24, 30, 34]. Several approaches [34, 50] directly fuse multiple frames without explicit temporal alignment, because the existence of blur increases the difficulty of motion estimation. Unlike these approaches, we attempt to acquire information from multiple frames using alignment, with a slight modification that an image deblurring module is added prior to alignment when there is a blur.

Deformable Convolution. Dai *et al.* [3] first propose deformable convolutions, in which additional offsets are learned to allow the network to obtain information away from its regular local neighborhood, improving the capability of regular convolutions. Deformable convolutions are widely used in various tasks such as video object detection [1], action recognition [52], semantic segmentation [3], and video super-resolution [40]. In particular, TDAN [40] uses deformable convolutions to align the input frames at the feature level without explicit motion estimation or image warping. Inspired by TDAN, our PCD module adopts deformable convolution as a basic operation for alignment.

Attention Mechanism. Attention has proven its effectiveness in many tasks [43, 47, 22, 23, 51]. For example, in video SR, Liu *et al.* [22] learn a set of weight maps to weigh the features from different temporal branches. Non-local operations [44] compute the response at a position as a weighted sum of the features at all positions for capturing long-range dependencies. Motivated by the success of these works, we employ both temporal and spatial attention in our TSA fusion module to allow different emphases on different temporal and spatial locations.

3. Methodology

3.1. Overview

Given $2N+1$ consecutive low-quality frames $I_{[t-N:t+N]}$, we denote the middle frame I_t as the reference frame and the other frames as neighboring frames. The aim of video restoration is to estimate a high-quality reference frame $\hat{O}_t$, which is close to the ground truth frame O_t. The overall framework of the proposed EDVR is shown in Fig. 2. It is a generic architecture suitable for several video restoration

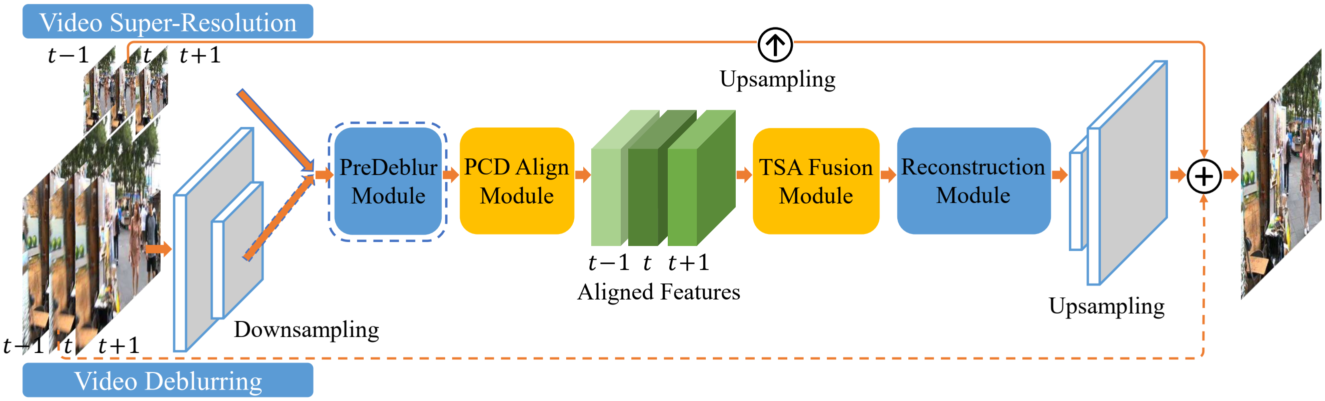

Figure 2. **The EDVR framework**. It is a unified framework suitable for various video restoration tasks, *e.g*., super-resolution and deblurring. Inputs with high spatial resolution are first down-sampled to reduce computational cost. Given blurry inputs, a PreDeblur Module is inserted before the PCD Align Module to improve alignment accuracy. We use three input frames as an illustrative example.

tasks, including super-resolution, deblurring, denoising, deblocking, *etc*.

Take video SR as an example, EDVR takes $2N{+}1$ low-resolution frames as inputs and generates a high-resolution output. Each neighboring frame is aligned to the reference one by the PCD alignment module at the feature level. The TSA fusion module fuses image information of different frames. The details of these two modules are described in Sec. 3.2 and Sec. 3.3. The fused features then pass through a reconstruction module, which is a cascade of residual blocks in EDVR and can be easily replaced by any other advanced modules in single image SR [46, 51]. The upsampling operation is performed at the end of the network to increase the spatial size. Finally, the high-resolution frame $\hat{O}_t$ is obtained by adding the predicted image residual to a direct upsampled image.

For other tasks with high spatial resolution inputs, such as video deblurring, the input frames are first downsampled with strided convolution layers. Then most computation is done in the low-resolution space, which largely saves the computational cost. The upsampling layer at the end will resize the features back to the original input resolution. A PreDeblur module is used before the alignment module to pre-process blurry inputs and improve alignment accuracy.

Though a single EDVR model could achieve state-of-the-art performance, we adopt a two-stage strategy to further boost the performance in NTIRE19 competition. Specifically, we cascade the same EDVR network but with shallower depth to refine the output frames of the first stage. The cascaded network can further remove the severe motion blur that cannot be handled by the preceding model. The details are presented in Sec. 3.4.

3.2. Alignment with Pyramid, Cascading and Deformable Convolution

We first briefly review the use of deformable convolution for alignment [40], *i.e*., aligning features of each neighboring frame to that of the reference one. Different from optical-flow based methods, deformable alignment is applied on features of each frame, denoted by $F_{t+i}, i{\in}[-N{:}+N]$. We use the modulated deformable module [53]. Given a deformable convolution kernel of K sampling locations, we denote w_k and $\mathbf{p}_k$ as the weight and the pre-specified offsets for the k-th location, respectively. For instance, a $3{\times}3$ kernel is defined with $K{=}9$ and $\mathbf{p}_k{\in}\{(-1,-1),(-1,0),\cdots,(1,1)\}$. The aligned features F^a_{t+i} at each position $\mathbf{p}_0$ can then be obtained by:

$$F^a_{t+i}(\mathbf{p}_0) = \sum_{k=1}^{K} w_k \cdot F_{t+i}(\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k) \cdot \Delta m_k. \quad (1)$$

The learnable offset $\Delta\mathbf{p}_k$ and the modulation scalar Δm_k are predicted from concatenated features of a neighboring frame and the reference one:

$$\Delta\mathbf{P}_{t+i} = f(\,[F_{t+i}, F_t]\,), \quad i \in [-N:+N] \quad (2)$$

where $\Delta\mathbf{P}{=}\{\Delta\mathbf{p}\}$, f is a general function consisting several convolution layers, and $[\cdot,\cdot]$ denotes the concatenation operation. For simplicity, we only consider learnable offsets $\Delta\mathbf{p}_k$ and ignore modulation Δm_k in the descriptions and figures. As $\mathbf{p}_0 + \mathbf{p}_k + \Delta\mathbf{p}_k$ is fractional, bilinear interpolation is applied as in [3].

To address complex motions and large parallax problems in alignment, we propose PCD module based on well-established principles in optical flow: pyramidal processing [31, 35] and cascading refinement [7, 8, 9]. Specifically, as shown with black dash lines in Fig. 3, to generate feature F^l_{t+i} at the l-th level, we use strided convolution filters to downsample the features at the $(l{-}1)$-th pyramid level by a factor of 2, obtaining L-level pyramids of feature representation. At the l-th level, offsets and aligned features are predicted also with the $\times 2$ upsampled offsets and aligned features from the upper $(l{+}1)$-th level, respectively (purple dash lines in Fig. 3):

$$\Delta\mathbf{P}^l_{t+i} = f(\,[F_{t+i}, F_t],\ (\Delta\mathbf{P}^{l+1}_{t+i})^{\uparrow 2}\,), \quad (3)$$

$$(F^a_{t+i})^l = g(\,\mathrm{DConv}(F^l_{t+i}, \Delta\mathbf{P}^l_{t+i}),\ ((F^a_{t+i})^{l+1})^{\uparrow 2}\,), \quad (4)$$

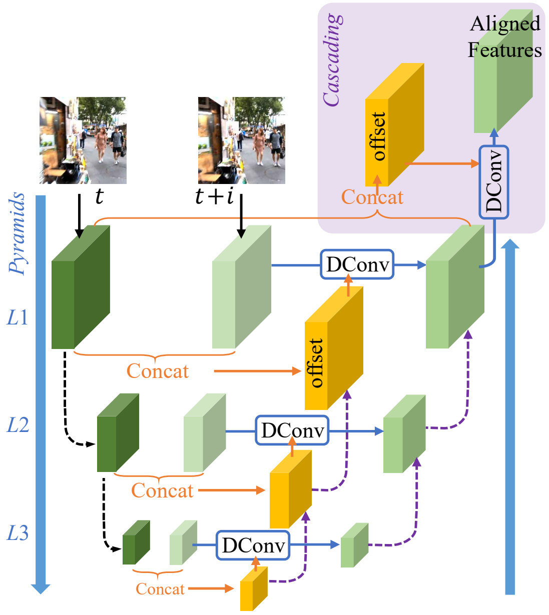

Figure 3. PCD alignment module with Pyramid, Cascading and Deformable convolution.

where $(\cdot)^{\uparrow s}$ refers to upscaling by a factor s, DConv is the deformable convolution described in Eqn. 1, and g is a general function with several convolution layers. Bilinear interpolation is adopted to implement the $\times 2$ upsampling. We use three-level pyramid, *i.e.*, L=3, in EDVR. To reduce computational cost, we do not increase channel numbers as spatial sizes decrease.

Following the pyramid structure, a subsequent deformable alignment is cascaded to further refine the coarsely aligned features (the part with light purple background in Fig. 3). PCD module in such a coarse-to-fine manner improves the alignment to the sub-pixel accuracy. We demonstrate the effectiveness of PCD in Sec. 4.3. It is noteworthy that the PCD alignment module is jointly learned together with the whole framework, without additional supervision [40] or pretraining on other tasks like optical flow [48].

3.3. Fusion with Temporal and Spatial Attention

Inter-frame temporal relation and intra-frame spatial relation are critical in fusion because 1) different neighboring frames are not equally informative due to occlusion, blurry regions and parallax problems; 2) misalignment and unalignment arising from the preceding alignment stage adversely affect the subsequent reconstruction performance. Therefore, dynamically aggregating neighboring frames in pixel-level is indispensable for effective and efficient fusion. In order to address the above problems, we propose TSA fusion module to assign pixel-level aggregation weights on each frame. Specifically, we adopt temporal and spatial attentions during the fusion process, as shown in Fig. 4.

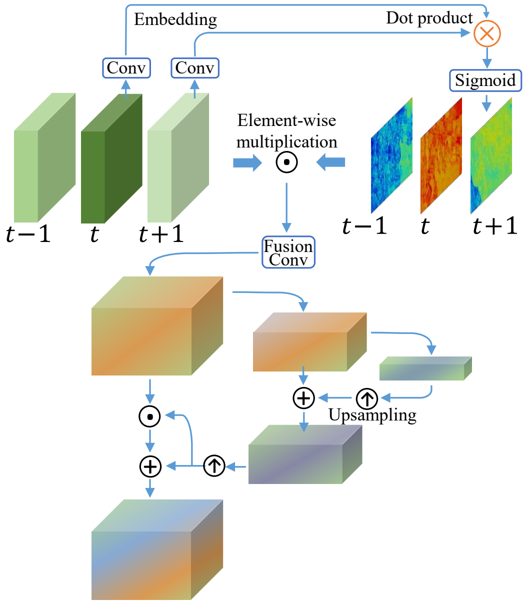

Figure 4. TSA fusion module with Temporal and Spatial Attention.

The goal of temporal attention is to compute frame similarity in an embedding space. Intuitively, in an embedding space, a neighboring frame that is more similar to the reference one, should be paid more attention. For each frame $i{\in}[-N{:}+N]$, the similarity distance h can be calculated as:

$$h(F^a_{t+i}, F^a_t) = \text{sigmoid}(\,\theta(F^a_{t+i})^T \phi(F^a_t)\,), \qquad (5)$$

where $\theta(F^a_{t+i})$ and $\phi(F^a_t)$ are two embeddings, which can be achieved with simple convolution filters. The sigmoid activation function is used to restrict the outputs in $[0, 1]$, stabilizing gradient back-propagation. Note that for each spatial location, the temporal attention is spatial-specific, *i.e.*, the spatial size of $h(F^a_{t+i}, F^a_t)$ is the same as that of F^a_{t+i}.

The temporal attention maps are then multiplied in a pixel-wise manner to the original aligned features F^a_{t+i}. An extra fusion convolution layer is adopted to aggregate these attention-modulated features $\tilde{F}^a_{t+i}$:

$$\tilde{F}^a_{t+i} = F^a_{t+i} \odot h(F^a_{t+i}, F^a_t), \qquad (6)$$

$$F_{\text{fusion}} = \text{Conv}(\,[\tilde{F}^a_{t-N}, \cdots, \tilde{F}^a_t, \cdots, \tilde{F}^a_{t+N}]\,), \qquad (7)$$

where $\odot$ and $[\cdot, \cdot, \cdot]$ denote the element-wise multiplication and concatenation, respectively.

Spatial attention masks are then computed from the fused features. A pyramid design is employed to increase the attention receptive field. After that, the fused features are modulated by the masks through element-wise multiplication and addition, similar to [45]. The effectiveness of TSA module is presented in Sec. 4.3.

3.4. Two-Stage Restoration

Though a single EDVR equipped with PCD alignment module and TSA fusion module could achieve state-of-the-art performance, it is observed that the restored images are not perfect, especially when the input frames are blurry or severely distorted. In such a harsh circumstance, motion compensation and detail aggregation are affected, resulting in inferior reconstruction performance.

Intuitively, coarsely restored frames would greatly mitigates the pressure for alignment and fusion. Thus, we employ a two-stage strategy to further boost the performance. Specifically, a similar but shallower EDVR network is cascaded to refine the output frames of the first stage. The benefits are two-fold: 1) it effectively removes the severe motion blur that cannot be handled in the preceding model, improving the restoration quality; 2) it alleviates the inconsistency among output frames. The effectiveness of two-stage restoration is illustrated in Sec. 4.4.

4. Experiments

4.1. Training Datasets and Details

Training datasets. Previous studies on video processing [21, 10, 34] are usually developed or evaluated on private datasets. The lack of standard and open video datasets restricts fair comparisons. REDS [26] is a newly proposed high-quality (720p) video dataset in the NTIRE19 Competition. REDS consists of 240 training clips, 30 validation clips and 30 testing clips (each with 100 consecutive frames). During the competition, since the test ground truth is not available, we select four representative clips (with diverse scenes and motions) as our test set, denoted by *REDS4*[1]. The remaining training and validation clips are re-grouped as our training dataset (a total of 266 clips). To be consistent with our methods and process in the competition, we also adopt this configuration in this paper.

Vimeo-90K [48] is a widely used dataset for training, usually along with Vid4 [21] and Vimeo-90K testing dataset (denoted by Vimeo-90K-T) for evaluation. We observe dataset bias when the distribution of training sets deviates from that of testing sets. More details are presented in Sec. 4.3.

Training details. The PCD alignment module adopts five residual blocks (RB) to perform feature extraction. We use 40 RBs in the reconstruction module and 20 RBs for the second-stage model. The channel size in each residual block is set to 128. We use RGB patches of size 64×64 and 256×256 as inputs for video SR and deblurring tasks, respectively. Mini-batch size is set to 32. The network takes five consecutive frames (*i.e.*, N=2) as inputs unless otherwise specified. We augment the training data with random horizontal flips and 90° rotations. We only adopt Charbonnier penalty function [17] as the final loss, defined by $\mathcal{L}=\sqrt{\|\hat{O}_t-O_t\|^2+\varepsilon^2}$, where ε is set to 1×10^{-3}.

We train our model with Adam optimizer [14] by setting β_1=0.9 and β_2=0.999. The learning rate is initialized as 4×10^{-4}. We initialize deeper networks by parameters from shallower ones for faster convergence. We implement our models with the PyTorch framework and train them using 8 NVIDIA Titan Xp GPUs.

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

We compare our EDVR with several state-of-the-art methods on video SR and video deblurring respectively. Two-stage and self-ensemble strategies [20] are *not* used. In the evaluation, we include all the input frames and do not crop any border pixels except the DUF method [10]. We crop eight pixels near image boundary for DUF due to its severe boundary effects.

Video Super-Resolution. We compare our EDVR method with nine algorithms: RCAN [51], DeepSR [19], BayesSR [21], VESPCN [2], SPMC [37], TOFlow [48], FRVSR [32], DUF [10] and RBPN [6] on three testing datasets: Vid4 [21], Vimeo-90K-T [48] and REDS4. Most previous methods use different training sets and different down-sampling kernels, making the comparisons difficult. Each testing dataset has different characteristics. Vid4 is commonly used in video SR. The data has limited motion. Visual artifacts also exist on its ground-truth (GT) frames. Vimeo-90K-T is a much larger dataset with various motions and diverse scenes. REDS4 consists of high-quality images but with larger and more complex motions. We observe dataset bias when training and testing sets diverge a lot. Hence, we train our models on Vimeo-90K when evaluated on Vid4 and Vimeo-90K-T.

The quantitative results on Vid4, Vimeo-90K-T and REDS4 are shown in Table 1, Table 2 and Table 3 (Left), respectively. On Vid4, EDVR achieves comparable performance to DUF and outperforms other methods by a large margin. On Vimeo-90K-T and REDS, EDVR is significantly better than the state-of-the-art methods, including DUF and RBPN. Qualitative results on Vid4 and Vimeo-90K-T are presented in Fig. 5 and Fig. 6, respectively. On both datasets, EDVR recovers more accurate textures compared to existing methods, especially in the second image of Fig. 6, where the characters can be correctly identified only in the outputs of EDVR.

Video Deblurring. We compare our EDVR method with four algorithms: DeepDeblur[27], DeblurGAN [16], SRN-Deblur [39] and DBN [34] on the REDS4 dataset. Quantitative results are shown in Table 3 (Right). Our EDVR outperforms the state-of-the-art methods by a large margin. We attribute this to both the effectiveness of our method and the challenging REDS dataset that contains complex blurring. Visual results are presented in Fig. 7, while most methods are able to address small blurring, only EDVR can successfully recover clear details from extremely blurry images.

[1] Specifically, REDS4 contains the 000, 011, 015 and 020 clips.

Table 1. Quantitative comparison on **Vid4** for 4× video SR. **Red** and blue indicates the best and the second best performance, respectively. Y or RGB denotes the evaluation on Y (luminance) or RGB channels. '*' means the values are taken from their publications.

Clip Name	Bicubic (1 Frame)	RCAN [51] (1 Frame)	VESPCN* [2] (3 Frames)	SPMC [37] (3 Frames)	TOFlow [48] (7 Frames)	FRVSR* [32] (recurrent)	DUF [10] (7 Frames)	RBPN* [6] (7 Frames)	**EDVR (Ours)** (7 Frames)
Calendar (Y)	20.39/0.5720	22.33/0.7254	-	22.16/0.7465	22.47/0.7318	-	24.04/0.8110	23.99/0.807	**24.05**/**0.8147**
City (Y)	25.16/0.6028	26.10/0.6960	-	27.00/0.7573	26.78/0.7403	-	**28.27**/**0.8313**	27.73/0.803	28.00/0.8122
Foliage (Y)	23.47/0.5666	24.74/0.6647	-	25.43/0.7208	25.27/0.7092	-	**26.41**/**0.7709**	26.22/0.757	26.34/0.7635
Walk (Y)	26.10/0.7974	28.65/0.8719	-	28.91/0.8761	29.05/0.8790	-	30.60/0.9141	30.70/0.909	**31.02**/**0.9152**
Average (Y)	23.78/0.6347	25.46/0.7395	25.35/0.7557	25.88/0.7752	25.89/0.7651	26.69/0.822	27.33/**0.8318**	27.12/0.818	**27.35**/0.8264
Average (RGB)	22.37/0.6098	24.02/0.7192	-/-	24.39/0.7534	24.41/0.7428	-/-	25.79/**0.8136**	-/-	**25.83**/0.8077

Table 2. Quantitative comparison on **Vimeo-90K-T** for 4× video SR. '†' means the values are taken from [48]. '*' means the values are taken from their publications.

Test on	Bicubic (1 Frame)	RCAN [51] (1 Frame)	DeepSR† [19] (7 Frames)	BayesSR† [21] (7 Frames)	TOFlow [48] (7 Frames)	DUF [10] (7 Frames)	RBPN* [6] (7 Frames)	**EDVR (Ours)** (7 Frames)
RGB Channels	29.79/0.8483	33.61/0.9101	25.55/0.8498	24.64/0.8205	33.08/0.9054	34.33/0.9227	-/-	**35.79**/**0.9374**
Y Channel	31.32/0.8684	35.35/0.9251	-/-	-/-	34.83/0.9220	36.37/0.9387	37.07/0.9435	**37.61**/**0.9489**

Table 3. Quantitative comparison on **REDS4**. **Left**: 4× Video SR (clean); **Right**: Video deblurring (clean). Test on RGB channels.

Method	Clip_000	Clip_011	Clip_015	Clip_020	Average
Bicubic	24.55/0.6489	26.06/0.7261	28.52/0.8034	25.41/0.7386	26.14/0.7292
RCAN [51]	26.17/0.7371	29.34/0.8255	31.85/0.8881	27.74/0.8293	28.78/0.8200
TOFlow [48]	26.52/0.7540	27.80/0.7858	30.67/0.8609	26.92/0.7953	27.98/0.7990
DUF [10]	27.30/0.7937	28.38/0.8056	31.55/0.8846	27.30/0.8164	28.63/0.8251
EDVR (Ours)	**28.01**/**0.8250**	**32.17**/**0.8864**	**34.06**/**0.9206**	**30.09**/**0.8881**	**31.09**/**0.8800**

Method	Clip_000	Clip_011	Clip_015	Clip_020	Average
DeblurGAN [16]	26.57/0.8597	22.37/0.6637	26.48/0.8258	20.93/0.6436	24.09/0.7482
DeepDeblur [27]	29.13/0.9024	24.28/0.7648	28.58/0.8822	22.66/0.6493	26.16/0.8249
SRN-Deblur [39]	28.95/0.8734	25.48/0.7595	29.26/0.8706	24.21/0.7528	26.98/0.8141
DBN [34]	30.03/0.9015	24.28/0.7331	29.40/0.8878	22.51/0.7039	26.55/0.8066
EDVR (Ours)	**36.66**/**0.9743**	**34.33**/**0.9393**	**36.09**/**0.9542**	**32.12**/**0.9269**	**34.80**/**0.9487**

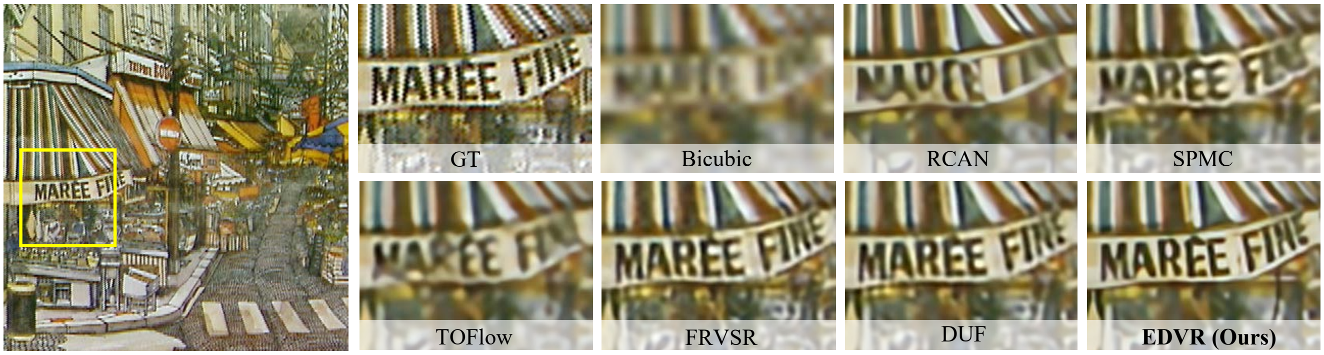

Figure 5. Qualitative comparison on the **Vid4** dataset for 4× video SR. **Zoom in for best view.**

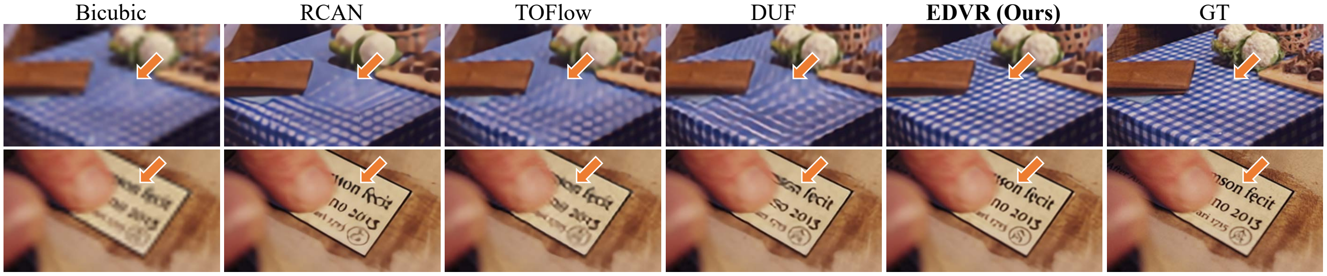

Figure 6. Qualitative comparison on the **Vimeo-90K-T** dataset for 4× video SR. **Zoom in for best view.**

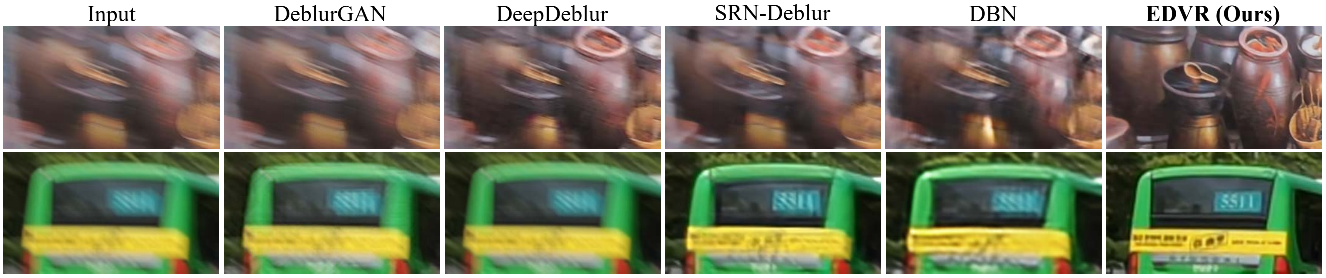

Figure 7. Qualitative comparison on the **REDS4** dataset for video deblurring. **Zoom in for best view.**

Table 4. Ablations on: **Left**: PCD and TSA modules (Experiments here adopt a smaller model with 10 RBs in the reconstruction module and the channel number is set to 64). FLOPs [25] are calculated on an image with the HR size of 1280 × 720. **Right**: the bias between the training and testing datasets.

Model	Model 1	Model 2	Model 3	Model 4
PCD?	✗ (1 DConv)	✗ (4 DConv)	✓	✓
TSA?	✗	✗	✗	✓
PSNR	29.78	29.98	30.39	30.53
FLOPs	640.2G	932.9G	939.3G	936.5G

Train \ Test	REDS4	Vid4 [21]	Vimeo90k [48]
REDS (5 frames)	**31.09/0.8800**	25.37/0.7956	34.33/0.9246
Vimeo-90K(7 frames)	30.49/0.8700	**25.83/0.8077**	**35.79/0.9374**
Δ	0.60/0.0100	-0.46/-0.0121	-1.46/-0.0128

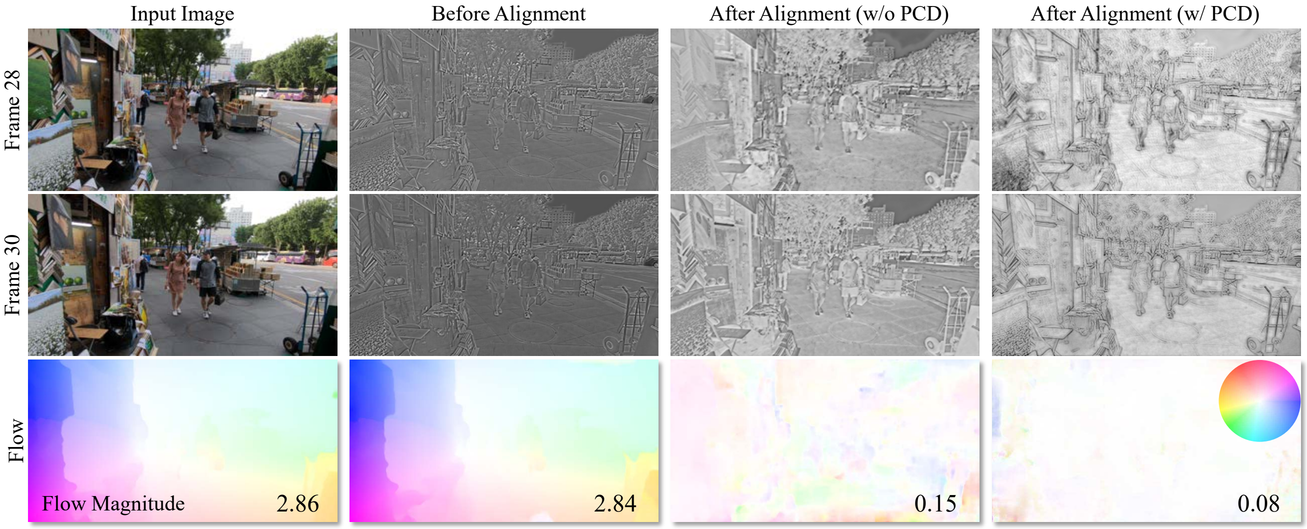

Figure 8. Ablation on the PCD alignment module. Compared with the results without PCD alignment, the flow of the PCD outputs is much smaller and cleaner, indicating that the PCD module can successfully handle large and complex motions. *Flow field color coding scheme* is shown in the right. The direction and magnitude of the displacement vector are represented by hue and color intensity, respectively.

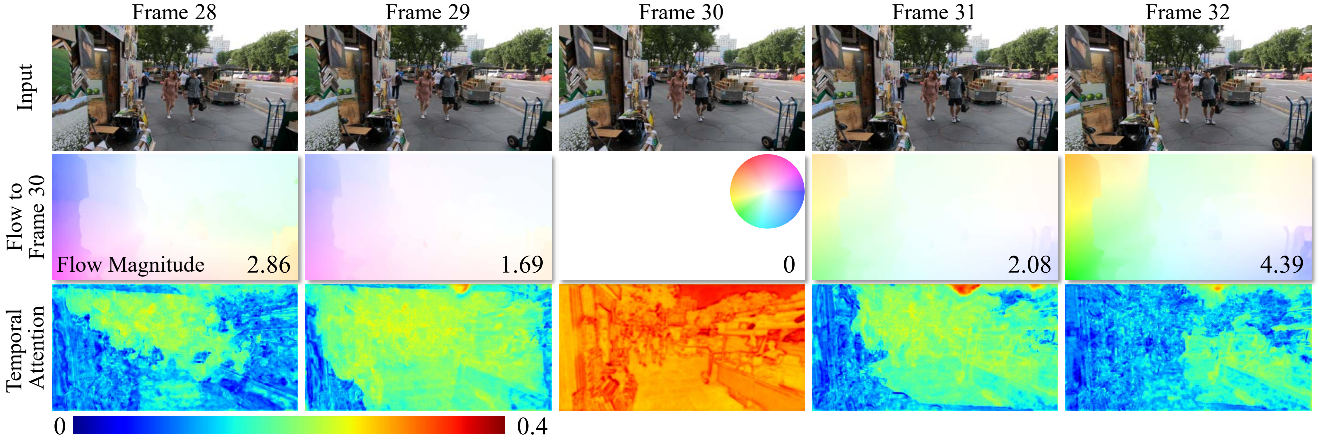

Figure 9. Ablation on the TSA fusion module. The frames and regions of lower flow magnitude tend to have more attention, indicating that the corresponding frames and regions are more informative.

4.3. Ablation Studies

PCD Alignment Module. As shown in Table 4 (Left), our baseline (Model 1) only adopts one deformable convolution for alignment. Model 2 follows the design of TDAN [40] to use four deformable convolutions for alignment, achieving an improvement of 0.2 dB. With our proposed PCD module, Model 3 is nearly 0.4 dB better than Model 2 with roughly the same computational cost, demonstrating the effectiveness of PCD alignment module. In Fig. 8, we show representative features before and after different alignment modules, and depict the flow (derived by PWCNet [35]) between reference and neighboring features. Compared with the flow without PCD alignment, the flow of the PCD outputs is much smaller and cleaner, indicating that the PCD module can successfully handle large and complex motions.

TSA Attention Module. As shown in Table 4 (Left), with the TSA attention module, Model 4 achieves 0.14 dB performance gain compared to Model 3 with similar computations. In Fig. 9, we present the flow between the reference and neighboring frames, together with the temporal atten-

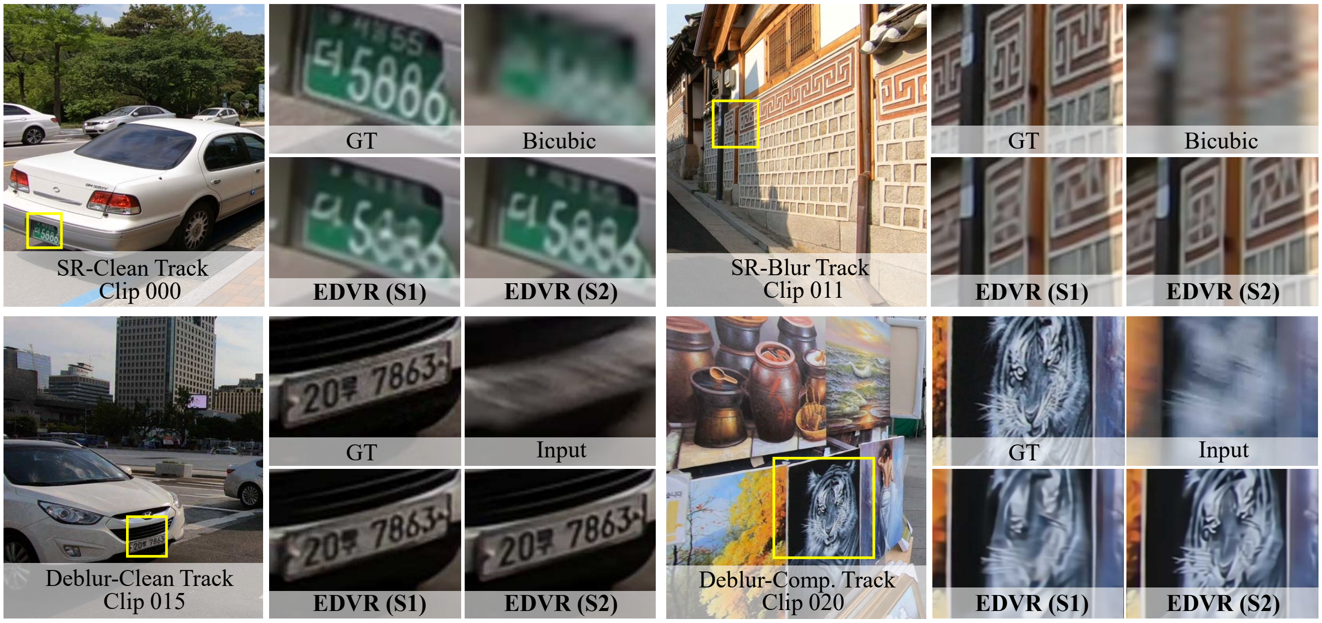

Figure 10. Qualitative results of our EDVR method on the four tracks in the NTIRE 2019 video restoration and enhancement challenges.

Table 5. Top5 methods in the NTIRE 2019 challenges on video restoration and enhancement. **Red** and blue indicates the best and the second best performance, respectively.

	SR		Deblur	
Method	Clean	Blur	Clean	Compression
EDVR (Ours)	**31.79/0.8962**	**30.17/0.8647**	**36.96/0.9657**	**31.69/0.8783**
$2nd$ method	31.13/0.8811	-/-	35.71/0.9522	29.78/0.8285
$3rd$ method	31.00/0.8822	27.71/0.8067	34.09/0.9361	29.63/0.8261
$4th$ method	30.97/0.8804	28.92/0.8333	33.71/0.9363	29.19/0.8190
$5th$ method	30.91/0.8782	28.98/0.8307	33.46/0.9293	28.33/0.7976

tion of each frame. It is observed that the frames and regions with lower flow magnitude tend to have higher attention, indicating that the smaller the motion is, the more informative the corresponding frames and regions are.

Dataset Bias. As shown in Table 4 (Right), we conduct different settings of training and testing datasets for video super-resolution. The results show that there exists a large dataset bias. The performance decreases 0.5-1.5 dB when the distribution of training and testing data mismatch. We believe that the generalizability of video restoration methods is worth investigating.

4.4. Evaluation on REDS Dataset

We participated in all the four tracks in the NTIRE19 video restoration and enhancement challenges [29, 28]. Quantitative results are presented in Table 5. Our EDVR wins the champions and outperforms the second place by a large margin in all tracks. In the competition, we adopt self-ensemble as [42, 20]. Specifically, during the test time, we flip and rotate the input image to generate *four* augmented inputs for each sample. We then apply the EDVR method on each, reverse the transformation on the restored outputs and average for the final result. The two-stage restoration strategy as described in Sec. 3.4 is also used to boost the performance. As shown in Table 6, we observe that the two-stage restoration largely improves the performance around 0.5 dB (EDVR(+) vs. EDVR-S2(+)). While the self-ensemble is helpful in the first stage (EDVR vs. EDVR+), it only brings marginal improvement in the second stage (EDVR-S2 vs. EDVR-S2+). Qualitative results are shown in Fig. 10. It is observed that the second stage helps recover clear details in challenging cases, *e.g.*, the inputs are extremely blurry.

Table 6. Evaluation on REDS4 for all the four competition tracks. '+' and '-S2' denote the self-ensemble strategy and two-stage restoration strategy, respectively.

Track		EDVR	EDVR-S2	EDVR+	EDVR-S2+
SR	Clean	31.09/0.8800	31.54/0.8888	31.23/0.8818	31.56/0.8891
	Blur	28.88/0.8361	29.41/0.8503	29.14/0.8403	29.49/0.8515
Deblur	Clean	34.80/0.9487	36.37/0.9632	35.27/0.9526	36.49/0.9639
	Comp.	30.24/0.8567	31.00/0.8734	30.46/0.8599	31.06/0.8741

5. Conclusion

We have introduced our winning approach in the NTIRE 2019 video restoration and enhancement challenges. To handle the challenging benchmark released in the competition, we propose EDVR, a unified framework with unique designs to achieve good alignment and fusion quality in various video restoration tasks. Thanks to the PCD alignment module and TSA fusion module, EDVR not only wins all four tracks in the NTIRE19 Challenges but also demonstrates superior performance to existing methods on several benchmarks of video super-resolution and deblurring.

Acknowledgement. We thank Yapeng Tian for providing the core codes of TDAN [40]. This work is supported by SenseTime Group Limited, Joint Lab of CAS-HK, the General Research Fund sponsored by the Research Grants Council of the Hong Kong SAR (CUHK 14241716, 14224316. 14209217), and Singapore MOE AcRF Tier 1 (M4012082.020).

References

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV*, 2018. 2

[2] Jose Caballero, Christian Ledig, Aitken Andrew, Acosta Alejandro, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017. 1, 2, 5, 6

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

[4] Qiqin Dai, Seunghwan Yoo, Armin Kappeler, and Aggelos K Katsaggelos. Dictionary-based multiple frame video super-resolution. In *ICIP*, 2015. 1

[5] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 1, 2

[6] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019. 1, 5, 6

[7] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Lite-flownet: A lightweight convolutional neural network for optical flow estimation. In *CVPR*, 2018. 2, 3

[8] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn–revisiting data fidelity and regularization. *arXiv preprint arXiv:1903.07414*, 2019. 3

[9] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 2, 3

[10] 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 *CVPR*, 2018. 1, 2, 5, 6

[11] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2016. 1

[12] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. Dynamic video deblurring using a locally adaptive blur model. *TPAMI*, 40(10):2374–2387, 2018. 2

[13] Tae Hyun Kim, Mehdi S M Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018. 1

[14] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[15] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. *arXiv preprint arXiv:1711.07064*, 2017. 1

[16] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 5, 6

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 5

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017. 1, 2

[19] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *CVPR*, 2015. 1, 5, 6

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 1, 2, 5, 8

[21] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *TPAMI*, 36(2):346–360, 2014. 5, 6, 7

[22] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, 2017. 1, 2

[23] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *NIPS*, 2018. 2

[24] Ziyang Ma, Renjie Liao, Xin Tao, Li Xu, Jiaya Jia, and Enhua Wu. Handling motion blur in multi-frame super-resolution. In *CVPR*, 2015. 2

[25] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference. *arXiv preprint arXiv:1611.06440*, 2016. 7

[26] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenges on video deblurring and super-resolution: Dataset and study. In *CVPRW*, June 2019. 1, 5

[27] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 5, 6

[28] Seungjun Nah, Radu Timofte, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Kyoung Mu Lee, Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, Chen Change Loy, et al. Ntire 2019 challenge on video deblurring: methods and results. In *CVPRW*, 2019. 8

[29] Seungjun Nah, Radu Timofte, Shuhang Gu, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Kyoung Mu Lee, Xintao Wang, Kelvin C.K. Chan, Ke Yu, Chao Dong, Chen Change Loy, et al. Ntire 2019 challenge on video super-resolution: Methods and results. In *CVPRW*, 2019. 8

[30] Liyuan Pan, Yuchao Dai, Miaomiao Liu, and Fatih Porikli. Simultaneous stereo video deblurring and scene flow estimation. In *CVPR*, 2017. 2

[31] Anurag Ranjan and Black J. Michael. Optical flow estimation using a spatial pyramid network. *arXiv preprint arXiv:1611.00850*, 2016. 3

[32] Mehdi S M Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 1, 2, 5, 6

[33] Oded Shahar, Alon Faktor, and Michal Irani. Space-time super-resolution from a single video. In *CVPR*, 2011. 1

[34] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. 2, 5, 6

[35] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 3, 7

[36] Hiroyuki Takeda, Peyman Milanfar, Matan Protter, and Michael Elad. Super-resolution without explicit subpixel motion estimation. *TIP*, 18(9):1958–1975, 2009. 1

[37] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *CVPR*, 2017. 1, 2, 5, 6

[38] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1

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

[40] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*, 2018. 1, 2, 3, 4, 7, 8

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Xintao Wang, Yapeng Tian, Ke Yu, Yulun Zhang, Wu Shixiang, Chao Dong, Yu Qiao, Chen Change Loy, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, 2017. 1, 2

[42] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *CVPR*, 2016. 8

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[45] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, 2018. 1, 4

[46] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *ECCVW*, 2018. 2, 3

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 2

[48] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *arXiv preprint arXiv:1711.09078*, 2017. 1, 2, 4, 5, 6, 7

[49] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *CVPR*, 2018. 2

[50] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Wei Liu, and Hongdong Li. Adversarial spatio-temporal learning for video deblurring. *TIP*, 28(1):291–301, 2019. 2

[51] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 1, 2, 3, 5, 6

[52] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Trajectory convolution for action recognition. In *NIPS*, 2018. 2

[53] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. *arXiv preprint arXiv:1811.11168*, 2018. 3