Deformable Kernel Convolutional Network for Video Extreme Super-Resolution

Xuan Xu[1][0000−0003−1057−3286], Xin Xiong[2], Jinge Wang[1], and Xin Li[1,†][0000−0003−2067−2763]

[1] West Virginia University, Morgantown WV 26505, USA
[2] Huazhong University of Science and Technology, Wuhan 430074, China
{xuxu,jnwang1}@mix.wvu.edu; xiong_xin@hust.edu.cn;
xin.li@mail.wvu.edu

Abstract. Video super-resolution, which attempts to reconstruct high-resolution video frames from their corresponding low-resolution versions, has received increasingly more attention in recent years. Most existing approaches opt to use deformable convolution to temporally align neighboring frames and apply traditional spatial attention mechanism (convolution based) to enhance reconstructed features. However, such spatial-only strategies cannot fully utilize temporal dependency among video frames. In this paper, we propose a novel deep learning based VSR algorithm, named Deformable Kernel Spatial Attention Network (DKSAN). Thanks to newly designed Deformable Kernel Convolution Alignment (DKC_Align) and Deformable Kernel Spatial Attention (DKSA) modules, DKSAN can better exploit both spatial and temporal redundancies to facilitate the information propagation across different layers. We have tested DKSAN on AIM2020 Video Extreme Super-Resolution Challenge to super-resolve videos with a scale factor as large as 16. Experimental results demonstrate that our proposed DKSAN can achieve both better subjective and objective performance compared with the existing state-of-the-art EDVR on Vid3oC and IntVID datasets.

Keywords: Video Super-Resolution, Deep Learning, Deformable Kernels, Deformable Convolution Network, Attention Mechanism.

1 Introduction

Video Super-Resolution (VSR) refers to the task of reconstructing high-resolution (HR) video frames from their corresponding low-resolution (LR) observation data. Similar to image super-resolution, VSR aims at faithful recovery of important image structures (e.g., edges and textures) and has been widely used in practical applications from video surveillance [26] and high-definition Television (HDTV) [24] to video coding and streaming [31]. Existing VSR research can be mainly classified into two subfields, enhancing spatial super-resolution and enhancing temporal super-resolution. The former focuses on super-resolving LR

[†]Corresponding author

video frames to approximate HR video frames to improve visual quality of video; while the later refers to interpolate new frames between neighboring frames for the purpose of increasing video frame rate (e.g., from 30fps to 60fps). Different from Single Image Super-Resolution (SISR) which only needs to consider the information from spatial domain, both spatial and temporal dependencies have to be utilized by VSR algorithms in order to optimize their performance. In particular, how to effectively exploit temporal redundancy by motion compensation techniques has remained one of the key technical challenges in the task of VSR.

In order to explore the potential benefit from temporal information of VSR, several existing approaches [5],[20],[23],[34] have used a sequence of consecutive LR frames (including one reference frame and several neighboring frames) as inputs to reconstruct the HR frame corresponding to the reference LR frame. To better exploit temporal dependency among multiple LR frames, the consecutive frames need to be aligned before the reconstruction of the HR frame. One of the most popular motion estimation methods, optical-flow estimation [11], is often considered and has been adopted by several VSR approaches [21],[25],[2]. However, VSR based on rigid motion estimation has to suffer from the potential problem arising from misalignment. For example, it is well known that there are two plagues with optical flow estimation: occlusion and aperture problems [29]. VSR based on incorrect motion estimation results may introduce undesired blurring and misregistration artifacts to the reconstructed HR frames.

In view of the weakness of rigid motion estimation approaches, alternative methods - namely deformable motion estimation - have been proposed as well. Recently, *deformable convolution* [4],[42] has become more and more popular as a supplementary module to video frame alignment. Several VSR works such as [35],[30],[33] have already successfully applied varying forms of deformable convolution alignment module to temporally align neighboring frames with respect to the reference frame, which demonstrates improved motion compensation when compared with optical-flow-based methods. However, existing deformable alignment modules still learn the motion parameters via several standard convolution layers with fixed kernel configurations, which can not extract accurate motion information especially in the presence of large and deformable motion (e.g., in sport video). By contrast, deformable kernels [8] can adapt effective receptive fields [22] (i.e., the support of filters) by weighting the per-pixel contribution, which is expected to be capable of characterizing more sophisticated motion information.

In this paper, we propose a novel Multi-Frame based Deformable Kernel Spatial Attention Network (DKSAN) for video *extreme* super-resolution (the upscaling factor is as large as 16). Inspired by EDVR [35] which applies deformable convolution [42] to temporally align neighboring frames with reference frame, we have designed a new module not based on optical flow estimation, called Deformable Kernel Convolution Alignment (DKC_Align) module, to enhance deformable convolution [42]. The key idea is to combine deformable kernel with deformable convolution to extract and improve not only global but also local edge and texture features while aligning the neighboring frames with respect

to the reference frame. Moreover, we have developed a Deformable Kernel Spatial Attention (DKSA) module to further enhance the spatial details of reconstructed feature maps, which extends the previous spatial attention works such as [35],[13],[38]. The novelty of DKSA module lies in that the deformable kernel [8] can better represent spatially-localized edge and texture features which are often important for the task of VSR than conventional convolution based spatial attention.

2 Related Works

Unlike image super-resolution which deals with reconstructing missing information in the spatial domain only, VSR has to not only reconstruct the missing high-frequency information in the spatial domain but also consider the motion-related consistency across different video frames in the temporal domain. In this section, we briefly review existing VSR approaches based on multi-frame such as [35],[30],[36],[13],[33],[2], optical flow [11] alignment and deformable convolution [42] alignment.

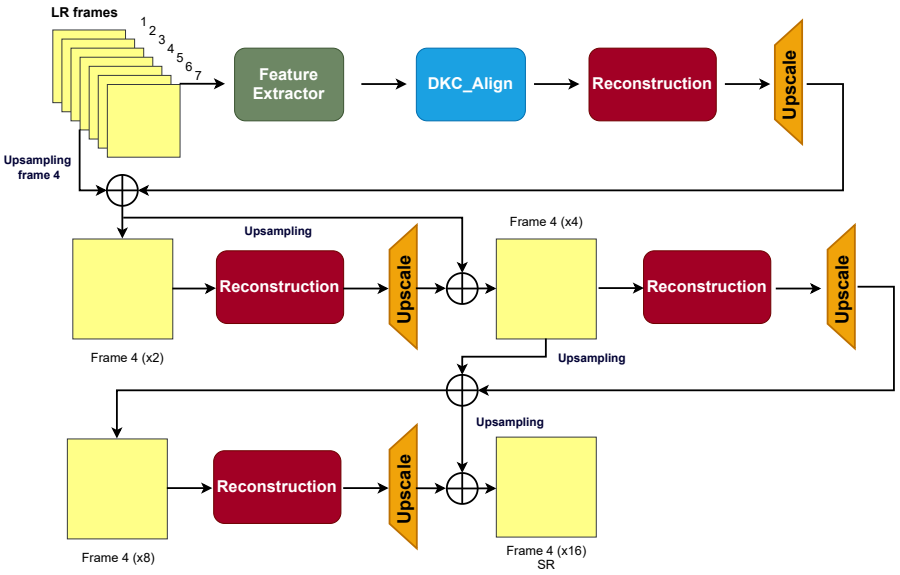

Fig. 1. Overview of DKSAN, $\oplus$ denotes element-wise sum.

2.1 Video Super-Resolution

One of the early works of applying optical-flow to VSR problems in order to utilize temporal and spatial information is [19]. In this work, a draft-ensemble

strategy was introduced to use two robust optical flow methods: TV-l_1 flow and MDP flow to overcome the difficulty with large motion variation and then combine SR drafts via a deep convolutional neural network to generate the final SR result. Later, [15] proposed to use optical-flow to estimate motion compensation of consecutive LR frames and wrapped them as inputs of the CNN to generate SR frames. Those two-stage approaches are not optimal solution since they separate the motion compensation from frame reconstruction. To explore potential benefits of end-to-end learning architecture for VSR problem, a novel end-to-end deep CNN to joint train the estimation of optical flow and spatio-temporal networks called ESPCN was developed in [2]. In [28], a new layer called sub-pixel motion compensation (SPMC) was introduced to handle inter-frame motion alignment; it also applied a ConvLSTM [37] architecture for reconstruction and testing. Another work [9] proposed a recurrent back-projection network (RBPN) with encoder-decoder mechanism to extract spatial and temporal information. In [14], dynamic upsampling filters (DUF) was developed to avoid use the explicit motion compensation by computing pixels of local spatio-temporal neighbors of LR frames to learn implicit motion compensation. Most recently, a novel temporal group attention (TGA) framework [13] was proposed to group the input frames (7 frames) as three groups then generate temporal spatial attention maps to reconstruct the missing details in the reference frame. Another recent work [40] proposed to learn self-supervised motion representation, task-oriented flow (TOFlow), instead of fix optical flow as the motion compensation module for VSR problem.

2.2 Deformable Convolution

The inherent limitation with traditional CNNs is the capability of modeling geometric transformations because of their fixed kernel shape. Although dilated convolution can alleviate this limitation to some degree, it is still difficult for standard fixed-shape convolutional kernels to align the key points or salient features in the input images. To solve this problem, a deformable convolution network has been developed in [4],[42] to improve the capability of modeling geometric transformations by adding flexible and learnable offsets. By acquiring information from other field rather than fixed local area, deformable convolution networks have been widely used by high-level vision tasks such as object detection [1] and segmentation [4]. Inspired by [42], a recent work [30] proposed a temporally-deformable alignment network (TDAN) to adapt deformable convolution to align the consecutive LR input frames at the feature level. Along this line of research, EDVR [35] designed a more aggressive alignment approach, PCD align module, to align the neighboring LR frames at different scale levels; also they proposed a temporal and spatial attention fusion module to future enhance important features. Another recent work [36] proposed a novel space-time video super-resolution framework to utilize deformable convolution and deformable ConvLSTM module to achieve temporal and spatial super-resolution at the same time. Most recently, [33] introduced another deformable convolution based VSR

framework called deformable non-local network (DNLN) with non-local attention module and hierarchical feature fusion block to enhance the global details between neighboring frames and references. Those deformable alignment based methods have shown better performance than optical-flow based networks.

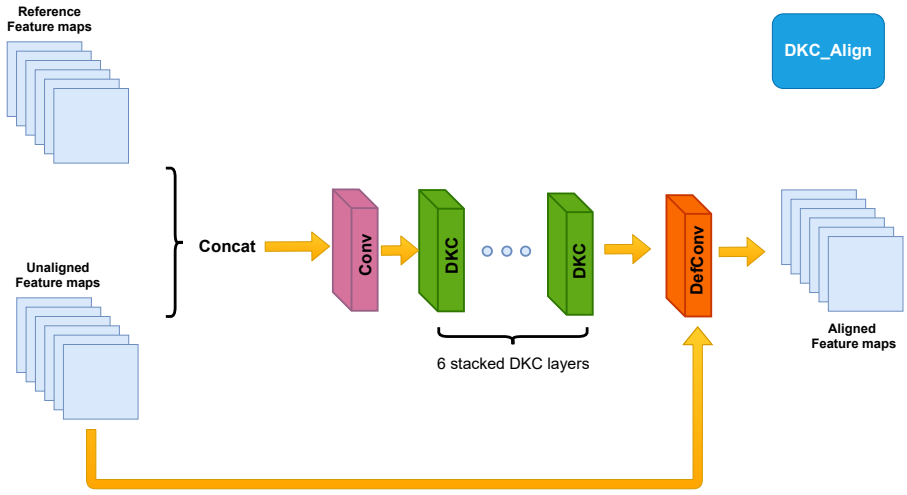

Fig. 2. Overview of DKC_Align module. Conv means convolution layer, DKC means deformable kernel convolution layer and DefConv stands for deformable convolution layer.

3 Proposed Methodology

The design of DKSAN network can be presented in the order of top-down hierarchy: DKSAN network (Fig. 1) → DKC_Align subnetwork (Fig. 2)→ reconstruction module (Fig. 3).

3.1 Overview: Deformable Kernel Spatial Attention Network

For multi-frame based VSR, we are given a group of $2N+1$ consecutive LR frames $I_T^{LR} = \{I_{r-N}^{LR}, \ldots, I_{r-1}^{LR}, I_r^{LR}, I_{r+1}^{LR}, \ldots, I_{r+N}^{LR}\}$, where I_r^{LR} is denoted as frame at the center or reference frame and I_{r-N}^{LR} or I_{r+N}^{LR} are the neighboring frames of I_r^{LR}. The goal of multi-frame based VSR is to reconstruct a HR frame $\hat{Y}_r$ from the LR sequence of I_T^{LR} by exploiting both spatial and temporal redundancies in the sequence. The overall diagram of our proposed networks DKSAN is shown in Fig. 1. It mainly includes four parts: *feature extraction, DKC_Align module, reconstruct module*, and *upscale module*. Different from traditional deep learning based multi-frame VSR architectures, this work aims at super-resolving the LR

videos at the extreme cases (e.g., with the scaling factor of 16). Due to large scaling factor constraint, it is difficult to upscale the LR feature maps to the target HR ones directly. One-time upscaling approaches such as [35],[30],[13] tend to introduce undesired blurring and artifacts to super-resolved HR video frames.

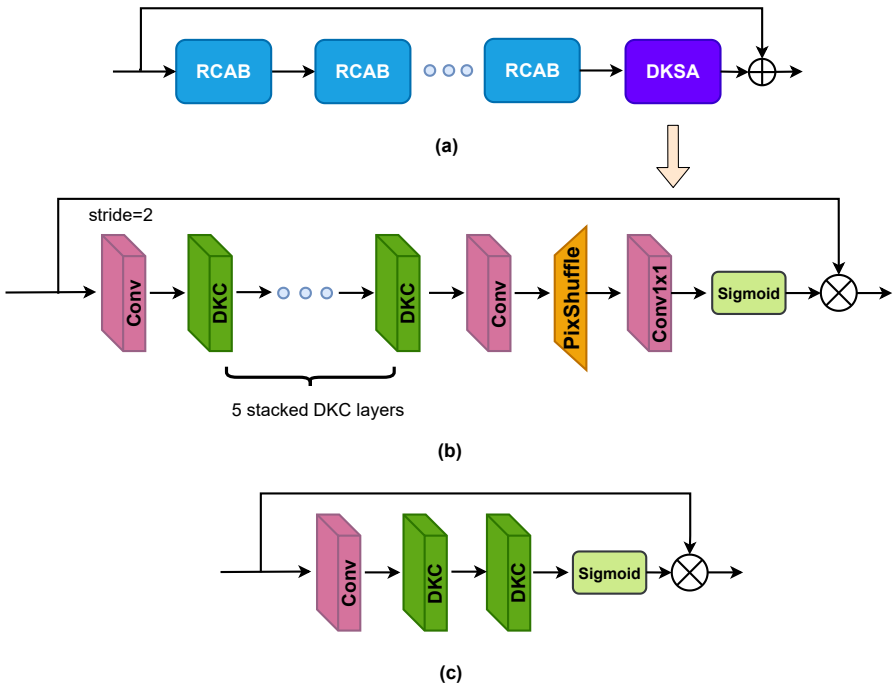

Fig. 3. Overview of reconstruction module; DKSA is deformable kernel spatial attention module shown in (b); a light version of DKSA is shown in (c); $\oplus$ and $\otimes$ denote element-wise sum and element-wise product, respectively.

To address this issue, we propose to construct a cascade of upscaling building blocks to iteratively super-resolve LR features several times (four times to reach the factor of $16 = 2^4$). Thanks to the cascade architecture, the LR frames can be super-resolved progressively to reconstruct the unknown HR frames more accurately than previous one-shot approaches. The whole problem of VSR can be formulated as follows:

$$\hat{Y}_r = \mathbb{F}(I_T^{LR}) \tag{1}$$

where I_T^{LR} denotes the consecutive LR frames and $\hat{Y}_r$ denotes the super-resolved reference frame I_r^{LR}. In particular, we extract the preliminary features of all input frames through the feature extraction which is stacked by several resblocks

[35] without batch normalization layers. This procedure can be represented by:

$$F_{fea} = E_{res}(I_T^{LR}) \tag{2}$$

where E_{res} denotes the preliminary feature extraction, the output F_{fea} is the extracted feature maps for all input frames. Let define F_n is the neighboring feature, and F_r is the reference feature separated from F_{fea}. To align the neighboring feature and the reference feature with the proposed DKC_Align module E_{DKC_Align}, we have

$$F_{Align} = E_{DKC\_Align}(F_n, F_r) \tag{3}$$

$$F_{fusion} = \mathbf{W}_E(F_{Align}) \tag{4}$$

where $n \in [t-N, t+N]$ and $n \neq r$, F_{Align} is the concatenated aligned feature maps for each neighboring frame feature with reference frame feature. The details about this alignment module will be elaborated in section 3.2; $\mathbf{W}_E \in \mathbb{R}^{1\times1\times C}$ is a 1×1 Conv layer. Conceptually similar to encoder-decoder configuration [3],[9], the aligned feature F_{Align} (encoder outputs) will be fed to the reconstruction module and upscale module for the first-level upscaling (decoder) operation:

$$\hat{Y}_r^{level1} = U_1(E_{Recon1}(F_{fusion})) + B_{2\times}(I_r^{LR}) \tag{5}$$

where E_{Recon1} denotes the first level reconstruction module, U_1 is the first level upscaling module and $B_{2\times}$ stands for the Bicubic interpolation with scale factor of 2; $\hat{Y}_r^{level1}$ is the $2\times$ SR frame. Finally, to get the extreme super-resolved frame $\hat{Y}_r$, we repeat another 3 times of reconstruction operation:

$$\hat{Y}_r^{level2} = U_2(E_{Recon2}(E_2(\hat{Y}_r^{level1}))) + B_{2\times}(\hat{Y}_r^{level1}) \tag{6}$$

$$\hat{Y}_r^{level3} = U_3(E_{Recon3}(E_3(\hat{Y}_r^{level2}))) + B_{2\times}(\hat{Y}_r^{level2}) \tag{7}$$

$$\hat{Y}_r = U_4(E_{Recon4}(E_4(\hat{Y}_r^{level3}))) + B_{2\times}(\hat{Y}_r^{level3}) \tag{8}$$

where E_2, E_3, E_4 are the preliminary feature extractors for each level; U_2, U_3, U_4 denote the upscaling module for each corresponding level, respectively. The details about the reconstruction module are described in section 3.3 including the DKSA module.

3.2 Deformable Kernel Alignment Module

Different from previous VSR works which applied optical flow to align neighboring frames with reference frame, [30] and [35] introduced to utilize modulated deformable convolution [42] to temporally align the given consecutive frames in order to add temporal information to VSR frameworks.

Deformable Convolution and Deformable Kernel Inspired by [8],[35], we propose a new alignment module, DKC_Align, to combine the deformable kernel [8] and deformable convolution [42] as shown in Fig. 2. First, let F_n and F_n^{align}

denote the input and output feature maps (not the reference frame feature), $\mathbf{W}_k$ represents the weight kernel and p_k is the pre-specified offsets for the k-th location (K is the total sampling location), then the modulated deformable convolution can be described as follows:

$$F_n^{align}(p) = \sum_{k \in K} \mathbf{W}_k \cdot F_n(p + p_k + \Delta p_k) \cdot \Delta m_k \tag{9}$$

where $F_n^{align}(p)$ and $F_n(p)$ indicate the feature location p from F_n^{align} and F_n, Δp_k and Δm_k stand for the learnable offset and the modulation scalar, respectively. With Δp_k and Δm_k, the convolution will get the ability to be irregularly dilated to work with important feature points without the shape limitation of conventional convolution. Such process of deformable convolution can be regarded as a strategy of adapting the local receptive field to a support of arbitrary shape.

To get Δp_k and Δm_k and align the neighboring feature with reference feature in particular, we first concatenate the neighboring frame feature and the reference frame feature then fuse them with one Conv2D layer and fed them into several deformable kernel layers:

$$\Delta P_n, \Delta M_n = \mathbb{D}(f([F_n, F_r])), n \in [t - N, t + N], n \neq r \tag{10}$$

where f denotes the one Conv2d layer to fuse F_n and F_r, $\mathbb{D}$ represents the deformable kernel convolution layer. To formally express deformable kernel convolution layer $\mathbb{D}$, let Δk denote a learnable offset of the kernel $\mathbf{W}$, then deformable kernel convolution layer can be formulated as:

$$\mathbb{D} = \sum_{k \in K} \mathbf{W}_{k+\Delta k} \cdot f([F_n, F_r])(p + p_k + \Delta p_k), n \in [t - N, t + N], n \neq r \tag{11}$$

According to [8], deformable convolution can only adapt theoretical receptive fields by deforming the conventional convolution, but it cannot evaluate the contribution of each grid point. As a complementary tool to deformable convolution [4],[42], deformable kernel [8] can weigh the contribution of each grid point to inform the network which point is more important (i.e., adaptive control of effective receptive fields). The advantage of combining deformable convolution with deformable kernel is to not only deform the convolution for extracting key grid points but also adaptively weigh the importance of each point (similar to the introduction of attention mechanism [32]). This way, the deformable convolution kernel layers will be more sensitive to the key feature points than traditional convolution layers and capable of extracting richer information to improve the alignment accuracy and reconstruction quality for our VSR task. Note that previous work such as EDVR [35] only studies the benefit of deformable convolution in VSR; while deformable kernel [8] was originally designed for high-level vision tasks such as object detection and classification. To the best of our knowledge, this work is the first to leverage of idea of combining deformable convolution with deformable kernel into the application of VSR.

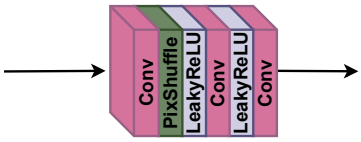

Fig. 4. The details of upscale module, the last Conv layer has only 3 feature maps output in order to generate RGB color frame.

3.3 Reconstruction Module

To get the super-resolved frame $\hat{Y}_r$, the output F_{fusion} from the DKC_Align module is fed into the reconstruction module. The reconstruction module includes several stacked RCAB blocks and the DKSA module (please refer to Fig. 3 (a)):

$$F_{recon} = E_{DKSA}(E_{RCABs}(F_{fusion})) + F_{fusion} \tag{12}$$

where F_{recon} is the final reconstruction features to be fed into upscale module (the architecture of upscale module is shown in Fig. 4 which includes several Conv layers, PixelShuffle and LeakyReLU), E_{RCABs} and E_{DKSA} denote the RCAB blocks and DKSA module. Note that RCAB module has the same structure as it proposed in RCAN [41] which includes resblock [10] and channel attention mechanism [12],[41],[39].

Deformable Kernel Spatial Attention Module In order to further calibrate output feature maps, we propose to construct a new Deformable Kernel based Spatial Attention (DKSA) module instead of traditional spatial attention mechanism. As shown in Fig. 3 (b), in DKSA, we first use one Conv layer to extract the output of the stacked RCAB blocks, then a couple of stacked Deformable Kernel Convolution (DKC) layers are placed to further extract key features from the naive feature map. As discussed in section 3.2, deformable kernels can better measure the effective receptive field than standard convolution kernels. Therefore, DKSA can generate improved spatial attention maps to enforce networks pay more attention to important features such as edges and textures. Note that Fig. 3 (c) shows a light version of DKSA which is used by the level-1 reconstruction module.

4 Experimental Results

In this section, we demonstrate the training and test datasets, network setting, training details, experimental results and ablation study of proposed video extreme super-resolution approach.

4.1 Datasets

In this work, the training data we have used is Vid3oC [16] dataset provided by AIM2020 Video Extreme Super-Resolution Challenge. The Vid3oC dataset

Video Name	Scale	Bicubic	EDVR	DKSAN (ours)
		PSNR (dB)	PSNR (dB)	PSNR (dB)
050	x16	25.36	26.75	**29.17**
051	x16	23.20	23.76	**24.72**
052	x16	20.57	20.92	**21.61**
053	x16	21.61	22.15	**22.63**
054	x16	20.08	20.56	**21.15**
055	x16	20.01	20.36	**21.48**
056	x16	21.44	21.33	**22.54**
057	x16	20.22	20.33	**21.66**
058	x16	19.55	19.80	**21.45**
059	x16	20.22	20.92	**21.90**
060	x16	20.13	20.30	**21.38**
061	x16	21.08	21.58	**22.22**
062	x16	21.54	21.58	**23.12**
063	x16	21.54	22.00	**23.26**
064	x16	20.46	21.04	**21.94**
065	x16	22.53	23.41	**24.80**
Average	x16	21.22	21.67	**22.81**
Parameters	-	-	20.6M	29.5M
Runtime(s/f)	-	-	0.87	0.95

Table 1. Quantitative comparison on Vid3oC dataset for scaling factor of 16; s/f means seconds per frame. **Bold** font indicates the best result.

Video Name	Scale	Bicubic	EDVR	DKSAN (ours)
		PSNR (dB)	PSNR (dB)	PSNR (dB)
050	x16	21.56	21.81	**23.06**
051	x16	23.02	24.13	**24.92**
052	x16	29.56	29.33	**31.87**
053	x16	24.05	24.51	**25.09**
054	x16	31.34	33.15	**36.18**
055	x16	24.39	25.01	**26.88**
056	x16	31.16	31.93	**34.22**
057	x16	34.35	35.20	**39.75**
058	x16	36.00	37.36	**38.15**
059	x16	30.49	31.37	**34.17**
Average	x16	28.59	29.38	**31.43**

Table 2. Quantitative comparison on IntVID dataset for scaling factor of 16. **Bold** font indicates the best result.

includes 50 videos for training, 16 sequences with 120 frames each for validation and 16 sequences with 120 frames each for testing. Note that the ground-truth of testing data are not released. Therefore, in this paper, we only show the validation results for Vid3oC dataset. In order to evaluate the validity of our network, we choose 10 videos (050 to 059) from another dataset, IntVID [16], as a secondary test dataset. For each video, we extract 14 consecutive frames for testing.

4.2 Implementation Details

In the proposed DKSAN networks, to compare with EDVR, we set the kernel size as 3×3 with 128 filters for most of Conv layers, all deformable kernel layers and all deformable convolution layers. The kernel size of feature fusion layers is 1×1. The reduction ratio of channel attention module is still $r = 16$ as [41]. 5 resblocks are in feature extractor. The number of RCAB blocks are set to 30, 20, 15, 10 for each level (from 1 to 4) of reconstruction module. The PixelShuffle layer is the same as [27]. The last Conv layer filter is set to 3 in order to output color frames.

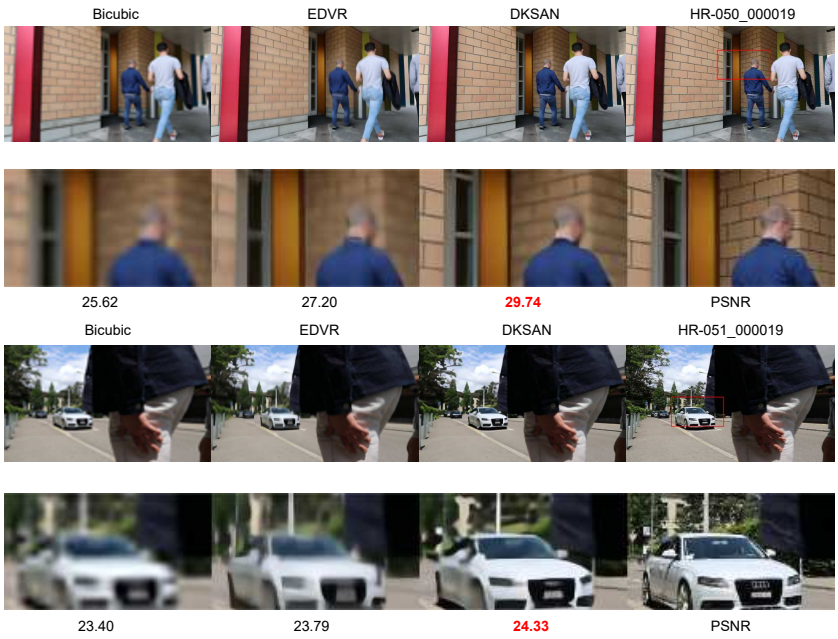

Fig. 5. Visual comparison results among competing approaches for Vid3oC dataset (video 050 and 051) at a scaling factor of 16.

In particular, we randomly crop the 7 low-res frames as small patches with the size of 32×32, and crop the corresponding 4th high-res frames with the size

of 512×512. The batch size is 16. We augment the training set by random flips and rotations. The optimizer we used is ADAM [17] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to 4×10^{-4}. The total training step is $115k$. The loss function we used is adapted Charbonnier penalty function [18]. The loss can be defined as eqn. 13 shown as follows:

$$Loss = \sqrt{||\hat{Y}_r - Y_r||^2 + \xi^2} \tag{13}$$

where $\xi = 1 \times 10^{-3}$, $\hat{Y}_r$ is super-resolved frame and Y_r is target frame (ground-truth). All experiments are trained on 4 NVIDIA Titan Xp GPUs with PyTorch framework Implementation.

Note that for fair comparison, we retrain EDVR with the same training dataset (Vid3oC) and keep most of EDVR setting as the same as the original implementation to run the experiment except setting the upscale module from factor 4 to factor 16 in order to make sure EDVR can generate extreme super-resolved frames.

4.3 Comparison Against State-of-the-Art

Because few existing works related to video extreme super-resolution (with a scale factor of 16), in this work, we have compared our proposed network against with Bicubic interpolation and state-of-the-art EDVR.

Table 1 shows the PSNR comparison results, number of parameters and running time (seconds per frame) of our approach with the competing methods, Bicubic interpolation and EDVR with the scaling factor of 16 on the validation set of Vid3oC [16]. From the Table, we can see that our DKSAN method has the best PSNR scores for all 16 testing videos. The significant PSNR gains (up to $2.4dB$) over previous state-of-the-art method EDVR. Since PSNR metrics cannot always evaluate the subjective quality of images, therefore, a qualitative result is shown in Fig. 5, we can easily observe that our proposed network DKSAN can better reconstruct the lines on the wall for "050_000019" and a clearer car for "051_000019" compared with EDVR.

To further verify the effectiveness of our proposed method, we selected another dataset, IntVid [16] as a secondary test dataset. From Table 2, we can easily find out that our proposed DKSAN has the best performance for all 10 testing videos compared with EDVR and bicubic interpolation. A qualitative result is shown in Fig. 6. For the subject "'050_0010", compared with EDVR, our DKSAN can better recover more details of the rear wing. Taking another example, in "054_0007", our DKSAN can reconstruct a much clearer face than EDVR does.

4.4 Ablation Studies

To investigate the effect of proposed DKC_Align module and DKSA module, we have conducted different strategies to remove the certain components from the

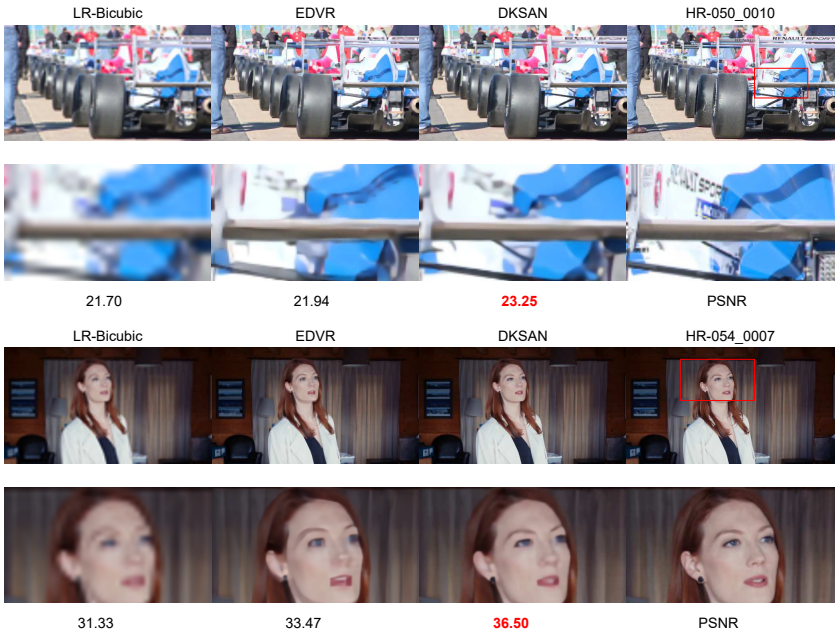

Fig. 6. Visual comparison results among competing approaches for IntVID dataset (video 050 and 054) at a scaling factor of 16.

final framework DKSAN. In particular, we have implemented four competing models for our ablation studies: 1) training with only resblocks, without channel attention, alignment and DKSA; 2) training without DKC_Align and DKSA module; 3) training with DKC_Align module but without DKSA module; 4) training with all modules (proposed DKSAN). Note that all experiments are trained under same dataset and conditions for fair comparison.

Table 3 shows the results, the number of parameters and running time (seconds per frame) of all four strategies mentioned previously with PSNR scores of each video and average. The backbone result is running only based on resblock, no channel attention, alignment and DKSA applied. From the results, we can see that the backbone has the worst performance; adding channel attention module but without DKC_Align and DKSA modules, the result is only 31.27 dB; after adding DKC_Align module, the result is improved to 31.32 dB; finally, we observe that after adding DKSA module (the full version of DKSAN), the result is further improved to 31.43 dB (**0.4dB** and **0.16dB** gained when compared with backbone and w/o Alignment & DKSA respectively) because of the effective module DKSA.

Video Name	Backbone	w/o Alignment & DKSA	w/o DKSA	DKSAN (ours)
	PSNR (dB)	PSNR (dB)	PSNR (dB)	PSNR (dB)
050	22.80	22.98	22.87	**23.06**
051	24.72	24.89	24.89	**24.92**
052	31.65	31.75	31.85	**31.87**
053	25.05	25.07	**25.18**	25.09
054	35.30	35.73	35.86	**36.18**
055	26.52	**26.89**	26.69	26.88
056	33.79	33.92	**34.33**	34.22
057	38.97	39.13	39.27	**39.75**
058	38.09	38.21	**38.30**	38.15
059	33.38	34.15	34.16	**34.17**
Average	31.03	31.27	31.32	**31.43**
Parameters	26.1M	26.3M	27.0M	29.5M
Runtime(s/f)	0.83	0.89	0.92	0.97

Table 3. Ablation Studies for DKSAN on IntVID dataset for scaling factor of 16. Backbone means only resblocks used; w/o Alignment & DKSA means DKC_Align and DKSA Module are not applied; w/o DKSA means only the DKSA module is not applied; s/f means seconds per frame. **Bold** font indicates the best result.

4.5 AIM 2020 Video Challenge

We have participated in the AIM2020 video extreme super-resolution challenges which is the second edition of AIM2019 challenges [6]. Our submissions won the **2nd place** for both track 1 and track 2 competitions. Note that track 1 is based on PSNR performance and track 2 is based on perceptual (see the AIM2020 challenge report [7] for more details).

5 Conclusions

In this work, we proposed a multi-frame based VSR networks DKSAN for extreme low-resolution videos. The novel temporal alignment module, DKC_Align, can help the networks to better learn and align the detailed features by improving both theoretical and effective receptive fields between reference frame and its neighboring frames. Furthermore, the DKSA module calibrated the reconstructed features to further enhance the edges and textures at the spatial domain. Thanks to the newly designed DKC_Align and DKSA modules, the proposed architecture can reconstruct high-quality HR frames from extreme LR frames and significantly improve both objective and subjective performance when compared with state-of-the-art approach EDVR [35].

Acknowledgment

This work is partially supported by the NSF under grants IIS-1908215 and OAC-1839909, the DoJ/NIJ under grant NIJ 2018-75-CX-0032, and the WV Higher Education Policy Commission Grant (HEPC.dsr.18.5).

References

1. Bertasius, G., Torresani, L., Shi, J.: Object detection in video with spatiotemporal sampling networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 331–346 (2018)
2. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4778–4787 (2017)
3. Cheng, G., Matsune, A., Li, Q., Zhu, L., Zang, H., Zhan, S.: Encoder-decoder residual network for real super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
4. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 764–773 (2017)
5. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE transactions on image processing **13**(10), 1327–1344 (2004)
6. Fuoli, D., Gu, S., Timofte, R., Tao, X., Li, W., Guo, T., Deng, Z., Lu, L., Dai, T., Shen, X., Xia, S., Dai, Y., Jia, J., Yi, P., Wang, Z., Jiang, K., Jiang, J., Ma, J., Zhong, Z., Wang, C., Jiang, J., Liu, X.: AIM 2019 challenge on video extreme super-resolution: Methods and results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3467–3475 (2019)
7. Fuoli, D., Huang, Z., Gu, S., Timofte, R., et al.: AIM 2020 challenge on video extreme super-resolution: Methods and results. In: European Conference on Computer Vision Workshops (2020)
8. Gao, H., Zhu, X., Lin, S., Dai, J.: Deformable kernels: Adapting effective receptive fields for object deformation. In: International Conference on Learning Representations (2020), `https://openreview.net/forum?id=SkxSv6VFvS`
9. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Horn, B.K., Schunck, B.G.: Determining optical flow. In: Techniques and Applications of Image Understanding. vol. 281, pp. 319–331. International Society for Optics and Photonics (1981)
12. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
13. Isobe, T., Li, S., Jia, X., Yuan, S., Slabaugh, G., Xu, C., Li, Y.L., Wang, S., Tian, Q.: Video super-resolution with temporal group attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8008–8017 (2020)

14. Jo, Y., Wug Oh, S., Kang, J., Joo Kim, S.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3224–3232 (2018)
15. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video super-resolution with convolutional neural networks. IEEE Transactions on Computational Imaging **2**(2), 109–122 (2016)
16. Kim, S., Li, G., Fuoli, D., Danelljan, M., Huang, Z., Gu, S., Timofte, R.: The Vid3oC and IntVID datasets for video super resolution and quality mapping. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 3609–3616. IEEE (2019)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Lai, W.S., Huang, J.B., Ahuja, N., Yang, M.H.: Deep laplacian pyramid networks for fast and accurate superresolution. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, p. 5 (2017)
19. Liao, R., Tao, X., Li, R., Ma, Z., Jia, J.: Video super-resolution via deep draft-ensemble learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 531–539 (2015)
20. Liu, C., Sun, D.: On bayesian adaptive video super resolution. IEEE transactions on pattern analysis and machine intelligence **36**(2), 346–360 (2013)
21. Liu, D., Wang, Z., Fan, Y., Liu, X., Wang, Z., Chang, S., Huang, T.: Robust video super-resolution with learned temporal dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2507–2515 (2017)
22. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: Advances in neural information processing systems. pp. 4898–4906 (2016)
23. Ma, Z., Liao, R., Tao, X., Xu, L., Jia, J., Wu, E.: Handling motion blur in multi-frame super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5224–5232 (2015)
24. Matsuo, Y., Sakaida, S.: Super-resolution for 2k/8k television using wavelet-based image registration. In: IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 378–382 (2017)
25. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6626–6634 (2018)
26. Seibel, H., Goldenstein, S., Rocha, A.: Eyes on the target: Super-resolution and license-plate recognition in low-quality surveillance videos. IEEE access **5**, 20020–20035 (2017)
27. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883 (2016)
28. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4472–4480 (2017)
29. Tekalp, A.M.: Digital video processing. Prentice Hall Press (2015)
30. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: TDAN: Temporally-deformable alignment network for video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)

31. Umeda, S., Yano, N., Watanabe, H., Ikai, T., Chujoh, T., Ito, N.: HDR video super-resolution for future video coding. In: 2018 International Workshop on Advanced Image Technology (IWAIT). pp. 1–4 (2018)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
33. Wang, H., Su, D., Liu, C., Jin, L., Sun, X., Peng, X.: Deformable non-local network for video super-resolution. IEEE Access **7**, 177734–177744 (2019)
34. Wang, W., Ren, C., He, X., Chen, H., Qing, L.: Video super-resolution via residual learning. IEEE Access **6**, 23767–23777 (2018)
35. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: EDVR: Video restoration with enhanced deformable convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
36. Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J.P., Xu, C.: Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3370–3379 (2020)
37. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. pp. 802–810 (2015)
38. Xu, X., Li, X.: SCAN: Spatial color attention networks for real single image super-resolution. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2024–2032 (2019)
39. Xu, X., Ye, Y., Li, X.: Joint demosaicing and super-resolution (JDSR): Network design and perceptual optimization. IEEE Transactions on Computational Imaging pp. 1–1 (2020)
40. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
41. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
42. Zhu, X., Hu, H., Lin, S., Dai, J.: Deformable convnets v2: More deformable, better results. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9308–9316 (2019)