Video Super-Resolution via a Spatio-Temporal Alignment Network

Weilei Wen, Wenqi Ren, *Member, IEEE*, Yinghuan Shi, Yunfeng Nie, Jingang Zhang, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Deep convolutional neural network based video super-resolution (SR) models have achieved significant progress in recent years. Existing deep video SR methods usually impose optical flow to wrap the neighboring frames for temporal alignment. However, accurate estimation of optical flow is quite difficult, which tends to produce artifacts in the super-resolved results. To address this problem, we propose a novel end-to-end deep convolutional network that dynamically generates the spatially adaptive filters for the alignment, which are constituted by the local spatio-temporal channels of each pixel. Our method avoids generating explicit motion compensation and utilizes spatio-temporal adaptive filters to achieve the operation of alignment, which effectively fuses the multi-frame information and improves the temporal consistency of the video. Capitalizing on the proposed adaptive filter, we develop a reconstruction network and take the aligned frames as input to restore the high-resolution frames. In addition, we employ residual modules embedded with channel attention as the basic unit to extract more informative features for video SR. Both quantitative and qualitative evaluation results on three public video datasets demonstrate that the proposed method performs favorably against state-of-the-art super-resolution methods in terms of clearness and texture details.**

***Index Terms*— Video super-resolution, temporal consistency, spatio-temporal adaptive filters.**

Manuscript received July 10, 2021; revised December 15, 2021; accepted January 7, 2022. Date of publication February 1, 2022; date of current version February 8, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102503; and in part by the National Natural Science Foundation of China under Grant 62172409, Grant 61971016, and Grant U1803264. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nikolaos Mitianoudis. *(Corresponding authors: Wenqi Ren; Jingang Zhang.)*

Weilei Wen is with the TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: wenwlmail@163.com).

Wenqi Ren is with the School of Cyber Science and Technology, Sun Yat-sen University at Shenzhen, Shenzhen 518107, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: rwq.renwenqi@gmail.com).

Yinghuan Shi is with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China (e-mail: syh@nju.edu.cn).

Yunfeng Nie is with the Brussels Photonics, Department of Applied Physics and Photonics, Vrije Universiteit Brussel, 1050 Brussels, Belgium (e-mail: ynie@b-phot.org).

Jingang Zhang is with the Intelligent Imaging Center, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: zhangjg@ucas.ac.cn).

Xiaochun Cao is with the School of Cyber Science and Technology, Sun Yat-sen University at Shenzhen, Shenzhen 518107, China (e-mail: caoxiaochun@iie.ac.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2022.3146625, provided by the authors.

Digital Object Identifier 10.1109/TIP.2022.3146625

I. Introduction

As THE demand for the resolution of pictures and videos in real scenes is getting higher and higher, the super-resolution (SR) of low-resolution (LR) videos becomes very important. For example, more and more ultra-high-definition televisions are introducing higher resolutions. However, video content that can catch up with the display capability of monitors is still relatively scarce, so the industry is paying more and more attention to video super-resolution.

Video SR aims to recover a high-resolution (HR) video frame from its corresponding low-resolution (LR) frame and the adjacent LR frames. In some high-level tasks, such as target recognition [1] and scene segmentation [2], super-resolution acts as a low-level auxiliary function, enabling the high-level tasks to be completed more accurately. Most of the existing video SR methods extract frames and then perform reconstruction through single image super-resolution (SR) algorithms. However, the results are always unsatisfactory since single image SR approaches cannot consider the information between frames, resulting in temporal inconsistency.

With the development of deep convolutional neural networks (CNNs), many image processing tasks, including image super-resolution [3]–[7], image deblurring [8], image denoising [9], have achieved significant improvements. The success of CNNs in image processing further advances the development of video SR. However, most CNNs-based SR algorithms only use the spatial correlation of each frame and cannot fully use the consecutive information. In contrast, the temporal relationship between consecutive frames is crucial to video super-resolution (VSR).

Previous VSR (or multi-image SR) methods align reference LR frame and neighboring frames with explicit motion compensation [11]–[13]. However, these approaches recover HR image count heavily on the accurate motion estimation. The deviation in the estimated motion compensation or wrapping operation may result in artifacts in the neighboring frames. Another possible problem with this type of approach is that the recovered HR frame, which is produced by mixing multiple motion-compensated LR frames through the convolutional layers, may contain a blurry phenomenon [13].

To alleviate the above problem, we propose a spatio-temporal alignment network (STAN) to impose

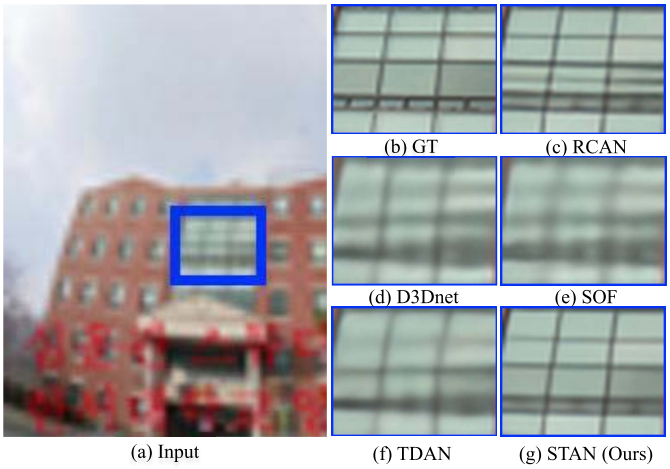


Fig. 1. VSR results on the "00006-0811" clip in the Vimeo-90k-T dataset [10]. The result generated by our method is clearer and closer to the ground truth than the state-of-the-arts. In particular, incorrect meshes were generated in the super-resolved result by RCAN.

one-stage spatio-temporal alignment in this paper. Unlike previous optical flow based methods [10], [12]–[14] which wrap the neighboring frames for alignment in an explicit motion estimation way, our algorithm can adaptively align the reference frame and neighboring frames at the feature level. Specifically, our STAN embeds the adaptive filters generated by the spatio-temporal neighborhood of each pixel in LR frames to align the neighboring LR features with the reference feature. Thus, for different reference LR feature pixels, the proposed network adaptively generates different filter kernels, allowing for maximum flexibility and capability to deal with various motions in temporal scenes. Furthermore, since our proposed STAN does not depend on explicit motion compensation and the mix of multiple frames, our method can generate realistic details and preserve temporally consistent in the recovered HR videos.

In addition, we propose a residual attention group (RAG) to exploit the multi-level and channel-wise features within the proposed STAN. The proposed RAG contains the channel attention mechanism and uses skip connections to prompt the fusion of information at different scales, which can adaptively learn to pay attention to informative features at different channels but also fully and effectively exploit the hierarchical features to maintain persistent memory.

In summary, our contributions in this paper are summarized as follows:

1) We propose a novel spatio-temporal alignment network (STAN) for feature-level alignment, which avoids explicit motion compensation for frame alignment and can explore more helpful inter-frame information through implicit alignment.
2) We construct an end-to-end efficient video SR architecture based on the proposed STAN. It integrates the video frame alignment and super-resolution into a unified framework without explicit flow estimation.
3) We develop a residual attention group (RAG) block to adaptively recalibrate the feature responses by explicitly modeling channel-wise feature interdependencies, and enhance the feature representation of the network by fusing information at different scales through share source skip connections.

We quantitatively and qualitatively evaluate the proposed network on various benchmark datasets. Extensive experiments demonstrate that our algorithm performs favorably against the state-of-the-art video SR approaches.

II. Related Work

In this section, we briefly review the relevant work of single image SR and video super-resolution (VSR). Since our proposed method is based on deep CNNs, we only focus on deep learning based approaches.

A. Single-Image Super-Resolution

To address the ill-posed SR problem, early methods employ interpolation techniques based on sampling theory [15]–[17]. However, those approaches show limitations in recovering detailed and realistic textures. Previous studies [18], [19] adopted natural image statistics to improve image quality. Since Dong *et al.* [20] first proposed the CNN-based super-resolution architecture (SRCNN), various more efficient and effective networks are presented. Kim *et al.* [21] propose a VDSR network that employs 20 layers to learn a residual mapping between the HR output and LR input. EDSR [4] shows that batch normalization layers do not affect super-resolution. Therefore, they remove batch normalization layers in the residual network to improve the SR effect. Although the idea of residuals theoretically allows the network to deepen indefinitely, the number of parameters and the computational effort of the CNN increases very fast as the network deepens. SRDenseNet [22] and RDN [23], [24] make full use the hierarchical features with dense connections, but the computational effort has also increased significantly. Drawing on the idea of recursion, DRRN [3] and DRCN [25] are proposed based on a recursive convolutional network to alleviate the phenomenon of gradient explosion and reduce the complexity of deep models. To handle the task of super-resolution with large scaling factors, some algorithms are proposed using a step-by-step method. For example, Lai *et al.* [26], [27] propose the LapSR network by employing a pyramidal framework to progressively generate $\times 8$ images.

In addition, SR algorithms based on generative adversarial networks (GANs) have also been proposed, and the generated results are more pleasant in terms of visual perception [4]. SRGAN [5] uses a GAN to predict high-resolution outputs by introducing a multi-task loss, which contains three loss functions: (1) an MSE loss; (2) a perceptual loss [28]; (3) an adversarial loss [29]. EnhanceNet [30] is another super-resolution network trained based on GAN. To avoid over-smoothing caused by relying only on the MSE loss, EnhanceNet utilizes two other loss functions: a) perceptual loss to constrain intermediate features, and b) texture matching loss to match low- and high-resolution maps and ensure local consistency of texture information. However, the algorithm may produce artifacts in texture-rich regions of the image. ESRGAN [30]

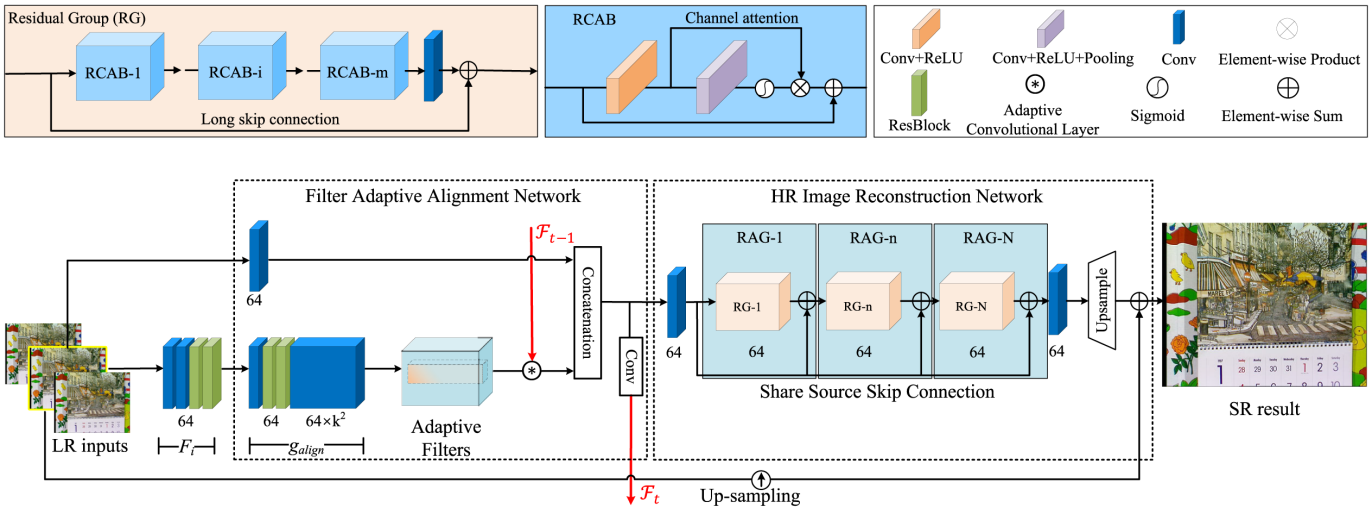


Fig. 2. Architecture of the proposed STAN. It consists of three main parts: the feature extraction module, the filter adaptive alignment module (FAAM), and the SR reconstruction module. Given three consecutive frames (I^L_{t-1}, I^L_t, and I^L_{t+1}), the features extracted from them are fed to the filter generation network g_{align} to generate adaptive filters $\mathcal{G}_\theta$. Then, employing the proposed adaptive convolutional layer, $\mathcal{G}_\theta$ aligns the previous time step features $\mathcal{F}_{t-1}$ with the current time step features. At last, a high-resolution image is recovered from the fused features using the HR image reconstruction network. Note that we only show one alignment module in the framework. In our implementation, three cascaded adaptive convolutional layers are used for alignment.

replaces the residual module in SRGAN with the residual dense blocks and uses the relativistic GAN loss so that the network learns to determine whether one image is more realistic than another, guiding the generator to recover more detailed textures. Another line of super-resolution research is based on attention mechanisms [31]. RCAN [32] and SAN [6] introduce channel attention to address the SR task, which outperformed previous networks by a substantial margin. They have blazed a new trail in single image SR. However, only the channel attention weights are considered in RCAN, the similarity of features between different convolutional layers is not exploited. To address this problem, Niu *et al.* [33] propose the layer attention and channel-spatial attention modules to extract informative features in the SR network.

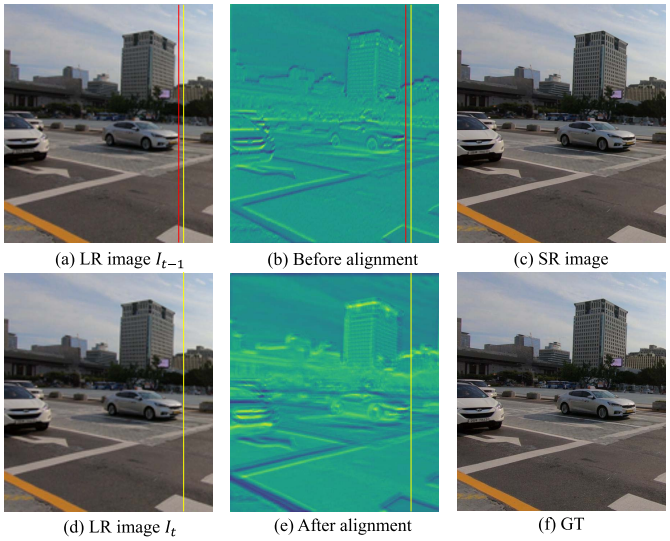
Although single-image SR algorithms achieve impressive results on public SR benchmarks [34]–[39], there are two issues in reconstructing high-resolution video. First, the information between adjacent frames cannot be fully utilized. Second, the reconstructed high-resolution video has obvious jitter and flicker, destroying the video consistency. Therefore, the video SR algorithms based on multi-frame fusion came into being.

B. Video Super-Resolution

We divide video SR algorithms into two categories: optical flow-based and implicit alignment approaches without optical flow estimation. There is also a type of methods that do not use alignment operations, such as RNN or 3D convolution based methods [40] that directly extract the temporal information of the video. Since these methods have many parameters that cannot process the video in real-time or have more limitations in extracting temporal information, this paper does not elaborate too much on this type of methods.

1) Video SR via Optical Flow: Liao *et al.* [41] use multiple optical flow estimation methods to predict SR drafts and then generate the final high-quality frame by a draft-ensemble network. Kappeler *et al.* [42] propose a two-stage strategy (VSR-net). This method estimates the optical flow between input LR frames and then warps neighboring frames to the reference frame by a spatial transformer. Finally, an HR frame is reconstructed through another deep network. Since this method is divided into motion estimation and motion compensation networks, the two-stage method greatly increases the time consumption, making it impossible to use in real scenes. VESPCN [11] uses a multi-scale spatial transformer network which consists of a coarse flow estimation and a fine flow estimation to compensate for the motion between frames. Since this method performs optical flow estimation and warping operation at different scales, each operation's errors are superimposed and lead to a final result with severe artifacts. Tao *et al.* [12] utilize a motion compensation transformer module to estimate optical flow, and propose a sub-pixel motion compensation (SPMC) module for simultaneous motion compensation and resolution enhancement. Sajjadi *et al.* [14] propose a recurrent framework that uses the HR estimation of the previous frame for generating the subsequent frame, which embeds the optical flow for motion compensation. Xue *et al.* [10] exploit a task-oriented flow estimation network to calculate the motion between frames, then warping input frames to the reference frames using a spatial transformer network, and aggregate the warped frames to generate an HR frame. Wang *et al.* [13] propose an optical flow reconstruction network to infer HR optical flows in a coarse-to-fine way to achieve motion compensation. The compensated LR inputs are fed into an SR network to produce high-quality video frames.

The above methods usually use optical flow for explicit temporal alignment. However, it is difficult to generate accurate optical flow, and flow warping operation also tends to introduce artifacts into the aligned frames.



(a) LR image I_{t-1} (b) Before alignment (c) SR image

(d) LR image I_t (e) After alignment (f) GT

Fig. 3. Effectiveness of the proposed filter adaptive alignment module. (a) and (d) are the low-resolution images of the previous frame I_{t-1} and the current frame I_t, respectively. (b) and (e) are the selected feature maps before and after FAAM alignment, respectively. (c) and (f) are the SR reconstructed image and ground truth of the current frame, respectively.

2) Video SR Without Explicit Motion Compensation: Huang *et al.* [43] exploit recurrent neural networks to capture long-term contextual information of sequential frames, which can avoid explicit motion estimation and compensation. However, this method fails to tackle large displacements and complicated motions.

The DUF [44] method uses 3D convolution to obtain spatio-temporal information. Then the dynamic up-sampling filter fuses spatio-temporal information and enlarges feature size to avoid explicit motion estimation and compensation. Besides, DUF employs another network to estimate the residual graph of the target frame and enhance the high-frequency information of the reconstructed results. TDAN [45] is introduced to exploit deformable convolutions to align the input frames at the feature level. However, this method is sensitive to the input patterns and may cause noticeable reconstruction artifacts due to unfaithful offsets. Like the TDAN algorithm, EDVR [46] also utilizes the deformable convolution for the adjacent frame alignment operation. They design the alignment module named PCD as a pyramidal cascade structure, which aligns the adjacent frames at different scales. Although the PCD module can align adjacent edges coarse-to-finely, the estimated errors are also superimposed at different levels, leading to artifacts.

In this paper, we generate the adaptive filters by the spatio-temporal neighborhood of each pixel in LR frames for aligning LR features with the reference. Extensive experiments demonstrate that our algorithm performs favorably against several state-of-the-art video SR approaches.

III. Our Method

In this section, we design an end-to-end trained video super-resolution network. The architecture of the proposed network can be found in Figure 2.

TABLE I

Temporal Consistency and Computational Efficiency Achieved on the Vid4 [47] Dataset. Params Represents the Number of Parameters. FLOPs is Computed Based on LR Frames With a Resolution of 320×180. Best Results are Shown in **Bold**. '–' Means That It is not Available

Method	T-MOVIE ↓	MOVIE ↓	Params (M) ↓	FLOPs (G) ↓	Runtime ↓
DUF [45]	30.85	6.68	5.82	1657.40	873*ms*
VESPCN [11]	18.81	4.49	–	–	–
DBPN [49]	21.43	5.50	10.43	5213.0	–
TGA [50]	18.77	4.53	7.06	700.10	–
SOF [13]	19.35	4.25	**1.64**	**108.90**	**108*ms***
TDAN [46]	18.87	4.11	1.97	288.02	2654*ms*
D3Dnet [51]	15.45	3.38	2.58	408.82	491*ms*
EDVR [47]	9.36	2.04	20.70	2492.18	336*ms*
BasicVSR [52]	**6.16**	**1.44**	6.29	2605.34	304*ms*
RCAN [33]	23.49	5.98	15.59	919.20	122*ms*
STAN	15.48	3.81	16.16	979.67	155*ms*

A. Network Overview

Given $2N+1$ consecutive low-resolution video frames $\{I^L_{t-N}, \ldots, I^L_{t-1}, I^L_t, I^L_{t+1}, \ldots, I^L_{t+N}\}$ as the input of the proposed network, our goal is to predict the high-resolution frame I^H_t from the reference LR frame I^L_t and 2N neighboring frames $\{I^L_{t-N}, \ldots, I^L_{t-1}, I^L_{t+1}, \ldots, I^L_{t+N}\}$. In this paper, we propose a spatio-temporal alignment network (STAN). It mainly consists of three parts: a feature extraction module, a Filter Adaptive Alignment Module (FAAM) to align the neighboring frames to the reference frame, and an HR image reconstruction network to restore the HR frame from LR frames.

B. Feature Extraction

Our proposed STAN takes the reference LR frame I^L_t and $2N$ neighboring frames as inputs to restore the corresponding HR output I^H_t. The $2N+1$ consecutive LR frames are first fed into the feature extraction module. The feature extraction module consists of two convolutional layers and k residual blocks which take LeakyReLU as the activation function. In our implementation, we imposed a modified residual block from EDSR [4] to extract feature maps of the LR inputs,

$$F_0 = \text{Conv}\left(\text{Concat}\left(I^L_{t-N}, \ldots, I^L_{t-1}, I^L_t, I^L_{t+1}, \ldots, I^L_{t+N}\right)\right), \tag{1}$$

and

$$F_i = RB_i\left(F_{i-1}\right), \quad i = 1, 2, \ldots n, \tag{2}$$

where F_0 is the shallow features extracted by the two convolution layers, F_i is the output of the i-th Resblock RB_i.

C. Filter Adaptive Alignment Module

In a recurrent neural network, the output of the current iteration of the network contains the information of the previous frame of the video. How to use the information in the previous frames to restore the current frame information is a key issue in video super-resolution. Unlike the previous optical flow and

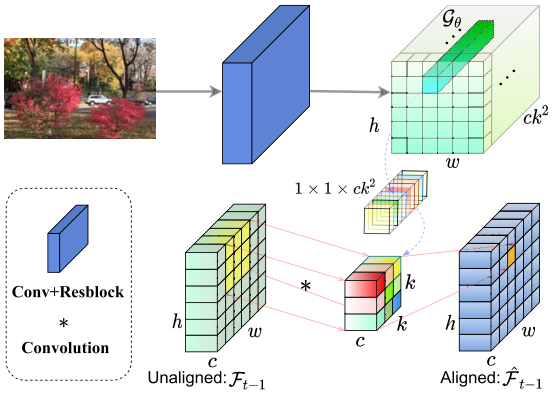


Fig. 4. Pipeline of the feature alignment process in the proposed filter adaptive alignment module (FAAM). The features extracted from the current frame I^t are used to generate the dynamic filter $\mathcal{G}_\theta$. Then we reshape $\mathcal{G}_\theta$ to the size of $h \times w \times c \times k^2$, so that it can be applied to each position of $\mathcal{F}_{t-1}$. Through the adjustment of the dynamic filter, the features $\mathcal{F}_{t-1}$ extracted from the previous frame can be well aligned with the features of the current frame.

deformable convolution-based alignment methods that may generate incorrect optical flow or cause noticeable reconstruction artifacts, we propose a Filter Adaptive Alignment Module (FAAM) for image alignment in the feature domain. FAAM avoids explicit motion compensation to explore more useful inter-frame information through implicit alignment.

The alignment module FAAM consists of two components, the filter generation network and the adaptive filter convolution layer.

We treat the features $\mathcal{F}_t \in R^{h\times w\times c}$ extracted from the continuous multi-frame images as the input of the filter generation network to generate the filters $\mathcal{G}_\theta$, $\theta \in R^{k^2\times c\times h\times w}$, where h, w, and k are height, width, and the filter size, respectively, c denotes the channel numbers of $\mathcal{F}_{t-1}$ and $\mathcal{F}_t$,

$$\mathcal{G}_\theta = g_{align}\left(F_n\right) = g_{align}\left(RB_n\left(F_{n-1}\right)\right). \tag{3}$$

Then applying the generated filter to the output intermediate features $\mathcal{F}_{t-1} \in R^{h\times w\times c}$ from the previous frame. The filter generation network g_{align} consists of one convolution layer and two residual blocks with kernel size of 3×3 , followed by a convolution layer with kernel size of 1×1 to expand the channels of the output to ck^2. Theoretically, the input of the proposed FAAM is a five-dimensional $(h \times w \times c \times k \times k)$ tensor, we reshape the generated filter $\mathcal{G}_\theta$ $(h \times w \times ck^2)$ into the same dimension as shown in Figure 4. It is worth noting that different filters are applied to different positions of $\mathcal{F}_{t-1}$. For each position (x, y, c_i) of the feature map $\mathcal{F}_{t-1}$, the aligned feature map $\hat{\mathcal{F}}_{t-1}$ is generated by imposing a specific local filter $\mathcal{G}_\theta(x, y, c_i)$ to the region centered around $\mathcal{F}_{t-1}(x, y, c_i)$,

$$\begin{aligned}\hat{\mathcal{F}}_{t-1}\left(x, y, c_i\right) &= \mathcal{G}_\theta\left(x, y, c_i\right) * \mathcal{F}_{t-1}\left(x, y, c_i\right) \\ &= \sum_{u=-r}^{r}\sum_{v=-r}^{r} \mathcal{G}_\theta\left(x, y, k^2c_i + ku + v\right) \\ &\quad \times \mathcal{F}_{t-1}\left(x-u, y-v, c_i\right),\end{aligned} \tag{4}$$

where $r = \frac{k-1}{2}$, $*$ denotes the convolution operation. The adaptive filter is not only channel specific but also position specific. As for each filter in the generated adaptive convolution layer, it should be strengthened if the previous time step features have a similar appearance as the current frame features. Otherwise, it should be repressed. The filter adaptive alignment network is trainable and efficient. Through adjustment of the adaptive filter, the proposed FAAM is able to align the LR features of previous time step $\mathcal{F}_{t-1}$ to the current time step features. In this way, we align and merge the features of the previous frame with the current frame to alleviate occlusion and video consistency issues.

The proposed FAAM produces adaptive filters that capture spatial and temporal consistent information about the neighboring frames and acts as a prior to regularize the features of the previous frame. Without the proposed FAAM to encode the temporal consistency of the features, the network can make erroneous local decisions that lead to artifacts as shown in Figure 3. We can observe that the proposed FAAM plays the role of feature alignment, aligning the features of the previous image frame with the current frame. In order to enhance the feature alignment capability, we concatenate g_{align} and adaptive convolutional layer three times in series. Feeding the features after one alignment into g_{align} to generate new dynamic filtering kernels, and continue to align the output of the features from the previous iteration. The effectiveness of the cascaded alignment modules can be found in Section IV-C.

D. Reconstruction Network

After the alignment of the features, we extract the reference frame feature through a convolutional layer and then send it to the subsequent reconstruction network together with the aligned features. We impose the residual channel attention block (RCAB) from [32] to extract the intermediate features. Since RCAB embeds the channel attention mechanism, the network can ignore irrelevant information and focus on important edges and textures in the features. However, different from [32], we introduce share source skip connections into every Residual Group (RG) which contains 20 RCABs to form our proposed Residual Attention Group (RAG) block. Unlike DenseNet [54] where the features of each layer are passed to all subsequent layers, we propose a share source structure where only the first layer is passed to all subsequent layers. The proposed share source structure ensures that the original features are not lost as the network deepens and reduces the computational effort relative to DenseNet. This structure strikes a balance between information transfer and computation.

The m-th residual block of the g-th RAG can be represented as

$$F_{1,1} = \mathrm{Conv}\left(\mathrm{Concat}\left(\mathrm{Conv}\left(I_t^L\right), \hat{\mathcal{F}}_{t-1}\right)\right) \tag{5}$$

and

$$F_{g,m} = \mathrm{RAG}_{g,m}\left(F_{g,m-1} + F_{1,1}\right), \tag{6}$$

where $F_{1,1}$ is the input of the first RAG, $\mathrm{RAG}_{g,m}$ denotes the function of m-th residual block of the g-th RAG. After fusing the information between adjacent frames through the

TABLE II

AVERAGE PSNR AND SSIM OF SUPER-RESOLVED RESULTS ON THE REDS4 DATASET. BEST RESULTS ARE SHOWN IN **BOLD**

Method	Clip_000 (RGB)		Clip_011 (RGB)		Clip_015 (RGB)		Clip_020 (RGB)		Average (RGB)		Average (Y)	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
VESPCN [11]	25.54	0.7035	27.38	0.7744	29.99	0.8441	26.63	0.7855	27.39	0.7769	28.76	0.7991
SPMC [12]	25.95	0.7246	27.81	0.7878	30.57	0.8591	26.93	0.7993	27.81	0.7927	29.19	0.8136
DBPN [49]	25.97	0.7245	28.68	0.8108	31.10	0.8721	27.43	0.8183	28.30	0.8064	29.68	0.8266
RCAN [33]	26.17	0.7376	29.34	0.8260	31.85	0.8884	27.74	0.8297	28.78	0.8204	30.12	0.8383
FRVSR [14]	26.27	0.7421	28.24	0.7990	30.92	0.8682	27.20	0.8094	28.78	0.8204	29.55	0.8252
TGA [50]	25.97	0.7274	28.71	0.8127	31.21	0.8764	27.39	0.8194	28.32	0.8090	29.67	0.8275
SOF [13]	26.49	0.7547	27.99	0.7963	31.09	0.8707	27.26	0.8114	28.21	0.8083	29.57	0.8276
TDAN [46]	**26.53**	**0.7541**	28.40	0.8037	31.07	0.8704	27.35	0.8145	28.34	0.8106	29.70	0.8301
STAN	26.21	0.7379	**29.35**	**0.8252**	**32.01**	**0.8890**	**27.84**	**0.8309**	**28.85**	**0.8207**	**30.22**	**0.8392**

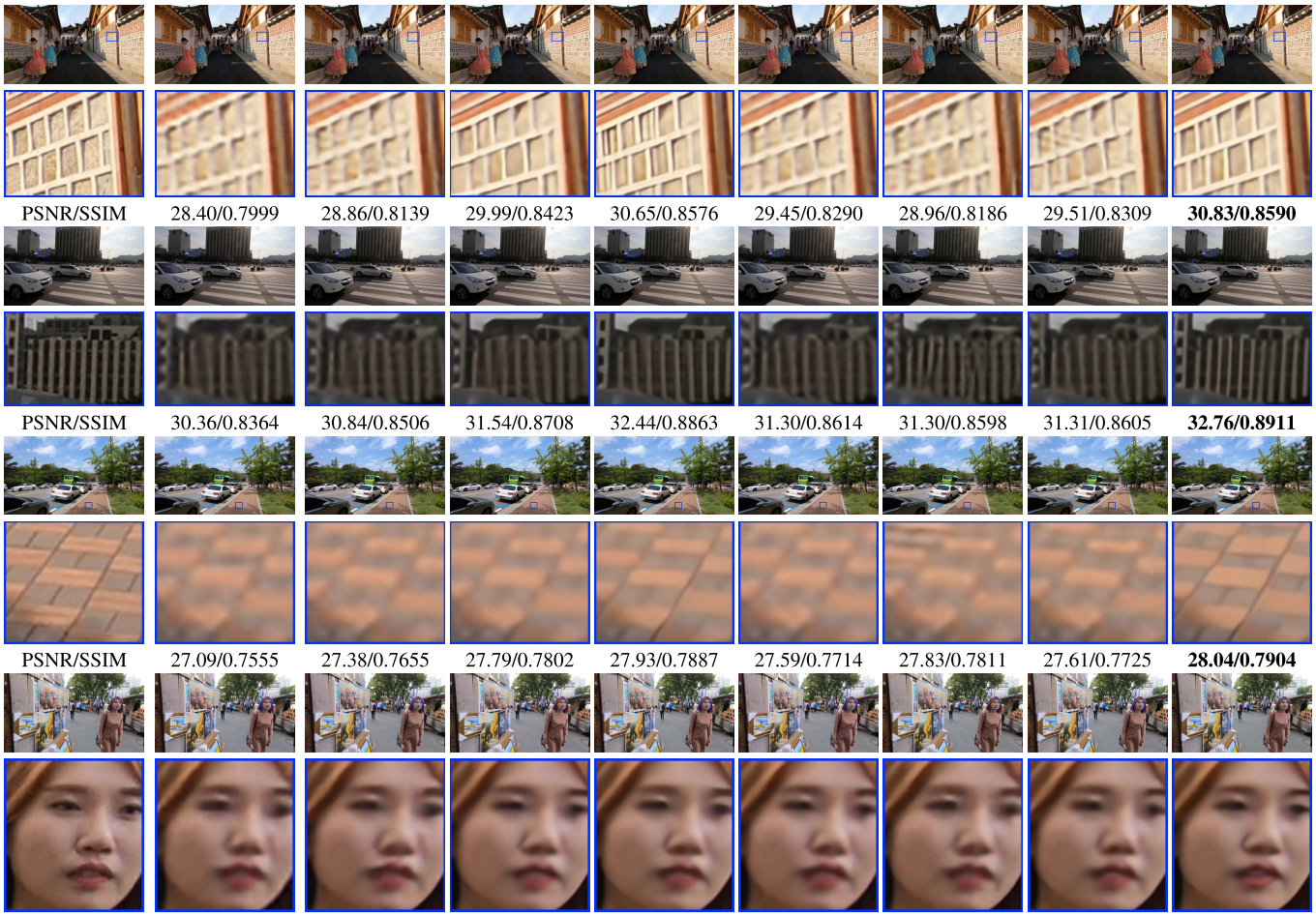


Fig. 5. Visual comparison for ×4 SR with BI degradation model on the REDS4 dataset. The proposed STAN obtains better visual quality and recovers more image details compared with other state-of-the-art SISR and VSR methods.

cascaded RAGs, we use the up-sampling module to generate the SR result. The up-sampling module includes a convolutional layer and a sub-pixel convolutional layer [55], transforming the scale sampling with a given magnification factor via pixel translation. In addition, we impose a long skip connection for transmitting shallow information to the deep layers,

$$F_{rec} = U_{\uparrow}\left(\text{Conv}\left(\text{Add}\left(F_{g,m}, I_{t\uparrow}^{L}\right)\right)\right) \tag{7}$$

where F_{rec} denotes the output of the reconstruction network, $U_{\uparrow}$ is the operation of sub-pixel convolution, $I_{t\uparrow}^{L}$ denotes the up-sampled image from I_t^L by bilinear interpolation.

To optimize our algorithm, we adopt the Charbonnier penalty function [26] to measure the difference between reconstructed frames I_t^H and ground-truth HR video frames I_t^{GT}, which is defined as:

$$\mathcal{L}_{sr} = \sqrt{\left\| I_t^H - I_t^{GT} \right\|^2 + \varepsilon^2} \tag{8}$$

TABLE III

AVERAGE PSNR AND SSIM OF 2× AND 4× SUPER-RESOLVED RESULTS ON THE VID4 DATASET. BEST RESULTS ARE SHOWN IN **BOLD**

Dataset	Method	Bicubic	DRCN [26]	LapSRN [27]	CARN [53]	VESPCN [11]	VSRnet [43]	VSRResNet [54]	**STAN** (Ours)
Vid4 ×2	PSNR ↑	28.42	31.57	31.41	31.96	–	31.29	31.87	**32.50**
	SSIM ↑	0.866	0.924	0.923	0.931	–	0.927	**0.943**	0.937
Vid4 ×4	PSNR ↑	23.75	24.94	24.98	25.27	25.35	24.81	25.51	**25.58**
	SSIM ↑	0.630	0.707	0.711	0.725	0.756	0.702	**0.753**	0.743

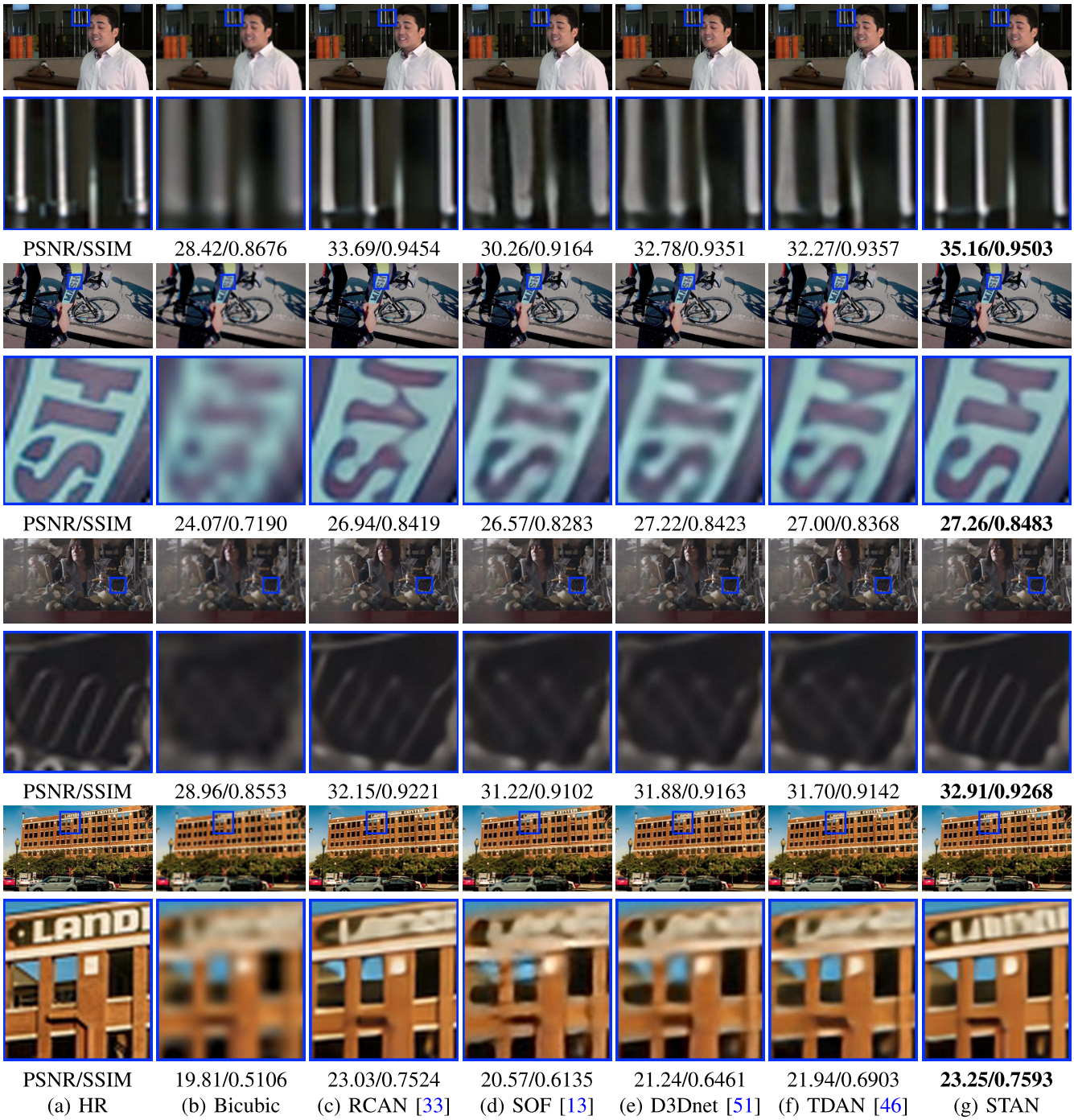


Fig. 6. Visual comparison for 4× SR on the Vimeo-90K-T dataset. The proposed STAN generates better visual quality and restores more image details compared with other SISR and VSR methods.

where ε is empirically set to 1×10^{-3}. Note that we do not use any other sophisticated loss functions such as adversarial loss [56], [57], perceptual loss [28], smoothness loss [58], and motion flow loss [13].

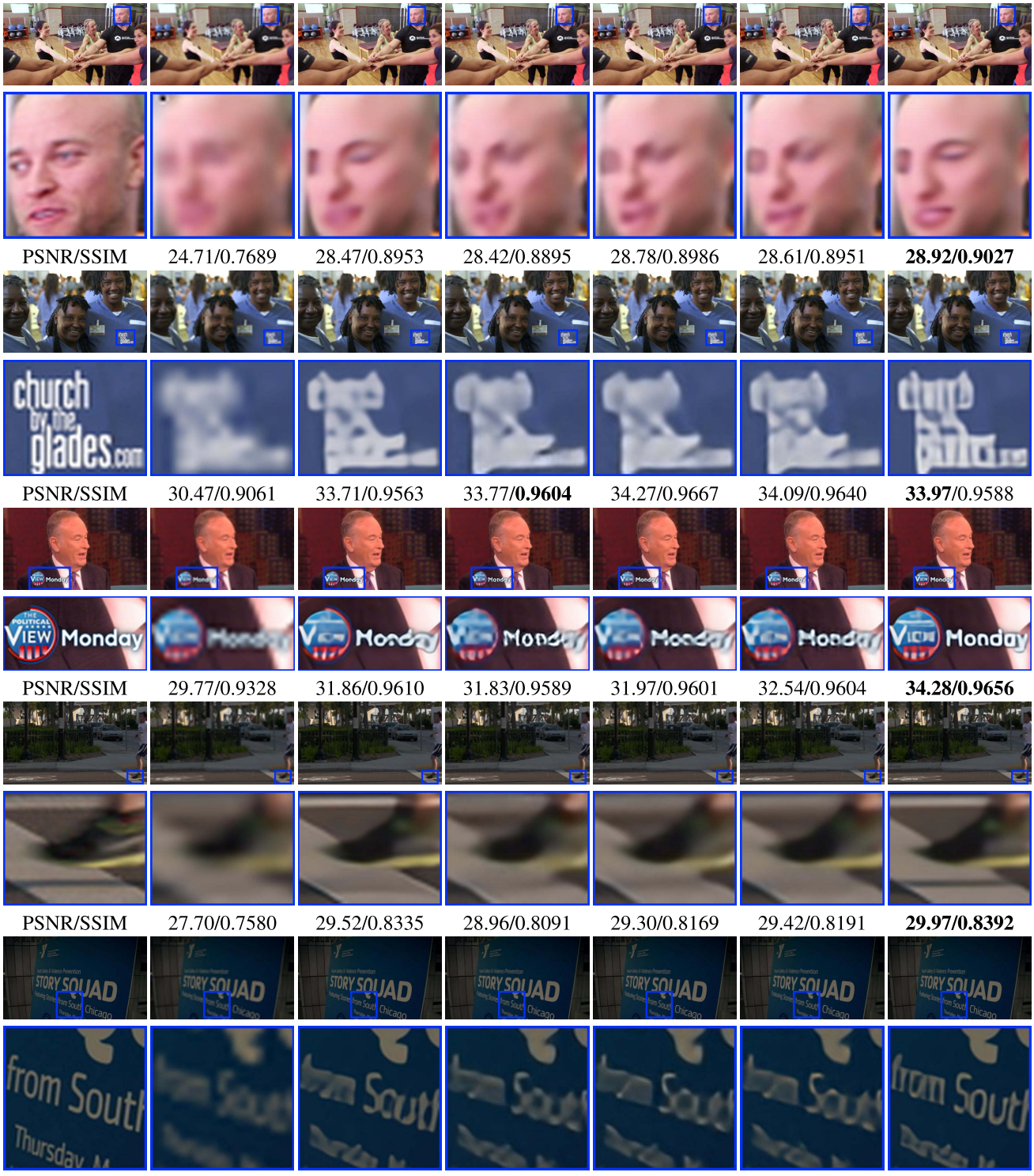


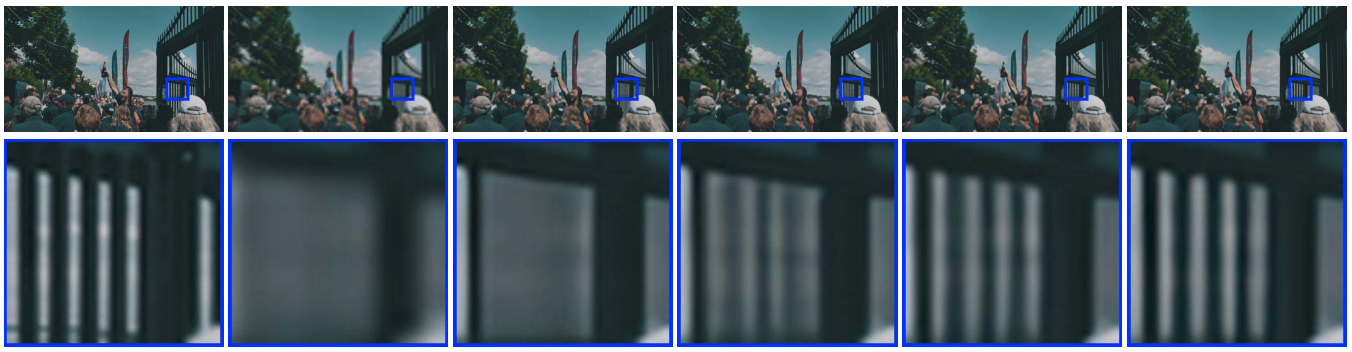
Fig. 7. Visual comparison for 4× SR on the Vimeo-90K-T dataset. The proposed STAN generates better visual quality and restores more image details compared with other SISR and VSR methods.

IV. Experiments

In this section, we first introduce the training datasets and implementation details. Then we compare the proposed network with several state-of-the-art methods on the Vid4 [47], REDS4 [59], and Vimeo-90K-T [10] datasets. Some video results can be found in the supplementary material.

A. Training Datasets

A large-scale dataset is important for training networks. Previous studies on VSR [13], [44] are usually trained or evaluated on private datasets. These datasets are all not publicly available. Recently, Xue *et al.* [10] released a Vimeo-90K super-resolution dataset with a fixed resolution of 448 × 256,



PSNR/SSIM | 25.84/0.7432 | 28.29/0.8525 | 28.24/0.8532 | 28.52/0.8582 | **28.94/0.8648**
(a) HR | (b) Bicubic | (c) RCAN [33] | (d) w/o FAAM | (e) w/ one FAAM | (f) STAN

Fig. 8. Ablation study for 4× SR with BI degradation model on the Vimeo-90K dataset. Our algorithm with three FAAM modules performs best than other configurations.

which contains 91701 training and testing samples (each with 7 consecutive frames). In the experiments, we use the training dataset from Vimeo-90K to train our network. We generate LR frames by bicubic with a down-sampling factor 4 and use consecutive 7 LR frames as input to predict the corresponding consecutive HR frames. For the testing data, we note that REDS [59] is a recently proposed video dataset with resolution of 1280×720, which consists of 240 training video clips and 30 validation video clips (each video clip contains 100 consecutive frames). Since the testing ground truth video clips is not available, recent work select four representative clips from the training and validation videos of REDS as the test set, denoted by REDS4. In this work, we also evaluate our method on the REDS4 dataset as [46]. In addition to REDS4, we also evaluate our method on the Vid4 [47] and Vimeo-90K testing dataset (denoted by Vimeo-90K-T) [10].

B. Implementation Details

We implement the proposed network using PyTorch platform. In our network, patch size is set as 128×128 . We use ADAM [60] optimizer with a batch size of 6 for training. The initial learning rate is set as 10^{-4}, after the first 60 epochs, the decay is carried out by 0.5, and then the decay is reduced every 30 epochs with each decay rate of 0.5. Default values of β_1 and β_2 are used, which are 0.9 and 0.999, respectively, and we set $\epsilon = 10^{-8}$. In addition to randomly rotating 90°, 180°, and 270°, and horizontal-flipping, we do not apply other data augmentation methods in the training process. The RAG block is selected as the basic unit for feature extraction, we set the number of RAG as $N = 10$. For all the results reported in the paper, we train the network for 200 epochs on an NVIDIA Titan RTX GPU with 24G RAM. The implementation code will be made available to the public.

C. Ablation Study

In this paper, we propose a dynamic convolution-based image feature alignment network. To verify effectiveness of the proposed feature adaptive alignment module, we conducted the following ablation experiments. We compare the SR results of the proposed network with one FAAM, three FAAM (i.e., STAN), and without the introduction of the alignment network, respectively. For the experiment without using the alignment module, we directly fuse the output feature maps of the last iteration with the current feature maps. As can be seen from Table IV, the ×4 SR result on the Vimeo-90K-T dataset is significantly improved with the introduction of the alignment network, and the performance tends to be better with more FAAM modules. In addition, we visualize the ablation experiments in Figure 8.

TABLE IV
EFFECTIVENESS OF THE PROPOSED FILTER ADAPTIVE ALIGNMENT MODULE FOR VIDEO SUPER-RESOLUTION ON THE VIMEO-90K-T DATASET

	RCAN	w/o FAAM	w/ one FAAM	**STAN**
PSNR	33.58	33.86	33.98	**34.08**
SSIM	0.911	0.914	0.915	**0.916**

Figure 8(d) shows that the super-resolved result without using the FAAM module still contains significant artifacts. In contrast, when we add one FAAM module, sharper details can be obtained as shown in Figure 8(e). In addition, we note that by adding the number (e.g., three) of FAAM modules, the proposed STAN can generate a sharper image as shown in Figure 8(f). These results prove the effectiveness of the proposed feature domain alignment module. Therefore, we use three FAAM in all the following experiments in this paper.

D. Comparisons With State-of-the-Art Methods

In this section, we compare our algorithm against previous SISR and VSR methods including DBPN [48], RCAN [32], DRCN [25], LapSRN [26], VESPCN [11], SPMC [12], CARN [52], VSRnet [42], 3DSRnet [62], VSRResNet [53], FRVSR [14], MEMC-Net [61], D3Dnet [50], TDAN [45], and SOF [13] on Vid4, REDS4 and Vimeo-90K-T datasets. For fair comparisons, we adopt the public implementations[1] of these methods and train all the approaches on the same Vimeo-90k training dataset.

[1] https://github.com/LoSealL/VideoSuperResolution

TABLE V

AVERAGE PSNR AND SSIM OF ENHANCED RESULTS ON THE VIMEO-90K-T DATASET. BEST RESULTS ARE SHOWN IN **BOLD**. '-' MEANS THAT IT IS NOT AVAILABLE

Method	Average (RGB)		Average (Y)	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Bicubic	29.73	0.8488	31.28	0.8683
RCAN [33]	33.58	0.9110	35.32	0.9256
MEMC-Net [64]	–	–	33.47	0.9470
TOFlow [10]	33.08	0.9054	34.83	0.9220
TGA [50]	33.48	0.9074	–	–
SOF [13]	32.88	0.9045	34.86	0.9234
D3Dnet [51]	33.39	0.9104	35.47	**0.9290**
TDAN [46]	33.56	0.9118	35.30	0.9265
STAN	**34.08**	**0.9162**	**35.75**	0.9288

The $\times 4$ quantitative reconstruction results on the REDS4 dataset are shown in Table II. Our algorithm achieves significant improvement in terms of PSNR and SSIM in almost all the video clips. Compared with the most recent methods of TDAN [45] and SOF [13], our algorithm achieves better results by up to 0.51 dB and 0.64 dB according to average RGB channel, respectively.

Since REDS4 only contains 4 videos, we evaluate all the methods on the Vimeo-90K-T dataset to prove the generalization ability of the proposed algorithm. The Vimeo-90K-T dataset contains 7814 video clips (we remove 10 all-black background video clips from the original 7824 testing samples). As shown in Table V, although the recent work of D3Dnet [50] yields the highest SSIM value, our method achieves better results by up to 0.48 dB than TDAN and 0.65 dB than D3Dnet on the Vimeo-90K-T dataset.

In addition, we also evaluate the proposed algorithm on the Vid4 test dataset according to $\times 2$ and $\times 4$ factors. The results are reported in Table III. As shown, our method performs favorably against other SISR and VSR methods for both $\times 2$ and $\times 4$ factors.

We then show the visual comparison of various methods on the REDS4 dataset for $\times 4$ SR in Figure 5. We note that most of the advanced SISR methods and the recent VSR networks could not accurately and clearly restore the grids and stripes of the building. Especially, most of the methods produce blurred artifacts for the "building" of the second image and the "floor" of the third image. TDAN and FRVSR can compensate for the details of the supporting frame, but there still are some fuzzy artifacts. In contrast, thanks to the feature domain alignment of the dynamic convolution, the proposed STAN can effectively suppress the artifacts caused by alignment and recovers finer details.

Figure 6 and Figure 7 are the $\times 4$ SR visualization results compared against three state-of-the-art VSR methods and a SISR approach on the Vimeo-90K-T dataset. As shown, RCAN [32] tends to generate some incorrect details (e.g., "SH" of the second image in Figure 6) in the restored results since RCAN does not use correlation information between adjacent frames. Although these three video SR methods of SOF, D3Dnet, and TDAN can avoid these erroneous information, their results still have some blurred artifacts as shown in Figure 6(d)-(f) and Figure 7(d)-(f).

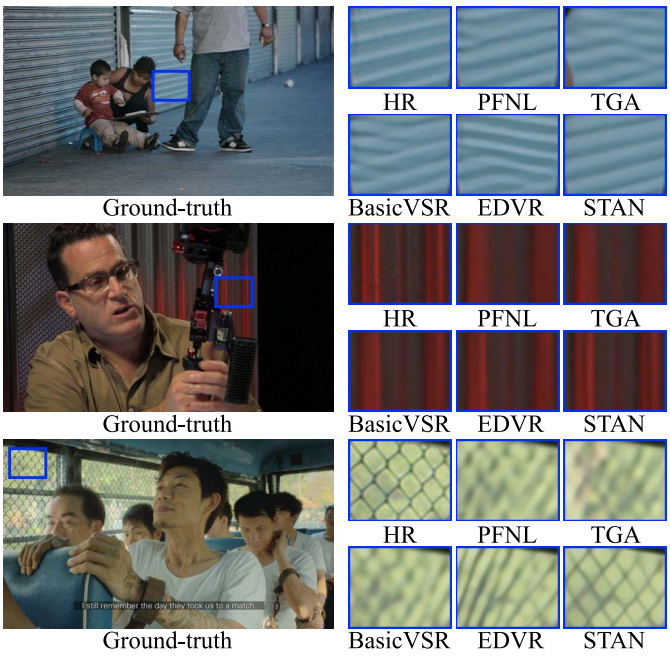


Fig. 9. Visual comparison for 4× SR on the Vimeo-90K-T dataset.

In contrast, the proposed algorithm recovers images with finer details and clearer structures. For example, our method clearly restores the shadow under the feet in the fourth image and the letters in the fifth image of Figure 7. Besides, We compared the proposed method with the recent algorithms BasicVSR [51], TGA [49], PNFL [63], and EDVR [46], as shown in Figure 9, these methods have some lines or mesh deformation. In contrast, our approach generates better reconstruction results of the dense lines and mesh areas. We analyze the reasons and attribute them to the following. First, for the methods based on deformable convolution such as EDVR, the pyramid structure will accumulate the erroneous offset estimation, resulting in texture deformation artifacts. For optical flow-based methods such as BasicVSR and BasicVSR++ [64], optical flow estimation mistakes will accumulate during the long-term bi-directional propagation, causing the deformation of lines and meshes. In contrast, our proposed alignment module based on adaptive filtering can adaptively align the features of adjacent frames to the current frame, which can better utilize the information of adjacent frames for reconstruction. Thus our method reconstruction results are more accurate.

E. Computational Efficiency and Temporal Consistency

In addition to the above comparison, we have executed a fair comparison of the number of parameters, computation, and inference time of the methods in the paper. We conducted inference time and calculations for all methods on a Tesla V100 with 32GB. All experiments are calculated with 180×320 LR video frames as input and HR reconstruction results of 720×1280 as output. As shown in Table I, although our method has many parameters, most of them are from the RCAB-based HR image reconstruction network. In addition,

the FLOPs of our approach are less compared to other implicit alignment-based and explicit alignment-based methods, such as DUF, EDVR, and BasicVSR. In comparing the inference time of various methods, we perform SR experiments on LR images with resolution of 180×320. The results show that the SOF algorithm has the fastest inference speed, taking about $108ms$ to infer a 180×320 image, followed by the SISR algorithm RCAN. Our approach is a bit slower than RCAN. The inference speed of our proposed algorithm is about twice faster than EDVR and BasicVSR.

The optical flow-based BasicVSR and pyramidal deformable convolution-based EDVR achieve the best video consistency results. Without using complex optical flow estimation and pyramid structure, our method achieves reliable consistency results by filter adaptive alignment module. For the excellent performance of BasicVSR and EDVR, we analyze that the optical flow estimation method can reasonably estimate the motion between video frames and maintain the consistency between video frames well by motion compensation. However, the optical flow estimation error will accumulate with long-term propagation and affect the image detail reconstruction. Thus, although the reconstructed video consistency of the BasicVSR method is excellent, the local detail texture recovery results are not well.

We attribute the perfect video consistency of EDVR to using the pyramid structure of the deformable convolution, which can align features from coarse to fine. The coarse layers mainly learn the image contour information, which is beneficial in maintaining the consistency between video frames. However, the continuous up-sampling of the coarse layers makes the error accumulate and cause artifacts in local details. Besides, the offset learning error is also superimposed in the pyramid structure. As the visualization results shown in Figure 9, the high-resolution image recovered by the EDVR method suffers from local texture artifacts. In contrast, our proposed method does relatively well in terms of texture detail processing.

V. Conclusion

In this paper, we propose a one-stage framework for VSR to reconstruct high-resolution videos without explicit motion compensation. To achieve this, we introduce a spatio-temporal alignment network (STAN) for feature domain temporal alignment, which effectively fuses the multi-frame information and improves the temporal consistency of the video. Furthermore, we propose a novel end-to-end VSR network which consists of an alignment network STAN and a reconstruction network for handling alignment and aggregating temporal information. With such a one-stage design, our network can explore inter-frame spatio-temporal information and adaptively learn to align adjacent video frames. Extensive experiments demonstrate that the proposed STAN is more efficient than existing two-stage networks and performs favorably against the state-of-the-art SISR and VSR methods.

References

[1] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[2] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3146–3154.

[3] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3147–3155.

[4] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 136–144.

[5] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.

[6] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11065–11074.

[7] L. Chen, J. Pan, and Q. Li, "Robust face image super-resolution via joint learning of subdivided contextual model," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5897–5909, Dec. 2019.

[8] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8174–8182.

[9] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[10] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, 2019.

[11] J. Caballero *et al.*, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4778–4787.

[12] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4472–4480.

[13] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.

[14] M. S. M. Sajjadi, R. Vemulapalli, and M. Brown, "Frame-recurrent video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6626–6634.

[15] J. Allebach and P. W. Wong, "Edge-directed interpolation," in *Proc. 3rd IEEE Int. Conf. Image Process.*, vol. 3, Sep. 1996, pp. 707–710.

[16] X. Li and M. T. Orchard, "New edge-directed interpolation," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1521–1527, Oct. 2001.

[17] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.

[18] J. Sun, Z. Xu, and H.-Y. Shum, "Image super-resolution using gradient profile prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[19] Y.-W. Tai, S. Liu, M. S. Brown, and S. Lin, "Super resolution using edge prior and single image detail synthesis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2400–2407.

[20] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. ECCV*, 2014, pp. 184–199.

[21] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.

[22] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4799–4807.

[23] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[24] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 7, pp. 2480–2495, Jul. 2020.

[25] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1637–1645.

[26] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 624–632.

[27] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2018.

[28] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[29] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014.
[30] M. S. M. Sajjadi, B. Schölkopf, and M. Hirsch, "EnhanceNet: Single image super-resolution through automated texture synthesis," 2016, *arXiv:1612.07919*.
[31] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–18.
[32] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
[33] B. Niu *et al.*, "Single image super-resolution via a holistic attention network," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 191–207.
[34] M. Bevilacqua, A. Roumy, C. Guillemot, and M.-L.-A. Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *Proc. Brit. Mach. Vis. Conf.*, 2012, pp. 1–10.
[35] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Curves Surf.* Cham, Switzerland: Springer, 2010, pp. 711–730.
[36] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Jun. 2001, pp. 416–423.
[37] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5197–5206.
[38] Y. Matsui *et al.*, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21811–21838, Oct. 2017.
[39] E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 114–125.
[40] S. Li, F. He, B. Du, L. Zhang, Y. Xu, and D. Tao, "Fast spatio-temporal residual network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10522–10531.
[41] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia, "Video super-resolution via deep draft-ensemble learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 531–539.
[42] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imaging*, vol. 2, no. 2, pp. 109–122, Jun. 2016.
[43] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 1015–1028, Apr. 2018.
[44] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3224–3232.
[45] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "TDAN: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3360–3369.
[46] X. Wang, K. C. K. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[47] C. Liu and D. Sun, "A Bayesian approach to adaptive video super resolution," in *Proc. CVPR*, Jun. 2011, pp. 209–216.
[48] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1664–1673.
[49] T. Isobe *et al.*, "Video super-resolution with temporal group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8008–8017.
[50] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.
[51] K. C. K. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "BasicVSR: The search for essential components in video super-resolution and beyond," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4947–4956.
[52] N. Ahn, B. Kang, and K.-A. Sohn, "Fast, accurate, and lightweight super-resolution with cascading residual network," 2018, *arXiv:1803.08664*.
[53] A. Lucas, S. López-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019, doi: 10.1109/TIP.2019.2895768.
[54] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[55] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Sep. 2016, pp. 1874–1883.
[56] X. Wang *et al.*, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. Workshops (ECCVW)*, Sep. 2018, pp. 1–16.
[57] Z. Zhang, Z. Wang, Z. Lin, and H. Qi, "Image super-resolution by neural texture transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7982–7991.
[58] L. Wang *et al.*, "Learning parallax attention for stereo image super-resolution," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12250–12259.
[59] S. Nah *et al.*, "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[61] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 933–948, Mar. 2021.
[62] Y. Xie, J. Xiao, T. Tillo, Y. Wei, and Y. Zhao, "3D video super-resolution using fully convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.
[63] P. Yi, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3106–3115.
[64] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment," 2021, *arXiv:2104.13371*.

Weilei Wen received the master's degree from Xidian University in 2021. He is currently pursuing the Ph.D. degree with Nankai University, Tianjin, China. His research interests include computer vision and deep learning. He mainly works on the area of generative models and image enhancement.

Wenqi Ren (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017. From 2015 to 2016, he was supported by the China Scholarship Council and working with Prof. Ming-Husan Yang as a Joint-Training Ph.D. Student with the Electrical Engineering and Computer Science Department, University of California at Merced. He is currently an Assistant Professor with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. His research interests include image processing and related high-level vision problems. He received the Tencent Rhino Bird Elite Graduate Program Scholarship in 2017 and the MSRA Star Track Program in 2018.

Yinghuan Shi received the B.Sc. and Ph.D. degrees from the Department of Computer Science, Nanjing University, China, in 2007 and 2013, respectively. He was a Visiting Scholar with the University of North Carolina at Chapel Hill and the University of Technology Sydney. He is currently an Associate Professor with the Department of Computer Science and Technology, Nanjing University. He has published more than 40 research papers in related journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, AAAI, ACMMM, MICCAI, and IPMI. His research interests include computer vision and medical image analysis. He serves as a program committee member for several conferences and a referee for several journals.

Yunfeng Nie received the Ph.D. degree in optical engineering from Vrije Universiteit Brussel, under the EU's FP7 Marie Curie Program "ADOPSYS" with exchanges at LPI Company, Madrid, Spain, and the University of Jena, Germany. She has been a full-time Researcher with the Faculty of Engineering, VUB, since 2014. She has been very active in freeform optical design algorithms, biomedical photonics, imaging spectrometers, and computational imaging.

Jingang Zhang is currently an Associate Professor with the University of Chinese Academy of Sciences (UCAS). He has presided over more than ten national and ministerial-level scientific research projects, such as the National Natural Science Foundation of China and the Joint Foundation Program of the Chinese Academy of Sciences for equipment pre-feasibility study. His research interests include image denosing, deblurring, and dehazing; image/video analysis and enhancement; and related high-level vision problems.

Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and coauthored more than 120 journal and conference papers. He is a fellow of IET. He is on the Editorial Board of the IEEE TRANSACTIONS ON IMAGE PROCESSING. His dissertation was nominated for the University of Central Florida's University-Level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.