TDAN: Temporally-Deformable Alignment Network for Video Super-Resolution

Yapeng Tian[1], Yulun Zhang[2], Yun Fu[2], Chenliang Xu[1]
[1]University of Rochester, [2]Northeastern University
{yapengtian,chenliang.xu}@rochester.edu, yulun100@gmail.com, yunfu@ece.neu.edu

Abstract

Video super-resolution (VSR) aims to restore a photo-realistic high-resolution (HR) video frame from both its corresponding low-resolution (LR) frame (reference frame) and multiple neighboring frames (supporting frames). Due to varying motion of cameras or objects, the reference frame and each support frame are not aligned. Therefore, temporal alignment is a challenging yet important problem for VSR. Previous VSR methods usually utilize optical flow between the reference frame and each supporting frame to warp the supporting frame for temporal alignment. However, both inaccurate flow and the image-level warping strategy will lead to artifacts in the warped supporting frames. To overcome the limitation, we propose a temporally-deformable alignment network (TDAN) to adaptively align the reference frame and each supporting frame at the feature level without computing optical flow. The TDAN uses features from both the reference frame and each supporting frame to dynamically predict offsets of sampling convolution kernels. By using the corresponding kernels, TDAN transforms supporting frames to align with the reference frame. To predict the HR video frame, a reconstruction network taking aligned frames and the reference frame is utilized. Experimental results demonstrate that the TDAN is capable of alleviating occlusions and artifacts for temporal alignment and the TDAN-based VSR model outperforms several recent state-of-the-art VSR networks with a comparable or even much smaller model size. The source code and pre-trained models are released in https://github.com/YapengTian/TDAN-VSR.

1. Introduction

The goal of video super-resolution (VSR) is to reconstruct a high-resolution (HR) video frame from its corresponding low-resolution (LR) video frame (reference frame) and multiple neighboring LR video frames (supporting frames). HR video frames contain more image details and are more preferred to humans. Therefore, the VSR technique is desired in many real applications, such as video surveillance and high-definition television (HDTV).

Figure 1. VSR results for a frame in the *walk* sequence. We find that our method can restore more accurate image structures and details than the recent DUF network.

To super-resolve the LR reference frame, VSR will exploit both the LR reference frame and multiple LR supporting frames. However, the LR reference frame and each supporting frame may not be aligned due to the motion of camera or objects. Thus, a vital issue in VSR is how to align the supporting frames with the reference frame.

Previous methods [2, 3, 4, 5, 6] usually exploit optical flow to predict motion fields between the reference frame and supporting frames, then warp the supporting frames using their corresponding motion fields. Therefore, the optical flow prediction is crucial for these approaches. Any errors in the flow computation or the image-level warping operation may introduce artifacts around image structures in the aligned supporting frames.

To alleviate the above issues, we propose a temporally-deformable alignment network (TDAN) in this paper that performs one-stage temporal alignment without using optical flow. Unlike previous optical flow-based VSR methods, our approach can adaptively align the reference frame and supporting frames at the feature level without explicit motion estimation and image warping operations. Therefore, the aligned LR video frames will have less annoying image artifacts, and the image quality of the reconstructed HR video frame will be improved. In specific, inspired by the deformable convolution [7], our TDAN uses features from both the reference frame and each supporting frame to dy-

namically predict offsets of sampling convolution kernels. These dynamic kernels are then applied on features from supporting frames to employ the temporal alignment. Here, given different reference and supporting frame pairs, the module will generate their corresponding sampling kernels, which makes TDAN have strong capability and flexibility to handle various motion conditions in temporal scenes. With the aligned supporting frames and the reference frame, a reconstruction network is utilized to predict an HR video frame corresponding to the LR reference frame.

Experimental results on the widely-used VSR benchmarks: Vid4 [8] and SPMCs-30 [4] show that our framework achieves promising performance with beyond 0.5dB improvements in terms of PSNR over recent ToFlow [5], FRVSR [6], and FSTRN [9] on Vid4 and beyond 0.6dB improvements over recent DUF [1] on SPMCs-30. In Fig. 1, we show a visual comparison to DUF, and we can see that our method reconstructs more image details.

The contributions of this paper are three-fold: (1) we propose a novel temporally-deformable alignment network (TDAN) for feature-level alignment, which avoids the two-stage process adopted by previous optical flow-based methods and is capable of explore image contexts for alleviating occlusions; (2) we propose an end-to-end trainable VSR framework based on the TDAN; and (3) our method achieves better performance than several recent state-of-the-art VSR performance on Vid4 and SPMCs-30 benchmark datasets. An early version of our work was firstly released to ArXiv [10] in 2018. After that, it has already made a good impact on our community and been followed and improved by recent works including EDVR [11] for video super-resolution and deblurring and Zooming SlowMo [12] for space-time video super-resolution.

2. Related Work

Single Image Super-Resolution (SISR): Respecting for the long-history research on SISR, we only survey deep learning-based methods in this section. Dong *et al*. [13] firstly proposed an end-to-end image super-resolution convolutional neural network (SRCNN). Kim *et al*. [14] introduced a 20-layer deep network: VDSR with residual learning. Shi *et al*. [15] learned an efficient sub-pixel convolution layer to upscale the final LR feature maps into the HR output for accelerating SR networks. Deeper networks like LapSRN [16], DRRN [17], and MemNet [18], were explored to further improve SISR performance. However, training images used in the previous methods have limited resolution, which makes the training of even deeper and wider networks very difficult. Recently, Timofte *et al*. introduced a novel large dataset (DIV2K) consisting of 1000 DIVerse 2K resolution RGB images in NTIRE 2017 Challenge [19]. Current state-of-the-art SISR networks [20, 21, 22, 23, 24, 25, 26] trained on the DIV2K and outperformed previous networks by a substantial margin. A recent survey is conducted in [27].

Video Super-Resolution (VSR): It has been observed that temporal alignment critically affects the performance of VSR systems. Previous methods usually adopted two-stage approaches based on optical flow. They conducted motion estimation by computing optical flow in the first stage and utilize the estimated motion fields to perform image warping/motion compensation in the second stage. For example, Liao *et al*. [28] used two classical optical flow methods, TV-L_1 and MDP flow [29], with different parameters to generate HR SR-drafts, and then predicted the final HR frame by a deep draft-ensemble network. Kappeler *et al*. [30] took interpolated flow-warpped frames as inputs of a CNN to predict HR video frames. However, both the pioneering methods used classical optical flow algorithms, which are separated from the frame reconstruction CNN and are much slower than the flow CNN during inference.

To address the issue, Caballero *et al*. [2] introduced the first end-to-end VSR network: VESCPN, which jointly trains flow estimation and spatio-temporal networks. Liu *et al*. [3] proposed a temporal adaptive neural network to adaptively select the optimal range of temporal dependency and a rectified optical flow alignment method for better motion estimation. Tao *et al*. [4] computed LR motion field based on optical flow network and designed a new layer to utilize sub-pixel information from motion and simultaneously achieve sub-pixel motion compensation (SPMC) and resolution enhancement. Xue *et al*. [5] exploited task-oriented motion cues via Task-Oriented Flow (TOFlow), which achieves better VSR results than fixed flow algorithms. Sajjadi *et al*. [6] extended the conventional VSR models to a frame-recurrent VSR framework. Kim *et al*. [31] introduced a spatio-temporal flow estimation network to capture long-range temporal dependencies. However, sufficient high-quality motion estimation is not easy to obtain even with state-of-the-art optical flow estimation networks. Even with accurate motion fields, the image-warping based motion compensation will produce artifacts around image structures, which may be propagated into final reconstructed HR frames. The proposed TDAN performs a feature-wise one-stage temporal alignment without relying on optical flow, which will alleviate the issues in these previous optical flow-based VSR networks.

Recently, Jo *et al*. [1] proposed to use dynamic upsclaing filter and Li *et al*. [9] utilized a 3D convolution-based residual network for VSR. However, without explicitly temporal alignment, they have limited capacity in handling various and diverse spatio-temporal visual patterns.

Deformable Convolution: CNNs have the inherent limitation in modeling geometric transformations due to the fixed kernel configuration. To enhance the transformation modeling capability of CNNs, Dai *et al*. [7] proposed a de-

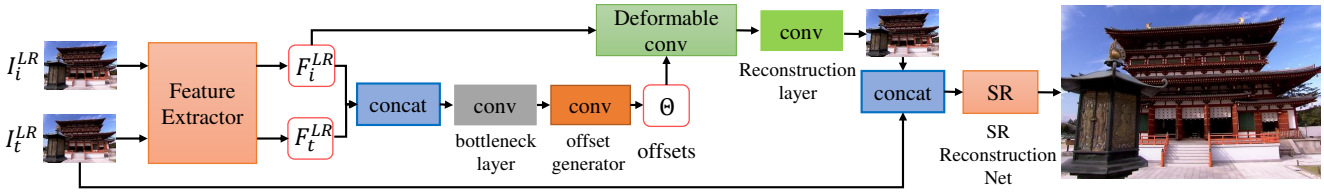


Figure 2. The proposed TDAN-based VSR framework. Here, we only show the framework with one supporting frame. In our implementation, 4 neighboring supporting frames are used for exploring more temporal information.

formable convolution operation. It has been applied to address several high-level vision tasks, such as object detection [7, 32], semantic segmentation [7], and human pose estimation [33]. Although the deformable convolution has shown superiority on these high-level vision tasks, it is rarely explored in low-level vision problems.

3. Method

3.1. Overview

Let $I_t^{LR} \in \mathbb{R}^{H\times W\times C}$ be the t-th LR video frame, and $I_t^{HR} \in \mathbb{R}^{sH\times sW\times C}$ be the corresponding HR video frame, where s is the upscaling factor, $H \times W$ denotes the frame size, and C refers to channel number. Our goal is to restore the HR video frame I_t^{HR} from the reference LR frame I_t^{LR} and $2N$ supporting LR frames $\{I_{t-N}^{LR}, ...I_{t-1}^{LR}, I_{t+1}^{LR}, ..., I_{t+N}^{LR}\}$. Therefore, our VSR framework takes the consecutive $2N+1$ frames $\{I_i^{LR}\}_{i=t-N}^{t+N}$ as the input to predict the HR frame I_t^{HR}, which is illustrated in Fig. 2. It consists of two main sub-networks: a temporally-deformable alignment network (TDAN) to align each supporting frame with the reference frame and a super-resolution (SR) reconstruction network to predict the HR frame.

The TDAN takes a LR supporting frame I_i^{LR} and the LR reference frame I_t^{LR} as inputs to predict the corresponding aligned LR frame $I_i^{LR'}$ of the supporting frame:

$$I_i^{LR'} = f_{TDAN}(I_t^{LR}, I_i^{LR}) \ . \tag{1}$$

Feeding the $2N$ supporting frames into the TDAN separately, we can obtain $2N$ correspondingly-aligned LR frames $\{I_{t-N}^{LR'}, ...I_{t-1}^{LR'}, I_{t+1}^{LR'}, ..., I_{t+N}^{LR'}\}$.

The $2N$ aligned frames along with the reference frame are then fed into the super-resolution (SR) reconstruction network. We can finally reconstruct the HR video frame:

$$I_t^{HR} = f_{SR}(I_{t-N}^{LR'}, ..., I_{t-1}^{LR'}, I_t^{LR}, I_{t+1}^{LR'}, ..., I_{t+N}^{LR'}) \ . \tag{2}$$

3.2. Temporally-Deformable Alignment Network

Given an LR supporting frame I_i^{LR} and the LR reference frame I_t^{LR}, the proposed TDAN will temporally align I_i^{LR} with I_t^{LR}. It mainly consists of three modules: feature extraction, deformable alignment, and aligned frame reconstruction.

Feature Extraction: This module extracts visual features F_i^{LR} and F_t^{LR} from I_i^{LR} and I_t^{LR}, respectively, via a shared feature extraction network. The network consists of one convolutional layer and k_1 residual blocks [34] with ReLUs as the activation functions. In our implementation, we adopted a modified residual structure from [20]. The extracted features are then used for feature-wise temporal alignment.

Deformable Alignment: The deformable alignment module takes the F_i^{LR} and F_t^{LR} as inputs to predict sampling parameters Θ for the feature F_i^{LR}:

$$\Theta = f_\theta(F_i^{LR}, F_t^{LR}) \ . \tag{3}$$

Here, $\Theta = \{\triangle p_n \mid n = 1, \ldots, |\mathcal{R}|\}$ refers to the offsets of the convolution kernels, where $\mathcal{R} = \{(-1,-1),(-1,0),...,(0,1),(1,1)\}$ donates a regular grid of a 3×3 kernel. With Θ and F_i^{LR}, the aligned feature $F_i^{LR'}$ of the supporting frame can be computed by the deformable convolution:

$$F_i^{LR'} = f_{dc}(F_i^{LR}, \Theta) \ . \tag{4}$$

More specifically, for each position p_0 on the aligned feature map $F_i^{LR'}$, we have:

$$F_i^{LR'}(p_0) = \sum_{p_n\in\mathcal{R}} w(p_n) F_i^{LR}(p_0 + p_n + \triangle p_n) \ . \tag{5}$$

The convolution will be operated on the irregular positions $p_n + \triangle p_n$, where the $\triangle p_n$ may be fractional. To address the issue, the operation is implemented by using bilinear interpolation, which is the same as that proposed in [7].

Here, the deformable alignment module consists of several regular and deformable convolutional layers. For the sampling parameter generation function f_θ, it concatenates F_i^{LR} and F_t^{LR}, and uses a 3×3 bottleneck layer to reduce the channel number of the concatenated feature map. Then, the sampling parameters are predicted by a convolutional layer with the kernel size $|\mathcal{R}|$ as the output channel number. Finally, the aligned feature $F_i^{LR'}$ is obtained from Θ and F_i^{LR} based on deformable convolution operation. In practice, besides the deformable convolution for alignment, we use 2 additional regular deformable convolutional layers before and 1 additional regular deformable convolutional layer after the f_{dc} for enhancing the transformation flexibility and capability of the module. Section 4.3 contains the ablation

study on the performance of the module with different numbers of the additional deformable convolutional layers.

We note that the feature of the reference frame F_t^{LR} is only used for computing the offset, and its information will not be propagated into the aligned feature of the supporting frame $F_i^{LR'}$. Besides, the adaptively-learned offset will implicitly capture motion cues and explore neighboring features within the same image structures for alignment.

Aligned Frame Reconstruction: Although the deformable alignment has the potential to capture motion cues and align F_i^{LR} with F_t^{LR}, the implicit alignment is difficult to learn without a supervision. Therefore, we restore an aligned LR frame $I_i^{LR'}$ for I_i^{LR} and utilize an alignment loss to enforce the deformable alignment module to sample useful features for accurate temporal alignment. The aligned LR frame $I_i^{LR'} \in \mathbb{R}^{H\times W\times C}$ can be reconstructed from the aligned feature map with a 3×3 convolutional layer.

After feeding $2N$ reference and supporting frame pairs into TDAN consecutively, we can obtain the corresponding $2N$ aligned LR frames, which will be used to predict the HR video frame I_t^{HR} in the SR reconstruction network.

3.3. SR Reconstruction Network

We use a SR reconstruction network to restore the HR video frame I_i^{HR} from the aligned LR frames and the reference frame. The network contains three modules: temporal fusion, nonlinear mapping, and HR frame reconstruction, which will aggregate temporal information from different frames, predict high-level visual features, and restore the HR frame for the LR reference frame, respectively.

Temporal Fusion: To fuse different frames cross the space-time, we directly concatenate the $2N + 1$ frames and then feed them into a 3×3 convolutional layer to output the fused feature map.

Nonlinear Mapping: The nonlinear mapping module with k_2 stacked residual blocks [20] will take the shadow fused features as input to predict deep features.

HR Frame Reconstruction: After extracting deep features in the LR space, inspired by the EDSR [20], we utilize an upscaling layer to increase the resolution of the feature map with a sub-pixel convolution as proposed by Shi *et al.* [15]. In practice, for $\times4$ upscaling, two sub-pixel convolution modules will be used. The final HR video frame estimation $I_t^{HR'}$ will be obtained by a convolutional layer from the zoomed feature map.

3.4. Loss Functions

Two loss functions $\mathcal{L}_{align}$ and $\mathcal{L}_{sr}$ are used to train the TDAN and SR reconstruction networks, respectively. Note that we have no ground-truth of the aligned LR frames. To optimize the TDAN, we utilize the reference frame as the label and make the aligned LR frames close to the reference frame:

$$\mathcal{L}_{align} = \frac{1}{2N} \sum_{i=t-N,\neq t}^{t+N} \| I_i^{LR'} - I_t^{LR} \|_2^2 \quad . \tag{6}$$

The objective function of the SR reconstruction network is defined via $\mathcal{L}_1$ reconstruction loss:

$$\mathcal{L}_{sr} = \| I_t^{HR'} - I_t^{HR} \|_2^2 \quad . \tag{7}$$

Combining the two loss terms, we have the overall loss function for training our VSR framework:

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{sr} \quad . \tag{8}$$

The two loss terms are simultaneously optimized when training our VSR framework. Therefore, our TDAN-based VSR network is end-to-end trainable. In addition, the TDAN can be trained with a self-supervision without requiring any annotations.

3.5. Analyses of the Proposed TDAN

Given a reference frame and a set of supporting frames, the proposed TDAN can employ temporal alignment to align the supporting frames with the reference frame. It has several merits:

One-Stage Temporal Alignment: Most previous temporal alignment methods are based on optical flow, which will split the temporal alignment problem into two sub-problems: flow/motion estimation and motion compensation. As discussed in the paper, the performance of these methods highly rely on the accuracy of flow estimation, and the flow-based image warping will introduce annoying artifacts. Unlike these two-stage temporal alignments, our TDAN is a one-stage approach, which aligns the supporting frames at the feature level. It implicitly captures motion cues via adaptive sampling parameter generation without explicitly estimating the motion field, and reconstructs the aligned frames from the aligned features.

Self-Supervised Training: The optical flow estimation is crucial for the two-stage methods. For ensuring the accuracy of flow estimation, some VSR networks [4, 5, 3] utilized additional flow estimation algorithms. Unlike these methods, there is no flow estimation inside the TDAN, and it can be trained in a self-supervised manner without relying on any extra supervisions.

Exploration: For each location in a frame, its motion field computed by optical flow only refers to one potential position p. It means that each pixel in a warpped frame will only copy one pixel at p or use an interpolated value for a fractional position. However, beyond utilizing information at p, our deformable alignment module can adaptively explore more features at sampled positions, which may share the same image structure as p, and it will help to aggregate more contexts for better-aligned frame reconstruction.

Methods	Bicubic	VSRnet [30]	VESCPN [2]	Liu *et al*. [3]	DBPN [22]	RDN [21]	RCAN [23]	TOFlow [5]	TDAN
Vid4	23.79/0.633	24.73/0.697	25.34/0.730	25.53/0.749	25.33/0.731	25.40/0.735	25.42/0.737	25.90/0.765	**26.42/0.789**
SPMCs-30	27.08/0.744	-/-	-/-	-/-	29.76/0.830	29.92/0.836	30.07/0.841	29.47/0.831	**30.38/0.854**

Table 1. PSNR (dB) and SSIM of different networks on Vid4 and SPMCs-30 with upscale factor 4 under BI configuration. The top-2 results are highlighted with red and blue colors.

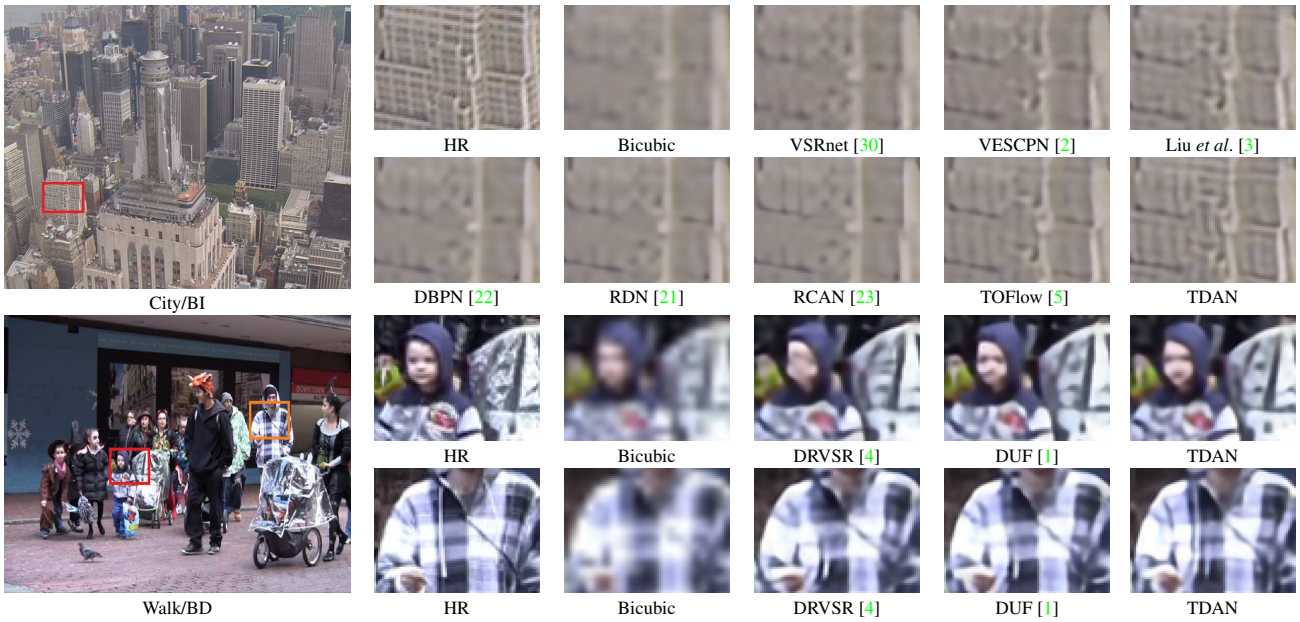


Figure 3. Visual comparisons for $4\times$ VSR on the Vid4 dataset. We observed that the proposed TDAN restores better image structures and details than other state-of-the-art VSR networks, which demonstrates the strong capability of the TDAN in temporal alignment leveraging informative pixels from LR supporting frames.

Therefore, the proposed TDAN has stronger exploration capability than optical flow-based models.

4. Experiments

4.1. Experimental Settings

Datasets: In our experiments, we used Vimeo Super-Resolution dataset *et al*. [5] containing 64612 training samples with 448×256 resolution as our training dataset and 31 frames from the Temple sequence [28] as the validation dataset. Same as other methods, we evaluated our models on the Vid4 benchmark [8], which contains four video sequences: *city*, *walk*, *calendar*, and *foliage*, and each sequence in the Vid4 has at least 30 720×480 video frames. In addition, we compare different methods on a larger testing set: SPMCs-30 [4]. It has 30 diverse and dynamic scenes and each sequence has 31 960×520 HR frames. Since either reconstructed frames or source code are not available for some methods, we will not report their results on the SPMCs-30.

Evaluation Metrics: PSNR, SSIM [35], and VQM_VFD [36] are used as evaluation metrics to compare different VSR networks quantitatively. We used PSNR between a reference frame and the corresponding aligned supporting frame as a metric to evaluate temporal alignment performance[1]. Following the evaluation from previous approaches [1, 6, 2], we crop 8 pixels near image boundary and ignore the first 3 and last 3 frames.

Degradation Methods: We compared our TDAN-based network with current state-of-the-art VSR and SISR networks: VSRnet [30], ESPCN [15], VESCPN [2], Liu *et al*. [3], TOFlow [5], DBPN [22], RDN [21], RCAN [23], DRVSR [4], FSRVSR [6], FSTRN [9] ,and DUF [1]. The first eight networks adopted the Matlab function *imresize* with the option bicubic (BI) to generate LR video frames. SPMC, FSRVSR, and DUF obtained LR frames by first blurring HR frames via a Gaussian kernel and then downsampling the blurred frames (denote as BD for short). Note that we compared the FRVSR-3-64 and DUF-16L models, which have similar model sizes as our TDAN-based VSR network. Recently, RBPN [37] has shown promising VSR results. However, its model size is more than 6 times larger than our TDAN, thus we do not include it into comparison. We trained two different TDAN models with the two different degradation methods for fair comparisons.

Implementation Details: In our implementation, $k_1 = 5$

[1]There is no ground-truth aligned frames, so we use the reference frame as a pseudo label to approximately measure the temporal alignment quality.

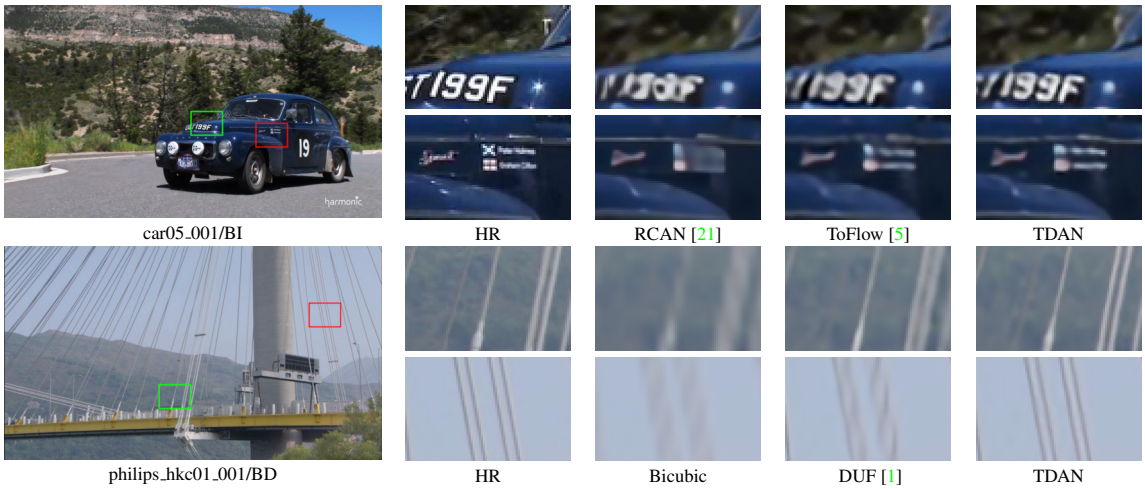


Figure 4. Visual comparisons for $4\times$ VSR on video frames from the SPMCs-30.

Methods	Bicubic	DRVSR [4]	FSRVSR [6]	FSTRN [9]	DUF [1]	TDAN
Vid4	23.47/0.616	26.03/0.775	26.17/0.798	24.76/0.720	26.85/0.816	26.86/0.814
SPMCs-30	26.68/0.730	29.89/0.840	-/-	-/-	30.14/0.857	30.80/0.869

Table 2. PSNR (dB) and SSIM of different networks with upscale factor 4 under the BD configuration. Our TDAN achieves significant improvements over the other methods on the SPMCs-30, which contains diverse and dynamic scenes.

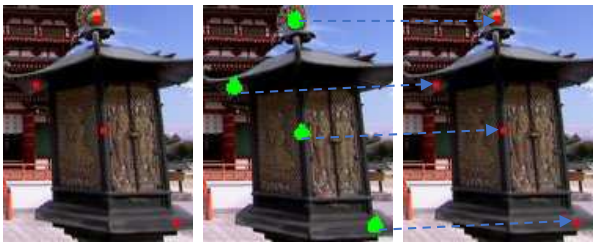


Figure 5. Visualization of the learned sampling positions. The proposed TDAN aligns supporting frames with the corresponding reference frame, and the aligned frame is reconstructed from features from the supporting frame based on learned sampling positions from both the reference and supporting frames. Green points in the supporting frame indicate sampling positions for predicting corresponding pixels labeled with red color in the aligned frame by TDAN. We used 3 layers with 3×3 kernels to sample features from feature maps of supporting frames and reconstruct aligned frames. So, we show 9^3 sampling points with green color for each output pixel (red point). Note that we directly show sampling on the supporting frame rather than feature maps, and center points of 5x5 red boxes refer to target pixels for better visualization.

and $k_2 = 10$ residual blocks are used in feature extraction and SR reconstruction networks, respectively. We used a downsampling factor: $s = 4$ in our experiments. Each training batch contains 64×5 LR RGB patches with the size 48×48, where 64 means the batch size and 5 refers to the number of consecutive input frames. We implemented our network with PyTorch [38] and adopt Adam [39] as the optimizer.

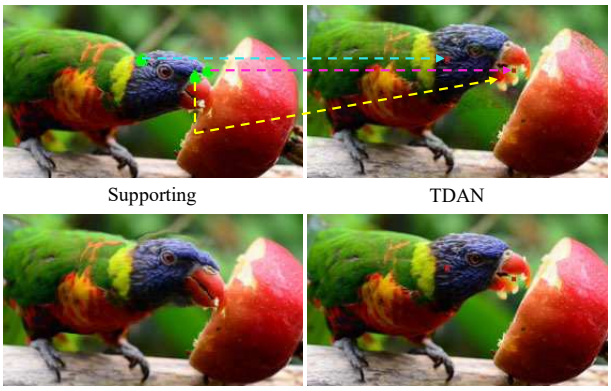


Figure 6. Temporal alignment results of FlowVSR and TDAN on a supporting/reference frame pair with a very large temporal gap (60 frames). The TDAN can exploit rich image contexts containing similar content (green regions) as target pixels (red points) from the supporting frame to employ accurately temporal alignment. Note that the alignment is performed on feature maps of supporting frames, so the real exploited image regions for alignment will be much larger than the green regions.

4.2. Comparisons

Results with BI Degradation: Table 1 shows quantitative comparisons on the BI configuration. Note that our TDAN used the same training dataset as TOFlow, and other VSR methods did not release their training data or training source code. Therefore we compared with their methods directly based on the provided results. We can see that our TDAN achieves the best performance among all compared state-of-the-art flow-based VSR networks and SISR networks.

Visual results for $4\times$ VSR on the BI configuration are illustrated in Fig. 3 and Fig. 4. We can find that the SISR networks without using the supporting frames: DPBN, RDN, and RCAN, fail to restore missing image details, such as the building structures in City image and numbers in Car image. With motion compensation, VESCPN [2], Liu *et al*. [3], and TOFlow [5] can compensate missing details from the supporting frames. Our TDAN recovers more fine image details than others, which demonstrates the effectiveness of the proposed framework.

Results with BD Degradation: Table 3 shows quantitative comparisons on the BD configuration. Our method outperforms the flow-based networks (*e.g*. DRVSR and FRVSR) and 3D convolution-based FSTRN, and achieves comparable results with DUF on the Vid4 dataset. When testing on a larger dataset: SPMCs-30, our TDAN is significantly better than the other methods with 0.64dB improvements over the DUF. It should be noted that DUF and FRVSR take 7 and 10 frames as inputs respectively, while our TDAN with 5 frames as inputs which exploits less sequence information.

Visual results on the BD configuration are shown in Fig. 3 and Fig. 4. In comparison with recent DUF [1], benefiting from the strong temporal alignment network, the proposed TDAN is more effective in exploiting information in the supporting frames. Therefore, it is more capable of restoring image structures, for example, the baby face in Walk and the bridge structures in philips_hkc01_001.

Video Quality Evaluation We further compare our TDAN to state-of-the-art VSR networks: Liu *et al*. [3], ToFlow [5], DRVSR [4], and DUF [1] on a video quality assessment metric: VQM_VFD [36] shown in Table 3. The TDAN outperforms recent flow-based methods: Liu *et al*. [3], ToFlow [5], and DRVSR [4], and achieves significantly better performance than DUF on the SPMCs-30. The results further demonstrate that our TDAN is more capable of restoring spatio-temporal structures in nature videos.

Methods	Liu *et al*.	ToFlow	TDAN-BI	DRVSR	DUF	TDAN-BD
Vid4	0.113	0.104	**0.094**	0.100	**0.084**	**0.084**
SPMCs-30	-	0.0049	**0.042**	-	0.043	**0.038**

Table 3. VQM_VFD [36] results under BI and BD settings. Note that the VQM_VFD can measure spatio-temporal consistency quality of restored videos and **smaller is better**.

Model Sizes: Table 4 shows parameter numbers of several networks with the leading VSR performance. We can see that the state-of-the-art SISR networks: RDN, RCAN, and TOFlow, have larger model sizes than the TDAN. Our proposed TDAN has comparable parameter number with the FRVSR and DUF. Even with such a light-weight model, the proposed TDAN still achieves promising VSR performance, which further validates the effectiveness of the proposed one-stage temporal alignment framework.

Methods	RDN	RCAN	TOFlow	FRVSR	DUF	TDAN
Param./M	22.30	15.50	6.20	2.00	1.90	1.97

Table 4. Parameter numbers ($\times 10^6$) of several networks with leading VSR performance.

Models	SISR	MFSR	FlowVSR	D2	D3	D4	D5
PSNR	30.07	30.97	31.17	31.06	31.21	31.32	31.39

Table 5. VSR performance of different baseline models and variants of TDAN on the validation video sequences.

4.3. Ablation Study

To further investigate our TDAN, we compare it with three models: SISR, MFSR, and FlowVSR, which are trained on the Vimeo Super-Resolution dataset same as the TDAN. The SISR model only uses the reference frame as the input, and the MFSR directly concatenates the supporting and reference frames as the input. The FlowVSR adopts optical flow to warp supporting frames, and then feeds the aligned supporting frames and the reference frame into SR reconstruction network. We use SpyNet [40] to predict optical flow for the FlowVSR as in ToFlow [5]. For fair comparisons, the MFSR and FlowVSR networks have a same SR reconstruction network as TDAN. Only the input channel number of the first convolutional layer in SISR is different from others, because only the reference frame is used in the SISR network. In addition, we compare our TDAN models with different numbers: 2, 3, 4, and 5 of deformable convolutional layers, and we denote these networks as D2, D3, D4, and D5, respectively. The four models all have 1 deformable convolutional layer for sampling features from supporting frames and 1 layer before the reconstruction layer in Fig. 2 for adaptively leveraging visual contexts. For D3, D4, and D5, to strength the capability of the offset generator, they have additional 1, 2, and 3 layers, respectively, before the convolutional offset generator in Fig. 2.

Effectiveness of TDAN for VSR Table 5 shows VSR performance of SISR, MFSR, FlowVSR, D2, D3, D4, and D5 networks. We can see that the MFSR outperforms the SISR; FlowVSR, D2, D3, D4, and D5 are better than the MFSR; our D3, D4, and D5 perform better than FlowVSR. These observations demonstrate that exploiting supporting frames even without temporal alignment can improve VSR performance; TDAN and flow-based warping are helpful in handling motion issues and exploiting useful information in supporting frames; the proposed TDAN (*e.g*. D3, D4, and D5) can achieve better performance than the optical flow-based FlowVSR model even with less parameters; more deformable layers can enhance the capability of TDAN. For setting TDAN with comparable model size as FRVSR and DUF, we used D4 in our experiments. From qualitative and quantitative comparisons in Sec. 4.2, we find that even D4 has achieved state-of-the-art VSR performance.

Why TDAN is Capable of Temporal Alignment? Fig-

Supporting Reference FlowVSR TDAN

Figure 7. Temporal alignment results of FlowVSR and TDAN on *Foliage*. Two visual occlusion regions are highlighted. A black car (red box) appears in the supporting frame but does not show in the reference frame; part of a white car shows in the reference frame but not in the supporting frame (green box). The proposed TDAN can effectively alleviate the occlusion issue and restore a photo-realistic image with fine details by exploiting rich visual contexts.

Vid4	City	Walk	Calendar	Foliage	Avg.
FlowVSR	34.49	26.28	30.04	30.73	30.50
TDAN	**49.63**	**48.14**	**44.74**	**46.77**	**47.32**

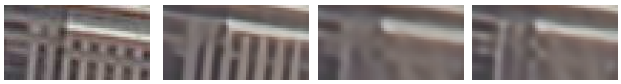
Table 6. Temporal alignment results of FlowVSR and TDAN on 620 LR video frames in the Vid4 dataset.

ure 5 shows visualization of sampling positions on supporting frames based on learned offsets and temporal alignment result on *Temple*. We see that sampling positions tend to capture visual regions with different shapes containing similar content as output pixels for temporal alignment rather than spanning over whole objects as in object detection [7], and the TDAN perfectly aligns the supporting frame with the reference frame. Another example with an additional visual comparison to FlowVSR is illustrated in Fig. 6. We see that the TDAN can handle visual occlusion (pink line) and large motion and deformation (the other two lines) with exploiting image contexts containing similar content as target pixels from the supporting frame, but the optical flow-based FlowVSR fails due to its limited exploration capacity (one pixel). The results demonstrate that the learnable sampling mechanism provides the TDAN strong ability to leverage rich and useful contextual information, which makes the TDAN be effective in employing temporal alignment. Figure 7 further compares temporal alignment performance of TDAN and FlowVSR on *Foliage*. Clearly, the FlowVSR generates a blurry aligned frame and fails to address the occlusion issue. In contrast, our TDAN can well handle visual occlusions with stronger exploration ability.

Table 6 shows quantitative results of different temporal alignment methods. We can find that the TDAN achieves significantly better temporal alignment performance than FLowVSR, which further demonstrates the superiority of the proposed temporal alignment framework.

5. Limitations

In this work, we only use a light-weight TDAN model with only 1.9 million parameters. Even though our TDAN can effectively leveraging temporal information, the smaller model might be not strong enough to recover certain image structures and details. One failure case of the TDAN is shown in Fig. 8. We can see that the TDAN fails to



HR RCAN TOFlow TDAN

Figure 8. A failure case of the TDAN. The very deep SISR network: RCAN can accurately recover the structures of the shown image region in the *city* video frame, but TOFlow and TDAN fail.

recover the structures in the building, but the very deep SISR network: RCAN can accurately reconstruct them, which demonstrates that the LR reference frame can provide enough cues for restoring the structures without requiring additional information from the LR supporting frames. Therefore, it is worth to learn a large model for more accurate structure and detail reconstruction.

In TDAN, we use the LR reference frame as the label to define the $\mathcal{L}_{align}$. However, the LR reference frame is not exactly same as real aligned LR frames, which will make the label noisy. Robust algorithms like [41] for learning under label noise can be considered to improve the $\mathcal{L}_{align}$.

6. Conclusion

In this paper, we propose a one-stage temporal alignment network: TDAN for video super-resolution. Unlike previous optical flow-based methods, which split the temporal alignment problem into two sub-problems: motion estimation and motion compensation, the TDAN implicitly captures motion cues via a deformable sampling module at the feature level and directly predicts aligned LR video frames from sampled features without image-wise warping operations. In addition, the TDAN is capable of exploring image contextual information. With the advanced one-stage temporal alignment design and the strong exploration capability, the proposed TDAN-based VSR framework outperforms the compared several state-of-the-art VSR networks.

Acknowledgments

This work was supported in part by NSF 1741472, 1813709, and 1909912. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018. 1, 2, 5, 6, 7

[2] Jose Caballero, Christian Ledig, Andrew P Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, volume 1, page 7, 2017. 1, 2, 5, 7

[3] Ding Liu, Zhaowen Wang, Yuchen Fan, Xianming Liu, Zhangyang Wang, Shiyu Chang, and Thomas Huang. Robust video super-resolution with learned temporal dynamics. In *ICCV*, pages 2526–2534. IEEE, 2017. 1, 2, 4, 5, 7

[4] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, pages 22–29, 2017. 1, 2, 4, 5, 6, 7

[5] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019. 1, 2, 4, 5, 6, 7

[6] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, pages 6626–6634, 2018. 1, 2, 5, 6

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, Oct 2017. 1, 2, 3, 8

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

[9] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5, 6

[10] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *arXiv preprint arXiv:1812.02898*, 2018. 2

[11] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 2

[12] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Allebach Jan, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time videosuper-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 2

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 2

[15] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, pages 1874–1883, 2016. 2, 4, 5

[16] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate superresolution. In *CVPR*, volume 2, page 5, 2017. 2

[17] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *CVPR*, volume 1, page 5, 2017. 2

[18] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *ICCV*, pages 4539–4547, 2017. 2

[19] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, et al. Ntire 2017 challenge on single image super-resolution: Methods and results. In *CVPR workshops*, pages 1110–1121. IEEE, 2017. 2

[20] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPR workshops*, volume 1, page 4, 2017. 2, 3, 4

[21] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2, 5, 6

[22] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep backprojection networks for super-resolution. In *CVPR*, 2018. 2, 5

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

[24] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[25] Wei Wang, Ruiming Guo, Yapeng Tian, and Wenming Yang. Cfsnet: Toward a controllable feature space for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4140–4149, 2019. 2

[26] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Lcscnet: Linear compressing-based skip-connecting network for image super-resolution. *IEEE Transactions on Image Processing*, 29:1450–1464, 2019. 2

[27] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019. 2

[28] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *ICCV*, December 2015. 2, 5

[29] Li Xu, Jiaya Jia, and Yasuyuki Matsushita. Motion detail preserving optical flow estimation. *TPAMI*, 34(9):1744–1757, 2012. 2

[30] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016. 2, 5

[31] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Schölkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, pages 111–127. Springer, 2018. 2

[32] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV*, September 2018. 3

[33] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, September 2018. 3

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 5

[36] Margaret H Pinson, Lark Kwon Choi, and Alan Conrad Bovik. Temporal video quality model accounting for variable frame delay distortions. *IEEE Transactions on Broadcasting*, 60(4):637–649, 2014. 5, 7

[37] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3897–3906, 2019. 5

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[39] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[40] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, pages 4161–4170, 2017. 7

[41] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, 2013. 8