Space-Time Video Super-Resolution Using Deformable Attention Network

Hai Wang[1], Xiaoyu Xiang[2]*, Yapeng Tian[3], Wenming Yang[1], Qingmin Liao[1]
[1]Tsinghua University [2]Meta Reality Lab [3]University of Rochester
wanghai19@mails.tsinghua.edu.cn, xiangxiaoyu@fb.com, yapengtian@rochester.edu
yang.wenming@sz.tsinghua.edu.cn, liaoqm@tsinghua.edu.cn

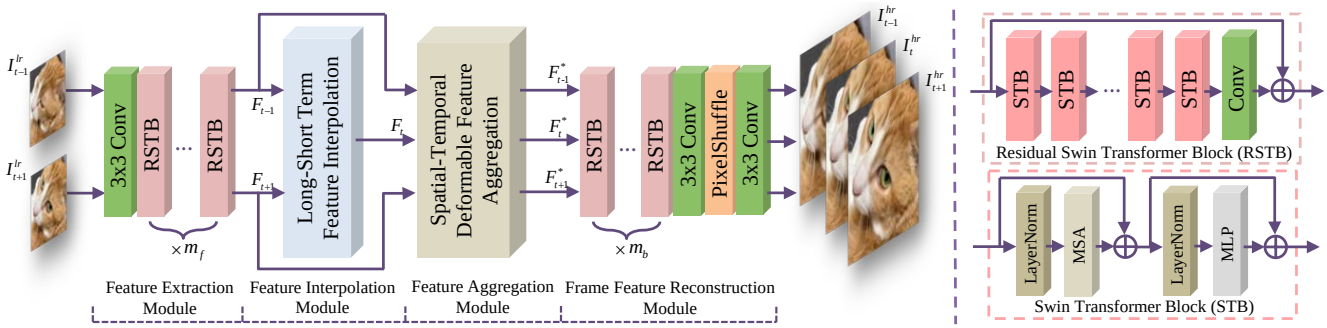


Figure 1. The proposed transformer-empowered spatial-temporal deformable attention network (STDAN) for space-time video super-resolution. Long-short term feature interpolation is capable of exploring neighboring LR frames to synthesize the intermediate frame in the feature space. Spatial-temporal deformable feature aggregation is utilized to capture spatial-temporal contexts by deformable attention.

Abstract

The target of space-time video super-resolution (STVSR) is to increase the spatial-temporal resolution of low-resolution (LR) and low frame rate (LFR) videos. Recent approaches based on deep learning have made significant improvements, but most of them only use two adjacent frames, that is, short-term features, to synthesize the missing frame embedding, which cannot fully explore the information flow of consecutive input LR frames. In addition, existing STVSR models hardly exploit the temporal contexts explicitly to assist high-resolution (HR) frame reconstruction. To address these issues, in this paper, we propose a deformable attention network called STDAN for STVSR. First, we devise a long-short term feature interpolation (LSTFI) module, which is capable of excavating abundant content from more neighboring input frames for the interpolation process through a bidirectional RNN structure. Second, we put forward a spatial-temporal deformable feature aggregation (STDFA) module, in which spatial and temporal contexts in dynamic video frames are adaptively captured and aggregated to enhance SR reconstruction. Experimental results on several datasets demonstrate that our approach outperforms state-of-the-art STVSR methods. The code is available at https://github.com/littlewhitesea/STDAN.

1. Introduction

Space-time video super-resolution (STVSR) aims to reconstruct photo-realistic high-resolution (HR) and high frame rate (HFR) videos from corresponding low-resolution (LR) and low frame rate (LFR) ones. STVSR methods have attracted much attention in the computer vision community since HR slow-motion videos provide more visually appealing content for viewers.

In recent years, deep neural networks have made great progress in solving STVSR. These approaches [2, 7, 8] can simultaneously handle the space and time super-resolution of videos in diverse scenes. Most of them only leverage corresponding two adjacent frames for interpolating the missing frame feature. However, other neighboring input LR frames can also contribute to the interpolation process. In addition, existing one-stage STVSR networks are limited in fully exploiting spatial and temporal contexts among various frames for SR reconstruction. To alleviate these problems, in this paper, we propose a STVSR network called STDAN based on Swin Transformer blocks. The cores of STDAN are (1) a feature interpolation module known as Long-Short Term Feature Interpolation (LSTFI), and (2) a feature aggregation module known as Spatial-Temporal Deformable Feature Aggregation (STDFA).

The LSTFI module containing long-short term cells

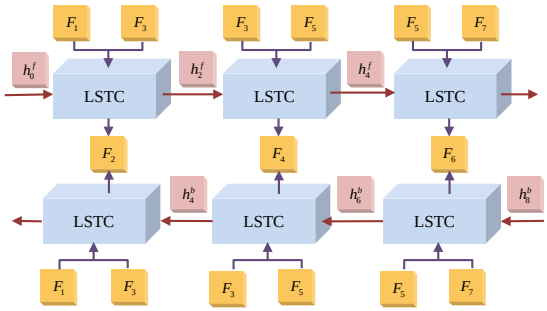


Figure 2. The framework of our long-short term feature interpolation (LSTFI) module. It consists of long-short term cells with bidirectional RNN, which can fully exploit the whole input video frame features during the interpolation process. The structure of the long-short term cell (LSTC) is illustrated in Figure 3(a). Note that the two neighboring frame features and the hidden state from previous LSTC provide short-term and long-term content for interpolation results, respectively.

(LSTCs), utilizes a bidirectional RNN [4] structure to synthesize features for missing intermediate frames. Specifically, to interpolate the intermediate feature, we adopt forward and backward deformable alignment [7] for dynamically sampling two neighboring frame features. Then, the preliminary intermediate feature in the current LSTC is mingled with the hidden state that contains long-range temporal context to synthesize the final interpolated features.

The STDFA module aims to capture spatial-temporal contexts among different frames to enhance SR reconstruction. To dynamically aggregate the spatial-temporal information, we propose to use deformable attention to adaptively discover and leverage relevant spatial and temporal information. The process of STDFA can be divided into two phases: cross-frame spatial aggregation and adaptive temporal aggregation. Through deformable attention, the cross-frame spatial aggregation phase dynamically fuses useful content from different frames. The adaptive temporal aggregation phase mixes the temporal contexts among these fused frame features further to acquire enhanced features.

2. Methodology

The architecture of our proposed network consists of four parts (Figure 1): feature extraction module, long-short term feature interpolation (LSTFI) module, spatial-temporal deformable feature aggregation (STDFA) module and frame feature reconstruction module. Given a low-resolution (LR) and low frame rate (LFR) video with N frames: $\{I^{lr}_{2t-1}\}^{N}_{t=1}$, our STDAN can generate $2N-1$ consecutive high-resolution (HR) and high frame rate (HFR) frames: $\{I^{hr}_{t}\}^{2N-1}_{t=1}$. The input frames are turned to shallow features $\{F_{2t-1}\}^{N}_{t=1}$ by the feature extraction module.

2.1. Long-Short Term Feature Interpolation

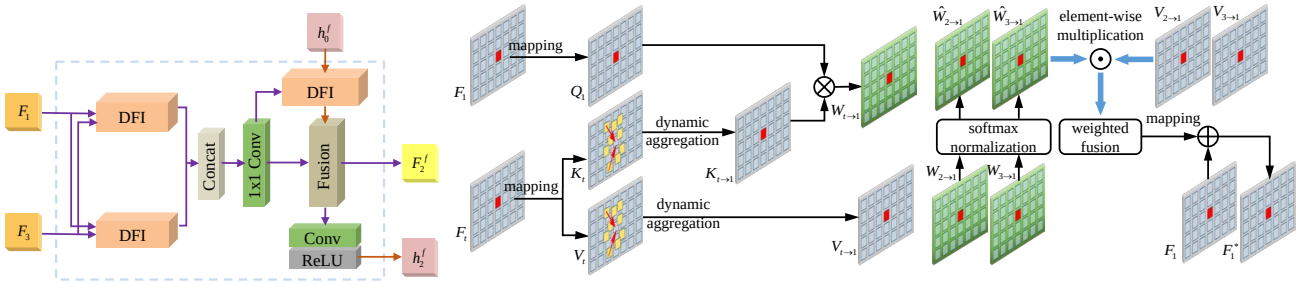
Given the two extracted features: F_1 and F_3, the feature interpolation module can synthesize the feature F_2 corresponding to the missing frame I^{lr}_2. To implement the super-resolution in the time dimension, Xiang *et al.* [7] applied multi-level deformable convolution [5] to perform frame feature interpolation. The learned offset used in deformable convolution can implicitly capture forward and backward motion information and achieve good performance. However, the synthesis of intermediate frame feature [7, 8] only utilizes the two neighboring frame features, which cannot fully explore the information from the other input frames to assist in the process. Unlike feature interpolation in previous STVSR algorithms [7, 8], we propose a long-short term feature interpolation (LSTFI) module to realize the intermediate frame in our STDAN, which is capable of exploiting helpful information from more input frames.

As illustrated in Figure 2, we adopt a bidirectional recurrent neural network (BRNN) [4] to construct the LSTFI module, which consists of two branches in forward and backward direction. Take the forward branch as an example. Two neighboring frame features and the hidden state from the previous long-short term cell (LSTC) are fed into each LSTC, and then the LSTC generates the corresponding intermediate frame feature and current hidden state used for subsequent LSTC. Here, the two neighboring frame features and hidden state serve as short-term and long-term information for the intermediate feature, respectively. However, each branch's hidden state only considers the unidirectional information flow. To fully mine the information flow of these frame features for the interpolation procedure, we fuse interpolation results from LSTCs in the forward and backward branches to acquire the final intermediate frame feature.

2.2. Spatial-Temporal Deformable Feature Aggregation

With the assistance of the LSTFI module, we now have $2N-1$ frame features, where the generation of $N-1$ intermediate frame features combines their adjacent frame features with hidden states. Although the hidden states can introduce certain temporal information, the whole interpolation procedure hardly explicitly explores the temporal information between various frames. In addition, the N input frame features are merely processed independently in the feature extraction module. However, these frame features $\{F_t\}^{2N-1}_{t=1}$ are consecutive, which means there are abundant temporal content without being exploited among these features.

To explore the abundant temporal contexts among frame features, we propose a spatial-temporal deformable feature aggregation (STDFA) module, which can mix cross-frame spatial information adaptively and capture the long-range



(a) The architecture of long-short term cell (b) Cross-Frame Spatial Aggregation (c) Adaptive Temporal Aggregation

Figure 3. The components in the long-short term feature interpolation (LSTFI) and spatial-temporal deformable feature aggregation (STDFA) modules. Note that we only show the case when the number of frame features is 3. Under the case, the value of t can be 2 and 3 for frame feature F_1.

temporal information. Specifically, we utilize the STDFA module to learn the residual auxiliary information from the remaining $2N-2$ frame features for each frame feature F_t. As presented in Figure 3, the processing of the STDFA module can be divided into two parts: spatial aggregation and temporal aggregation. To adaptively fuse cross-frame spatial content of frame feature F_i from the other frame features, we perform deformable attention to each pair: F_i and F_j ($j \in [1, 2N-1]$, $j \neq i$). In detail, frame feature F_i passes through a linear layer to get embedded feature Q_i. Similarly, frame feature F_j is fed into two linear layers to obtain embedded features K_j and V_j, respectively.

To implement deformable attention between F_i and F_j, we first predict the offset map:

$$\Delta M_{j\to i} = H_{og}([Q_i, K_j]) \,, \tag{1}$$

where H_{og} indicates offset generation function consisting of several convolutional layers with $k \times k$ kernel. The offset map $\Delta M_{j\to i}$ at position $\mathbf{p}_o$ is expressed as:

$$\Delta M_{j\to i}(\mathbf{p}_o) = [\Delta\mathbf{p}_1, \Delta\mathbf{p}_2, \cdots, \Delta\mathbf{p}_\xi, \cdots, \Delta\mathbf{p}_{k^2}] \,. \tag{2}$$

Then the offsets $\Delta M_{j\to i}(\mathbf{p}_o)$ are combined with k^2 pre-specified sampling locations to perform deformable sampling. Here, we denote the pre-specified sampling location as $\mathbf{p}_\xi$, and the value set of $\mathbf{p}_\xi$ of $k \times k$ kernel is defined as:

$$\mathbf{p}_\xi \in \left\{ (-\left\lfloor \frac{k}{2} \right\rfloor, -\left\lfloor \frac{k}{2} \right\rfloor), \cdots, (\left\lfloor \frac{k}{2} \right\rfloor, \left\lfloor \frac{k}{2} \right\rfloor) \right\} , \tag{3}$$

where $\lfloor \cdot \rfloor$ denotes rounding down function.

With the offsets $\Delta M_{j\to i}(\mathbf{p}_o)$, the embedded feature vector $Q_i(\mathbf{p}_o)$ can attend k^2 related points in K_j. Nevertheless, not all the information of these k^2 points is helpful for $Q_i(\mathbf{p}_o)$. In addition, each point on embedded feature Q_i needs to search k^2 points, which inevitably causes a large storage occupation. To avoid irrelevant points and reduce storage occupation, we only choose the first T points that are most relevant. To select the T points, we calculate the inner product between two embedded feature vectors as the relevance score:

$$RS_{j\to i}(\mathbf{p}_o, \xi) = Q_i(\mathbf{p}_o) \cdot K_j(\mathbf{p}_o + \mathbf{p}_\xi + \Delta\mathbf{p}_\xi) \,, \tag{4}$$

The larger the score, the more relevant the two points are. According to this criterion, we can determine the T points. In the following, to distinguish the selected T points from original k^2 points, we denote the pre-specified sampling location and learned offset of the T points as $\overline{\mathbf{p}}_\xi$ and $\Delta\overline{\mathbf{p}}_\xi$, respectively.

To adaptively mingle the spatial information from the T locations for each embedded feature vector $Q_i(\mathbf{p}_o)$, we first adopt softmax function to calculate the weight of these points:

$$w_\xi = \frac{e^{Q_i(\mathbf{p}_o)\cdot K_j(\mathbf{p}_o + \overline{\mathbf{p}}_\xi + \Delta\overline{\mathbf{p}}_\xi)}}{\sum_{\xi=1}^{T} e^{Q_i(\mathbf{p}_o)\cdot K_j(\mathbf{p}_o + \overline{\mathbf{p}}_\xi + \Delta\overline{\mathbf{p}}_\xi)}} \,. \tag{5}$$

Then, with the weights and the embedded feature vector $K_j(\mathbf{p}_o + \overline{\mathbf{p}}_\xi + \Delta\overline{\mathbf{p}}_\xi)$, we can obtain corresponding updated embedded feature vector:

$$K_{j\to i}(\mathbf{p}_o) = \sum_{\xi=1}^{T} w_\xi \cdot K_j(\mathbf{p}_o + \overline{\mathbf{p}}_\xi + \Delta\overline{\mathbf{p}}_\xi) \,. \tag{6}$$

Same as $K_{j\to i}(\mathbf{p}_o)$, the updated vector $V_{j\to i}(\mathbf{p}_o)$ can be also achieved with the weight w_ξ. Finally, we calculate the updated relevant weight map $W_{j\to i}$ at each position $\mathbf{p}_o$ between Q_i and $K_{j\to i}$ for the followed temporal aggregation:

$$W_{j\to i}(\mathbf{p}_o) = Q_i(\mathbf{p}_o) \cdot K_{j\to i}(\mathbf{p}_o) \,. \tag{7}$$

To capture the temporal contexts of frame feature vector $F_i(\mathbf{p}_o)$ from the remaining $2N-2$ features, we also apply softmax function to adaptively aggregating feature vectors $V_{j\to i}(\mathbf{p}_o)$. Specifically, the normalized temporal weight of each vector $V_{j\to i}(\mathbf{p}_o)$ ($j \in [1, 2N-1]$, $j \neq i$) is expressed as:

$$\hat{W}_{j\to i}(\mathbf{p}_o) = \frac{e^{W_{j\to i}(\mathbf{p}_o)}}{\sum_{j=1}^{2N-1, j\neq i} e^{W_{j\to i}(\mathbf{p}_o)}} \,. \tag{8}$$

Then, through fusion embedded feature vector $V_{j\to i}(\mathbf{p}_o)$ ($j \in [1, 2N-1]$, $j \neq i$) with the corresponding normalized weight, we can attain the embedded feature V_i^* that aggregates the spatial and temporal contexts from other $2N-2$ embedded features. The weighted fusion process is defined

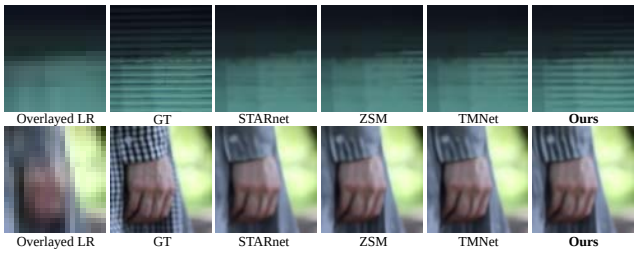


Figure 4. Visual comparisons of different STVSR approaches on Vimeo. We can see that our model can recover more structures.

Table 1. Quantitative comparisons of our STDAN and other SOTA methods. The Top-2 results are highlighted in red and blue colors.

STVSR Method	Vid4		Vimeo-Slow		Vimeo-Medium		Vimeo-Fast		Params
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	(M)
STARnet [1]	25.99	0.7819	33.10	0.9164	34.86	0.9356	36.19	0.9368	111.61
ZSM [7]	26.14	0.7974	33.36	0.9138	35.41	0.9361	36.81	0.9415	11.10
TMNet [8]	26.23	0.8011	33.51	0.9159	35.60	0.9380	37.04	0.9435	12.26
STDAN (Ours)	26.28	0.8041	33.66	0.9176	35.70	0.9387	37.10	0.9437	8.29

as:

$$V_i^*(\mathbf{p}_o) = \sum_{j=1}^{2N-1, j\neq i} \hat{W}_{j\to i}(\mathbf{p}_o) \cdot V_{j\to i}(\mathbf{p}_o)\,. \quad (9)$$

In the tail of STDFA module, the embedded feature V_i^* is sent into a linear layer to acquire the residual auxiliary feature F_i^{res}. Finally, we add frame feature F_i and residual auxiliary feature F_i^{res} to get the enhanced feature F_i^* that aggregates spatial and temporal contexts from the other $2N-2$ frame features.

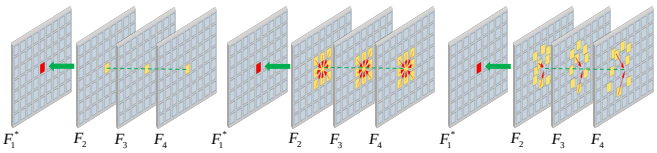
3. Example Results

We use the Vimeo-90K dataset [9] to train our network. Its three test subsets: *Vimeo-Slow*, *Vimeo-Medium* and *Vimeo-Fast* serve as the evaluation datasets. In addition, we also report the results on Vid4 [3] of different approaches. To compare diverse STVSR networks quantitatively, Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [6] are adopted as evaluation metrics.

We compare our STDAN with existing state-of-the-art (SOTA) one-stage STVSR approaches: STARnet [1], ZSM [7] and TMNet [8]. Quantitative results of various STVSR methods are shown in Table 1. From the table, we can see that our STDAN with the least parameters obtains SOTA performance on both Vid4 [3] and Vimeo [9].

Visual comparison of different models are displayed in Figure 4. We observe that our STDAN, with the proposed LSTFI and STDFA modules, restores more accurate structures and is more capable of handling motion blurs compared with other STVSR approaches.

Ablation Study on Feature Aggregation To valid the effect of the proposed spatial-temporal deformable feature aggregation (STDFA) module, we establish a baseline: model Ω_1. It only adopts short-term information to perform interpolation, and then directly reconstructs HR video frames through the feature reconstruction module without feature aggregation process. In contrast, we compare three different models: Ω_2, Ω_3 and Ω_4 with feature aggregation. For the spatial-temporal feature aggregation process in the model Ω_2, illustrated in Figure 5(a), each feature vector aggregates the information at the same position of other frame features, that is, the feature vector attends the valuable spatial content in a 1×1 window. We enlarge the window size of the model Ω_3 to 3. Considering large motions between frames. A deformable window is applied in the model Ω_4. As shown in Figure 5(c), model Ω_4 adopts the STDFA module to perform feature aggregation.



(a) STFA in a 1x1 fixed window (b) STFA in a 3x3 fixed window (c) STFA in a deformable window

Figure 5. Three different aggregation methods in the feature aggregation module. 'STFA' refers to spatial-temporal feature aggregation. Note that we only show 4 frames for a illustration, and STFA in a deformable window denotes our STDFA module.

Table 2. Ablation study on the proposed modules. Our long-short term feature interpolation leverages more input LR frames to assist in the interpolation process. The proposed spatial-temporal feature aggregation in the deformable window can adaptively capture spatial-temporal contexts among different frames for HR frame reconstruction.

Method		Ω_1	Ω_2	Ω_3	Ω_4	Ω_5
Params (M)		5.44	5.54	5.54	5.82	8.29
Feature Interpolation	Short-term	✓	✓	✓	✓	
	Long-short term					✓
Feature Aggregation	1×1 fixed window		✓			
	3×3 fixed window			✓		
	deformable window				✓	✓
Vid4 (slow motion)		25.27	25.69	25.85	25.97	26.28
Vimeo-Fast (fast motion)		35.88	36.22	36.41	36.63	37.10

Quantitative results on Vid4 [3] and *Vimeo-Fast* [9] datasets are shown in Table 2. From the table, we know that: (1) Feature aggregation module can improve the reconstruction results; (2) The larger the spatial range of feature aggregation, the more useful information can be captured to enhance recovery quality of HR frames.

Ablation Study on Feature Interpolation To investigate the effect of the proposed long-short term feature interpolation (LSTFI) module, we compare two models: Ω_4 and Ω_5. As shown in Figure 2, the model Ω_5 with LSTFI can exploit short-term information of two neighboring frames and long-term information of hidden states from other LSTCs. In comparison, model Ω_4 only uses two adjacent frames to interpolate the feature of the intermediate frame. From Table 2, combining long-term and short-term information can achieve better feature interpolation results, which leads to high-quality HR frames with more details.

References

[1] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *CVPR*, 2020. 4

[2] Jaeyeon Kang, Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. Deep space-time video upsampling networks. In *ECCV*, 2020. 1

[3] Ce Liu and Deqing Sun. On bayesian adaptive video super resolution. *TPAMI*, 2013. 4

[4] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *TSP*, 1997. 2

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

[6] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 4

[7] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *CVPR*, 2020. 1, 2, 4

[8] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *CVPR*, 2021. 1, 2, 4

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