Bi-RSTU: Bidirectional Recurrent Upsampling Network for Space-Time Video Super-Resolution

Hai Wang, Wenming Yang, Qingmin Liao, and Jie Zhou

***Abstract*—One-stage space-time video super-resolution (STVSR) aims to directly reconstruct high-resolution (HR) and high frame rate (HFR) video from its low-resolution (LR) and low frame rate (LFR) counterpart. Due to the wide application, one-stage STVSR has drawn much attention recently. However, existing one-stage methods suffer from ineffective exploration of the auxiliary information from adjacent time steps that may be useful to STVSR at the current time step. To address this issue, we propose a novel Bidirectional Recurrent Space-Time Upsampling network called Bi-RSTU for one-stage STVSR to utilize auxiliary information at various time steps. Specifically, an efficient channel attention feature interpolation (ECAFI) module is devised to synthesize the intermediate frame's LR feature by exploiting its two neighboring LR video frame features. Subsequently, we fuse the information from the previous time step into these intermediate and neighboring features. Finally, second-order attention spindle (SOAS) blocks are stacked to form the feature reconstruction module that learns a mapping from LR fused feature space to HR feature space. Experimental results on public datasets demonstrate that our Bi-RSTU shows competitive performance compared with current two-stage and one-stage state-of-the-art STVSR methods.**

***Index Terms*—Space-time video super-resolution, bidirectional recurrent neural network, feature interpolation, feature reconstruction.**

I. Introduction

THE goal of space-time video super-resolution (STVSR) is to reconstruct a high-resolution (HR) and high frame rate (HFR) video from its corresponding low-resolution (LR) and low frame rate (LFR) video. Since HR and HFR videos are more visually appealing to users than LR and LFR ones, STVSR algorithms are employed to improve the resolution and frame rate of legacy videos displayed on high-definition television (HDTV). Existing methods used to realize STVSR can be divided into two categories: two-stage methods and one-stage methods. Two-stage approaches usually mean synthesizing intermediate video frames first to increase the frame rate of LR and LFR videos through time super-resolution (SR) and then implementing space SR for the processed LR and HFR videos. In general, the procedure involves two models, one is used for time SR, and the other is responsible for space SR. In contrast, one-stage methods directly reconstruct HR and HFR videos from their corresponding LR and LFR videos without generating intermediate LR frames. Generally, these one-stage algorithms conduct time SR and space SR by one single model simultaneously.

Fig. 1. Results of space-time video super-resolution (STVSR). Given two low-resolution (LR) frames, STVSR can reconstruct their intermediate high-resolution (HR) frame. The two LR frames' overlayed result, the ground truth (GT) of the intermediate HR frame, and outputs of several STVSR methods are presented. Compared with other STVSR methods, our Bi-RSTU network restores more image details.

Time SR in two-stage methods, also called video frame interpolation (VFI), can generate non-existent video frames to increase the frame rate. Owing to its wide range of applications, such as visual quality enhancement and slow-motion synthesis, VFI has drawn much attention in the past few years [1], [2], [3]. For space SR in two-stage methods, it aims to recover a HR video frame from its corresponding LR frame. Concretely, space SR consists of single image super-resolution (SISR) and video super-resolution (VSR). The main difference between SISR and VSR is the number of input frames. Since space SR is a classic low-level visual task, it has been investigated extensively [5], [8], [7], [9]. Through combining these time and space SR algorithms, STVSR can be achieved by sequentially carrying out the two models separately.

Considering that the space and time dimensions of videos are intra-related, several one-stage traditional methods [10], [11], [12] have been proposed to realize STVSR. Nevertheless, one-stage STVSR is a more serious ill-posed problem than space SR, because it also requires synthesizing the

This work was partly supported by the National Natural Science Foundation of China(No.62171251), the Natural Science Foundation of Guangdong Province(No.2020A1515010711), the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(Nos. JCYJ20200109143010272, JCYJ20200109143035495, CJGJZD20210408092804011 & JSGG20211108092812020) and Oversea Cooperation Foundation of Tsinghua.

H. Wang, W. Yang and Q. Liao are with Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University, Shenzhen, China (e-mail: wanghai19@mails.tsinghua.edu.cn; yangelwm@163.com; liaoqm@tsinghua.edu.cn).

J. Zhou is with Department of Automation, Tsinghua University, Beijing, China (e-mail: jzhou@tsinghua.edu.cn).

Manuscript received April 19, 2005; revised August 26, 2015.

intermediate video frame except for increasing the spatial resolution of the video frame. These traditional methods with strong assumptions cannot cope with videos well in diverse scenes. Thanks to the powerful nonlinear fitting ability, deep learning has greatly promoted the performance of low-level and high-level visual tasks [13], [14], [15] in recent years. Particularly, Xiang *et al.* [16] proposed Zooming Slow-Mo, a one-stage STVSR network based on deep learning. Owing to the effective network modules and the close connections between time SR and space SR, Zooming Slow-Mo surpasses many two-stage STVSR methods in terms of reconstruction quality and operation efficiency. Although Zooming Slow-Mo achieves good STVSR performance on several public datasets, it cannot effectively explore the auxiliary information from adjacent time steps that may be useful to STVSR at the current time step due to its network architecture.

To make use of auxiliary information at various time steps, we propose a one-stage STVSR network named Bi-RSTU. It consists of basic recurrent space-time upsampling (RSTU) cells in a bidirectional recurrent neural network (RNN) [45] manner. Specifically, we put forward an efficient channel attention feature interpolation (ECAFI) module and a reconstruction module composed of second-order attention spindle (SOAS) blocks to constitute the RSTU cell. The ECAFI module can effectively distinguish the importance of different feature channels, thereby exploiting two LR neighboring frame features to synthesize their intermediate frame feature. To leverage the auxiliary information from the previous RSTU cell to assist the space-time super-resolution process at the current RSTU cell, we fuse the auxiliary information into the intermediate and neighboring features at the current RSTU cell. Then, SOAS blocks are stacked to form the feature reconstruction module that learns a mapping from LR fused feature space to HR feature space. Finally, our Bi-RSTU generates HR and HFR video frames from LR and LFR video frames sequentially with the modules mentioned above.

The contributions of this paper are summarized as follows:

- We design a one-stage STVSR network called Bi-RSTU based on the bidirectional RNN structure, which utilizes auxiliary information from adjacent time steps to assist the space-time super-resolution at the current time step. To the best of our knowledge, this is the first time to construct the framework of bidirectional RNN for one-stage STVSR.
- An efficient channel attention feature interpolation (ECAFI) module is proposed in this paper to perform frame interpolation in the feature space without additional optical flow, which makes our Bi-RSTU effectively synthesize the intermediate video frame.
- Experimental results on several public datasets demonstrate that our Bi-RSTU achieves competitive performance compared with other two-stage and one-stage state-of-the-art STVSR methods.

The rest of the paper is organized as follows. Section II presents a comparative review of related works. The proposed Bi-RSTU network is described in detail in Section III. Section IV shows quantitative and qualitative results of various methods on public datasets. Section V discusses possible works in the future. Finally, conclusions are drawn in Section VI.

II. Related Work

A. Space Super-Resolution

Space super-resolution (SR) algorithms have been proposed to convert low-resolution (LR) frames into high-resolution (HR) frames. According to the number of input frames, these algorithms can be divided into single image super-resolution (SISR) and video super-resolution (VSR) methods. SISR means that the SR process only involves one frame, while VSR indicates that the SR process contains multiple consecutive frames as inputs. SRCNN [13] is the pioneering work of SISR based on deep learning (DL), and it significantly exceeds the reconstruction quality of traditional methods. From then on, more complex network structures [6], [18], [19], [50] and deeper network layers [5], [17] are explored to improve the performance of SISR. Compared to SISR, the utilization of temporal information is vital for VSR. Therefore, VSR algorithms based on DL pay more attention to the alignment and fusion of information between frames. Several VSR methods [20], [21], [22] adopt optical flow to accomplish the alignment. However, the calculation cost of optical flow is high, and the wrong optical flow may seriously affect the quality of SR reconstruction. Hence, Tian *et al.* [23] proposed TDAN to achieve the alignment of two video frames in the feature space with deformable convolution (DC) [24], [25], avoiding the use of optical flow. Considering that TDAN only utilizes a single scale of DC, EDVR [9] adopts a multi-scale DC [24], [25] to perform alignment and then uses a spatial-temporal attention module to realize the fusion of various frame features. For other VSR algorithms [8], [30] without alignment process, advanced modules, such as the dynamic convolution and non-local module, are employed to mix information between frames. Recently, Isobe *et al.* [48] proposed RRN based on a unidirectional recurrent neural network (RNN) structure for VSR. In detail, RRN achieves the temporal alignment through the unidirectional transmission of hidden states and obtains satisfactory performance. Similarly, we also adopt the RNN structure to realize the temporal alignment through hidden states' transmission. Specifically, we put forward a bidirectional RNN structure in our network. Compared to the unidirectional structure used in RRN, this structure enables our network to explore more temporal contexts among video frames.

B. Time Super-Resolution

Time super-resolution, also called video frame interpolation (VFI), aims to synthesize non-existent frames in-between video frames. VFI is essential for the scenarios that require high frame rate (HFR) videos since it takes enormous costs to directly acquire HFR videos through capture devices. In recent years, VFI methods based on deep learning have made significant improvements. SepConv [1] employs a separate adaptive convolution to estimate convolution kernels and then generates an intermediate video frame with these kernels. To handle the occlusion, Niklaus *et al.* developed CtxSyn

[32] which introduces context cues to assist the process of synthesizing video frames. However, the approach to cope with occlusion in CtxSyn [32] is implicit, so a depth-aware module in DAIN [4] is proposed to explicitly detect occlusion for video frame interpolation. Recently, based on the residual channel attention (RCA) block [5], Choi *et al.* [33] put forward an end-to-end trainable VFI network, called CAIN, without requiring additional information, such as optical flow. Owing to the powerful attention mechanism, CAIN has achieved promising performance on the VFI task. Inspired by CAIN, our one-stage STVSR network utilizes residual efficient channel attention (RECA) blocks to interpolate the intermediate frame in the feature space. Compared to the RCA block, the RECA block in our model is more efficient.

C. Space-time Super-Resolution

The two tasks: time super-resolution and space super-resolution, are closely related. Feature maps with a higher resolution can provide finer structure details for synthesizing the intermediate frame. Meanwhile, a video with higher frame rate can offer more accurate pixel motion information to assist the alignment of features between frames. Hence, Shechtman *et al.* [10] originally proposed a one-stage traditional space-time video super-resolution (STVSR) algorithm through combining spatial and temporal information from multiple video sequences. Later, several methods [11], [12] are developed to achieve STVSR. However, these traditional methods are often applied to specific videos due to the hypotheses made. Recently, Xiang *et al.* [16] put forward Zooming Slow-Mo, a one-stage STVSR algorithm based on powerful deep learning (DL) without any assumptions. Therefore, the algorithm can tackle videos in different scenes. Specifically, Zooming Slow-Mo adopts a multi-scale deformable convolution [24], [25] to perform video frame interpolation in the feature space and then enhance the consistency between video frames with a deformable ConvLSTM module in a bidirectional manner. With these effective modules, Zooming Slow-Mo achieves state-of-the-art (SOTA) performance. Based on the structure of Zooming Slow-Mo, Xu *et al.* [28] further put forward a temporal modulation module for controllable STVSR, which can flexibly interpolate intermediate frames. In addition, Haris *et al.* [27] utilized spatial-temporal informative relationships to construct STVSR network with an extra motion estimation branch. However, these one-stage STVSR methods suffer from ineffective exploration of the auxiliary information between adjacent time steps owing to their adopted network structures. Therefore, we propose a one-stage STVSR method based on a bidirectional recurrent neural network (RNN) structure [45] to exploit the auxiliary information for a better STVSR result. Specifically, we fuse the information from the previous time step into the frame features that are processed at the current time step, and then the fused features pass through our feature reconstruction module to generate high-resolution (HR) outputs.

III. Proposed Method

In this section, we first introduce our bidirectional recurrent space-time upsampling network (Bi-RSTU) for one-stage STVSR in detail. Then, the efficient channel attention feature interpolation (ECAFI) module is elaborated. Finally, we describe the feature reconstruction module of Bi-RSTU.

A. Network Architecture

Given a low-resolution (LR) and low frame rate (LFR) video $V_{LL} = \{LR_{2t+1}\}_{t=0}^{n}$, the goal of Bi-RSTU is to estimate the corresponding high-resolution (HR) and high frame rate (HFR) video $V_{HH} = \{HR_t\}_{t=1}^{2n+1}$. The overall pipeline of Bi-RSTU is shown in Fig. 2(a). We can see it is composed of basic recurrent space-time upsampling (RSTU) cells in a bidirectional manner. The final SR reconstruction results are jointly determined by the outputs of the forward and backward branches.

At time step t, the architecture of RSTU cell is presented in Fig. 2(b). Given two low-resolution frames LR_{2t-1} and LR_{2t+1}, one feature extraction module is adopted to extract shallow features F_{2t-1} and F_{2t+1} from the two frames. Then, we employ the feature interpolation module H_{fi} to synthesize the shallow feature of their intermediate frame:

$$F_{2t} = H_{fi}(F_{2t-1}, F_{2t+1}). \tag{1}$$

Considering that the video frames obtained at adjacent time steps are closely similar for the STVSR task, we believe that the information from the previous time step is beneficial for space-time super-resolution (STSR) at the current time step. Hence, we fuse the information O_{2t-3}, O_{2t-2} and h_{t-1} of the STSR process at time step $t-1$ into the three shallow features F_{2t-1}, F_{2t} and F_{2t+1} at time step t to produce fused features:

$$F_{2t-1}^{*,0} = W_{Conv}\{Concat[O_{2t-3}, h_{t-1}, F_{2t-1}, F_{2t}]\}, \tag{2}$$

$$F_{2t}^{*,0} = W_{Conv}\{Concat[O_{2t-2}, F_{2t-1}, F_{2t}, F_{2t+1}]\}, \tag{3}$$

where O_{2t-3} and O_{2t-2} are HR features, h_{t-1} is the hidden state that is equivalent to feature F_{2t-2} at time step $t-1$, $Concat\,[,]$ denotes the concatenation operation along the channel dimension, and W_{Conv} refers to a convolutional layer. Here, the fused features $F_{2t-1}^{*,0}$ and $F_{2t}^{*,0}$ integrate the information of their adjacent frames and adopt the previously generated HR features as a guide for the next reconstruction process.

The fused features $F_{2t-1}^{*,0}$ and $F_{2t}^{*,0}$ pass through the feature reconstruction module H_{rec} and a convolutional layer sequentially to generate HR features O_m at time step t:

$$O_m = W_{Conv}\{H_{rec}(F_m^{*,0})\}, \quad m \in \{2t-1, 2t\}. \tag{4}$$

In addition, to acquire the hidden state h_t at time step t, we apply a convolutional layer and an activation layer to the output of the reconstruction module about feature $F_{2t}^{*,0}$.

$$h_t = \sigma(W_{Conv}\{H_{rec}(F_{2t}^{*,0})\}), \tag{5}$$

where σ denotes the ReLU function. At the end of the RSTU cell, the HR feature O_m is upscaled via a shuffle module H_{sf} to obtain the unidirectional HR video frame at time step t:

$$HR_m^u = H_{sf}(O_m), \quad m \in \{2t-1, 2t\}, \tag{6}$$

where the shuffle module is only composed of a PixelShuffle layer [49].

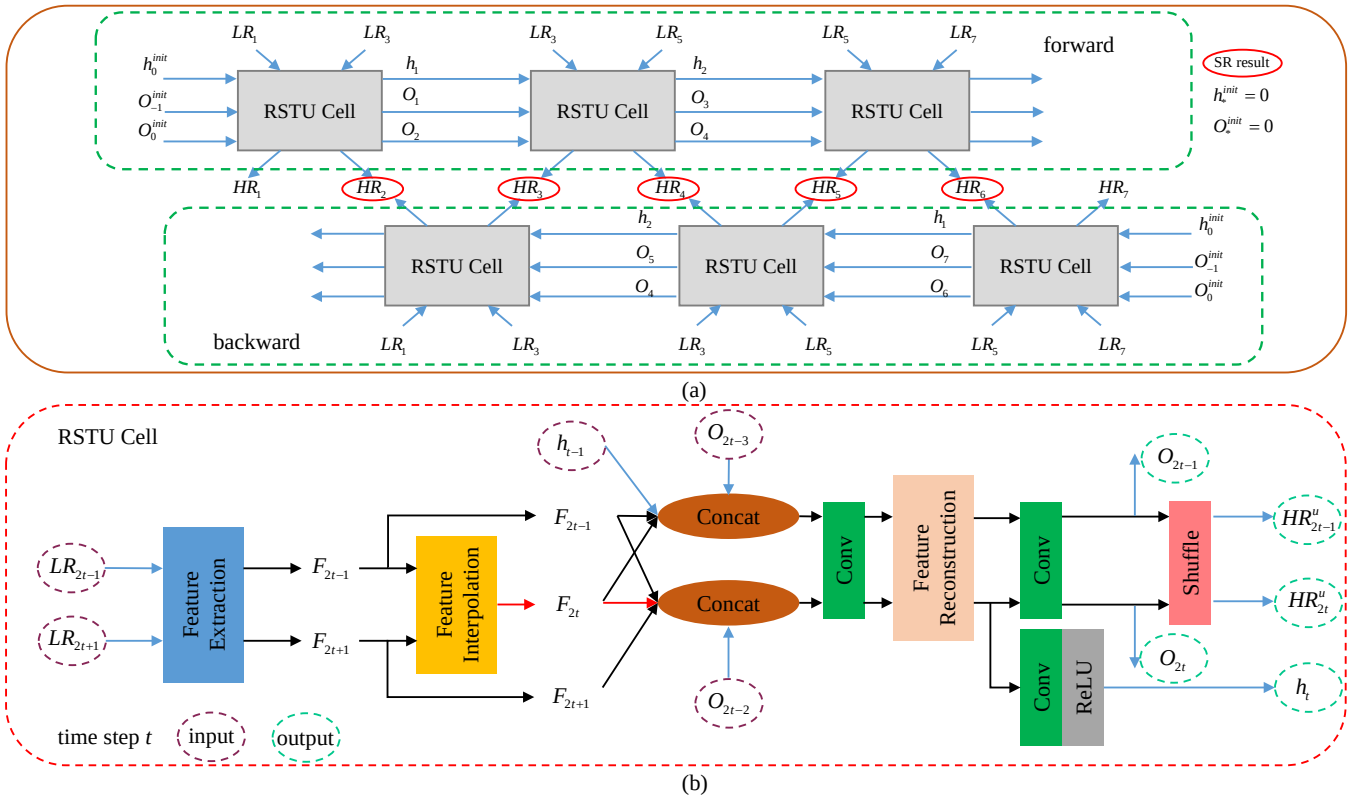

Fig. 2. (a) The pipeline of our Bi-RSTU network. Bi-RSTU is constructed in a bidirectional RNN manner, and it consists of basic recurrent space-time upsampling (RSTU) cells in the forward and backward branches. Note that we only present three RSTU cells in each branch for an illustration; (b) The constituted modules, inputs and outputs of the RSTU cell at time step *t*. Specifically, two HR features (O_{2t-3}, O_{2t-2}), hidden state (h_{t-1}) from time step *t-1*, and two neighboring LR frames (LR_{2t-1}, LR_{2t+1}) serve as the inputs of the RSTU cell at time step *t*, while the corresponding outputs of the RSTU cell are two HR features (O_{2t-1}, O_{2t}), hidden state (h_t) used for the next time step *t+1*, and two unidirectional HR frames (HR^u_{2t-1}, HR^u_{2t}), respectively.

The RSTU network consisting of RSTU cells in a unidirectional manner only uses the information at the previous time step from one direction. However, the outputs of adjacent time steps are similar for STVSR tasks, and the utilization of the information from two directions may be beneficial for the reconstruction of HR frames at the current time step. Thus, we adopt the RSTU network in a bidirectional manner to form our Bi-RSTU.

Here, we define the outputs HR^u_m from the forward and backward directions of Bi-RSTU as $HR^{u,f}_m$ and $HR^{u,b}_m$, respectively. The final HR output frame of Bi-RSTU is expressed as:

$$HR_m = H_{fh}(Concat[HR^{u,f}_m, HR^{u,b}_m]), \tag{7}$$

where $H_{fh}(\cdot)$ refers to the fusion block used for the HR video frames from the two directions, and it is made up of two convolutional layers and a PReLU function [43].

To optimize our proposed network, we adopt L1 penalty function as our final loss. Given K LR and LFR videos $V^i_{LL} = \{LR^i_{2t+1}\}^n_{t=0}$ and their corresponding HR and HFR videos $V^i_{GT} = \{GT^i_t\}^{2n+1}_{t=1}$, the goal of training the proposed network is to optimize loss function:

$$L_{stvsr} = \frac{1}{K}\sum_{i=1}^{K}\left[\frac{1}{2n+1}\sum_{t=1}^{2n+1}||HR^i_t - GT^i_t||_1\right], \tag{8}$$

where HR^i_t and GT^i_t are t-th output of Bi-RSTU and the corresponding ground truth about video V^i_{LL}, respectively. Based on the loss function, our Bi-RSTU can be end-to-end trainable and achieves the purpose of one-stage STVSR.

B. Feature Interpolation Module

Given the two shallow features F_{2t-1} and F_{2t+1}, feature interpolation module H_{fi} aims to interpolate the feature F_{2t} of their intermediate frame LR_{2t}. In general, optical flow [3] is explored to complete the interpolation process. However, the calculation cost of optical flow is high, and incorrectly estimated optical flow may seriously affect the interpolation quality. Therefore, Choi *et al.* [33] proposed CAIN, a video frame interpolation (VFI) network based on channel attention (CA) without using optical flow, and it achieves impressive performance on the VFI task. Recently, Wang *et al.* [34] put forward an efficient channel attention (ECA) module. It has been proven that the ECA module is more efficient than the CA module in image classification and instance segmentation. Therefore, inspired by CAIN [33] and the ECA module [34], we develop an efficient channel attention feature interpolation (ECAFI) module as H_{fi} in our STVSR network.

The architecture of the ECAFI module is shown in Fig. 3. Note that the interpolation module is composed of two efficient channel attention groups (ECAGs) with skip connections. By means of the residual-in-residual structure, the ECAFI module

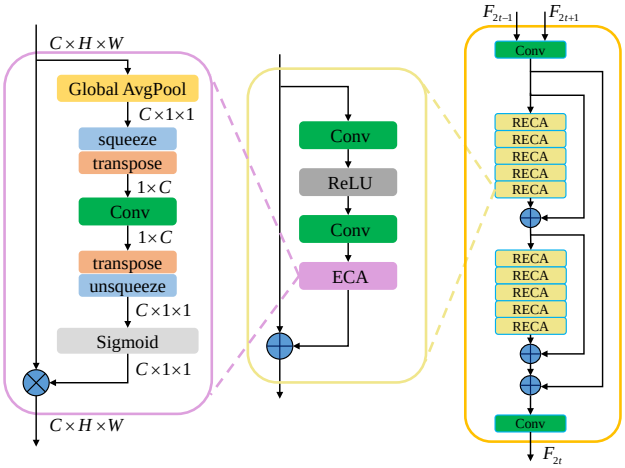

Fig. 3. The architecture of the efficient channel attention feature interpolation (ECAFI) module. 'squeeze' and 'unsqueeze' denote the operation of compressing and expanding one dimension of a tensor, respectively.

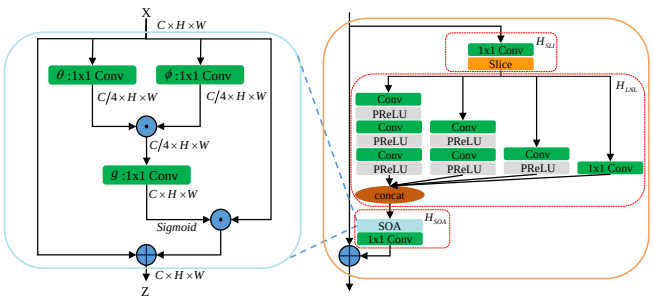

Fig. 4. The architecture of the second-order attention spindle (SOAS) block. The three basic units of the SOAS block are H_{SLI}, H_{LNL} and H_{SOA}, respectively. Specifically, H_{SLI} refers to the slicing operation. Linear and nonlinear information of features are excavated in H_{LNL}. H_{SOA} is used to enhance pixel-wise features further.

can explore hierarchical information from two neighboring frame features to synthesize the intermediate frame feature. Specifically, the ECAG consists of five residual efficient channel attention (RECA) blocks, and each RECA block comprises two convolutional layers, a ReLU activation function, and one ECA module [34]. Note that the RECA block follows a similar architecture as the RCA block [5]. Owing to effective RECA blocks and the residual-in-residual structure, the ECAFI module can fully exploit hierarchical information of the two neighboring frame features to synthesize the intermediate frame feature.

With the assistance of the ECAFI module, we can interpolate the feature of the intermediate frame. Specifically, the two shallow features F_{2t-1} and F_{2t+1} are concatenated along the channel dimension. Subsequently, we reduce the number of channels by half with a 3×3 convolutional layer to obtain the original intermediate frame feature:

$$F_{2t}^{0} = W_{Conv}\{Concat\,[F_{2t-1}, F_{2t+1}]\}. \tag{9}$$

Then, the output F_{2t}^{1} and F_{2t}^{2} of the first and the second ECAG can be formulated as:

$$F_{2t}^{1} = F_{2t}^{0} + H_{reca}^{5}(H_{reca}^{4}(\cdots(H_{reca}^{1}(F_{2t}^{0}))\cdots)), \tag{10}$$
$$F_{2t}^{2} = F_{2t}^{1} + H_{reca}^{10}(H_{reca}^{9}(\cdots(H_{reca}^{6}(F_{2t}^{1}))\cdots)), \tag{11}$$

where H_{reca}^{i} indicates the i-th RECA block in the ECAFI module. At the tail of the ECAFI module, we adopt a convolutional layer to acquire the shallow intermediate frame feature:

$$F_{2t} = W_{Conv}\{F_{2t}^{0} + F_{2t}^{2}\}. \tag{12}$$

C. Feature Reconstruction Module

For the LR fused features, the target of the feature reconstruction module is to map them to the HR feature space. To construct our feature reconstruction module that learns a better mapping, we propose a second-order attention spindle (SOAS) block, where the linear and nonlinear information of the fused features are fully exploited. As shown in Fig. 4, the SOAS block mainly consists of a spindle subblock [35] and a second-order attention subblock [42]. Here, the core idea of SOAS is to integrate linear and nonlinear LR features through the spindle subblock and then employ the second-order attention (SOA) subblock to enhance features pixel-wisely further.

Given the fused features $F_{m}^{*,0}$ $(m \in \{2t-1, 2t\})$ defined as Eq.(2) and Eq.(3), the output of SOAS can be defined as:

$$F_{m}^{*,1} = F_{m}^{*,0} + W_{Conv}\{(H_{SOA}(H_{LNL}(H_{SLI}(F_{m}^{*,0}))))\}, \tag{13}$$

where H_{SLI}, H_{LNL} and H_{SOA} are three basic units of the SOAS block. Specifically, H_{SLI} is composed of a convolutional layer followed by a slice operation. The output of the convolutional layer is sliced into four parts evenly along the channel dimension. Then, for the first three parts of $H_{SLI}(F_{m}^{*,0})$, H_{LNL} adopts diverse combinations of the convolutional layer and PReLU function [43] to fully explore the nonlinear information. Meanwhile, it employs a single convolutional layer to excavate the linear information of the fourth part. Next, we concatenate the nonlinear and linear information along the channel dimension in the last step of H_{LNL}. Considering that pixels in different locations have various importance for SR reconstruction, at the end of the SOAS block, we utilize a second-order attention (SOA) subblock followed by a 1×1 convolutional layer to enhance the output feature of H_{LNL} pixel-wisely.

The operation of the SOA subblock is clearly described in Fig. 4. Given a tensor X with a size of $C \times H \times W$, the overall process of the SOA subblock can be expressed as:

$$Z = X + X \cdot Sigmoid(g(\theta(X) \cdot \phi(X))), \tag{14}$$

where θ, ϕ, and g are 1×1 convolutional layers with different parameters. With the SOA subblock, we can obtain the output Z strengthening important pixel-wise features of input X for SR reconstruction.

In this work, we stack D SOAS blocks to form our feature reconstruction module H_{rec}. The detailed process of H_{rec} can be presented as:

$$O_{m}^{*} = H_{soas}^{D}(\cdots(H_{soas}^{1}(F_{m}^{*,0}))\cdots), \quad m \in \{2t-1, 2t\}, \tag{15}$$

where H_{soas}^{i} denotes the i-th SOAS block, and O_{m}^{*} is the output of feature reconstruction module H_{rec}. Afterwards, we

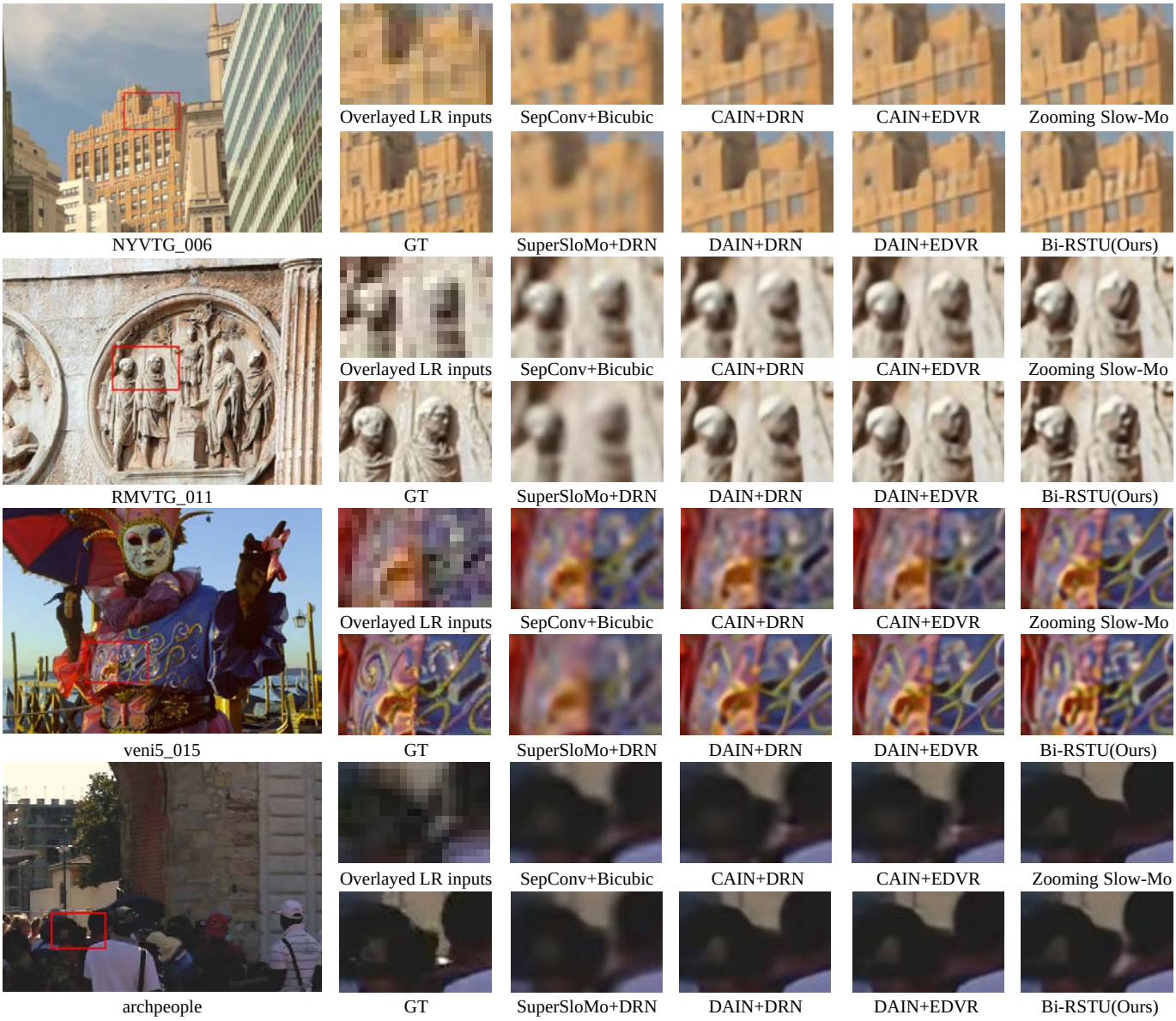

Fig. 5. Qualitative comparisons of different STVSR methods on SPMC-11 [39] and UDM10 [40] datasets. Compared with six two-stage STVSR algorithms and the one-stage STVSR network: Zooming Slow-Mo, our one-stage Bi-RSTU network obtains better visual quality and restores more structures and texture details.

can acquire HR features O_m $(m \in \{2t-1, 2t\})$ and hidden state h_t at time step t through Eq.(4) and Eq.(5).

IV. Experiments

In this section, we first introduce the datasets and evaluation metrics adopted in our experiments. Next, the implementation details of Bi-RSTU are elaborated. Also, we compare our proposed network with current two-stage and one-stage state-of-the-art (SOTA) STVSR methods on several public datasets. Finally, ablation studies are applied to the proposed modules in Bi-RSTU.

A. Datasets

We use the Vimeo-90K dataset [20] to train our proposed network. Specifically, the Vimeo-90K dataset contains more than 60,000 videos, and each video has seven frames that are regarded as our HR and HFR supervisions. The corresponding four LR and LFR frames are downscaled by a factor of 4 with bicubic interpolation from these odd-numbers supervisions. Same as many video super-resolution (VSR) methods [9], [21], we use Vid4 [38] as the evaluation dataset. Besides, we also compare the STVSR results of various networks on SPMC-11 [39] and UDM10 [40] datasets. The two datasets contain more videos of different scenes than Vid4. In detail, SPMC-11 includes eleven videos with a size of 960×536, and UDM10 consists of ten videos with a size of 1272×720. Each video of the two datasets has 31 HR frames.

B. Evaluation Metrics

To compare various STVSR networks quantitatively, we use PSNR and SSIM [47] as evaluation metrics, same as Zooming Slow-Mo [16]. In addition, considering that the calculation

TABLE I
QUANTITATIVE RESULTS OF OUR BI-RSTU NETWORK AND OTHER STVSR METHODS ON THE THREE PUBLIC DATASETS. BESIDES, WE CALCULATE THREE VIDEO QUALITY ASSESSMENT (VQA) METRICS AND THE WHOLE RUNTIME OF DIFFERENT APPROACHES ON THE ENTIRE VID4 [38] DATASET (THE BEST RESULTS ARE SHOWN IN BOLD)

VFI Method	SR Method	Vid4		SPMC-11		UDM10		VQA Metrics			Parameters (Million)	Runtime VFI (s)	Runtime SR (s)	Total Runtime (s)
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	T-MOVIE	MOVIE	VMAF				
SuperSloMo [2]	Bicubic	22.81	0.5754	24.90	0.6866	30.81	0.8731	37.20	9.89	21.12	19.8	32.08	-	-
SuperSloMo [2]	DRN [50]	23.85	0.6395	26.55	0.7529	33.20	0.9012	24.70	6.25	40.07	19.8+9.8	32.08	191.65	223.73
SuperSloMo [2]	EDVR [9]	24.39	0.6691	27.01	0.7688	33.87	0.9076	23.04	5.59	43.55	19.8+20.7	32.08	66.43	96.51
CAIN [33]	Bicubic	23.47	0.6195	25.60	0.7198	31.79	0.8891	31.58	8.31	34.55	42.8	57.26	-	-
CAIN [33]	DRN [50]	24.82	0.7066	27.72	0.8029	34.85	0.9257	19.13	5.20	63.66	42.8+9.8	57.26	191.65	248.91
CAIN [33]	EDVR [9]	25.70	0.7608	28.93	0.8434	35.35	0.9345	13.78	3.39	70.65	42.8+20.7	57.26	64.43	121.69
SepConv [1]	Bicubic	23.52	0.6253	25.66	0.7253	31.80	0.8917	31.25	8.2	34.91	21.7	618.20	-	-
SepConv [1]	DRN [50]	25.10	0.7287	28.25	0.8237	35.20	0.9332	17.17	4.71	67.92	21.7+9.8	618.20	191.65	809.85
SepConv [1]	EDVR [9]	26.10	0.7873	**29.72**	0.8671	35.95	0.9434	11.15	2.71	75.90	21.7+20.7	618.20	64.43	682.63
DAIN [4]	Bicubic	23.55	0.6259	25.67	0.7255	31.88	0.8923	31.21	8.17	35.86	24.0	37.55	-	-
DAIN [4]	DRN [50]	25.13	0.7287	28.24	0.8234	35.37	0.9339	17.01	4.68	68.30	24.0+9.8	37.55	191.65	229.20
DAIN [4]	EDVR [9]	26.11	0.7826	29.62	0.8655	36.13	0.9437	11.94	2.78	75.18	24.0+20.7	37.55	64.43	101.98
STARnet [27]		26.04	0.7821	29.08	0.8502	36.49	0.9454	14.61	3.53	75.89	111.61	-	-	43.35
Zooming Slow-Mo [16]		26.35	0.7973	28.79	0.8627	35.75	0.9491	12.48	2.93	78.46	11.10	-	-	36.97
TMNet [28]		26.43	0.8010	29.47	0.8634	35.94	**0.9500**	12.10	2.84	79.44	12.26	-	-	41.57
Bi-RSTU(Ours)		**26.51**	**0.8063**	29.48	**0.8674**	**36.85**	0.9483	**10.78**	**2.54**	**80.19**	**10.32**	-	-	144.80

process of PSNR and SSIM only involves each reconstructed frame independently, they cannot well measure the spatial-temporal consistency quality of generated videos. Therefore, we use video quality assessment (VQA) metrics: motion-based video integrity evaluation index (MOVIE) and temporal MOVIE (T-MOVIE) [46] to evaluate the reconstructed video consistency quality of different methods. The smaller values of MOVIE and T-MOVIE denote the higher spatial-temporal consistency quality of the video. In addition, through combining multiple elementary quality metrics, Netflix developed a metric called VMAF that could well reflect the human perception of video quality [51]. For this reason, we also introduce VMAF as another VQA metric in our experiments. In detail, the scores of VMAF range from 0 to 100, and a higher value of VMAF means better perceptual video quality.

C. Implementation Details

In our proposed Bi-RSTU network, the feature extraction module is only a 3×3 convolutional layer, and the number of feature channels is set to be 128 in the whole network. The number D of the SOAS blocks in the feature reconstruction module is 10. During the training phase, we augment the training frames by randomly flipping horizontally and 90° rotations. Then, we crop the inputs with a size of 64×64 at random to the network, and the batch size is set to be 16. Our model is trained by Adam [36] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The initial learning rate is 1e-4 and then decreases by ten every 60 epochs. The total number of epochs is set to be 140. We implement the Bi-RSTU network with PyTorch and train our model on 4 NVIDIA GTX-1080Ti GPUs. At the inference stage, we utilize an ensemble strategy similar to [44] to boost our STVSR results.

D. Comparison to State-of-the-art Methods

We compare our Bi-RSTU network with two-stage and one-stage STVSR methods. Specifically, two-stage STVSR methods consist of state-of-the-art (SOTA) time SR and space SR algorithms. Four VFI networks: SepConv [1], SuperSloMo [2], DAIN [4] and CAIN [33] are chosen to perform time SR. As for space SR, we select SOTA single image super-resolution (SISR) approach: DRN [50] and SOTA video super-resolution (VSR) method: EDVR [9]. The recently proposed Zooming Slow-Mo [16], STARnet [27] and TMNet [28] serve as the one-stage methods.

Quantitative comparisons on the three public test datasets: Vid4 [38], SPMC-11 [39] and UDM10 [40] are shown in Table I. We can see that: (1) The combination of VFI and VSR networks achieves better STVSR performance than VFI and SISR. For example, CAIN+EDVR achieves 0.88dB gain compared to CAIN+DRN on the Vid4 dataset, demonstrating that the utilization of information between frames is beneficial for the super-resolution (SR) reconstruction process; (2) The results of two-stage STVSR methods highly depend on the output of VFI networks. Compared to two-stage STVSR methods using the SuperSloMo network, the performance of two-stage STVSR methods with a more advanced VFI algorithm: DAIN is better. In detail, the PSNR value of DAIN+EDVR is 29.62dB on the SPMC-11 dataset. In contrast, the PSNR value of SuperSloMo+EDVR is 27.01dB. The former is 2.61dB higher than the latter; (3) Our Bi-RSTU network surpasses the SOTA one-stage approach: TMNet on the Vid4 and SPMC-11 datasets in terms of PSNR and SSIM. On the UDM10 dataset, Bi-RSTU outperforms TMNet by 0.91dB in terms of PSNR, while both of them have similar SSIM values. The results demonstrate that the auxiliary information from previous time steps is beneficial to the space-time super-resolution (STSR) process at the current time step. Furthermore, our network can reconstruct better HR outputs with the proposed bidirectional RNN structure. In addition, we compare our proposed model with 12 two-stage STVSR methods. Our model achieves the best performance on the Vid4 and UDM10 datasets.

We also investigate the T-MOVIE, MOVIE [46] and VMAF [51] of reconstructed HR videos from various STVSR algorithms on the Vid4 [38] dataset. Meanwhile, each method's model size and runtime are analyzed on the entire Vid4 [38]

Fig. 6. Visual comparisons of various models on Vid4 [38] in our ablation study. The three models: *ML1*, *ML2*, and *ML3*, in a unidirectional manner, generate inaccurate video frame content, such as the pigeon's vibrating wings. Thanks to the bidirectional mechanism, the *ML4* and *ML5* restore the results closer to the ground truth (GT). We can see that the *ML5* produces better image structures than the *ML4*.

TABLE II
ABLATION STUDIES ON THE PROPOSED MODULES IN BI-RSTU. 'UNI' MEANS A UNIDIRECTIONAL MANNER, WHILE 'BI' REFERS TO A BIDIRECTIONAL MANNER. H_{fi} AND H_{rec} DENOTE THE FEATURE INTERPOLATION AND RECONSTRUCTION MODULES, RESPECTIVELY. FOR A FAIR COMPARISON, WE REMOVE THE LAST FRAME OF BIDIRECTIONAL NETWORKS' OUTPUTS FOR EACH VIDEO

Model	Parameters (Million)	H_{fi}			H_{rec}		Vid4 [38]	
		PR	RCA	RECA	PR	SOAS	PSNR	SSIM
ML1(uni)	7.20	✓	✗	✗	✓	✗	24.72	0.7449
ML2(uni)	7.24	✗	✓	✗	✓	✗	24.73	0.7453
ML3(uni)	5.20	✗	✓	✗	✗	✓	24.92	0.7472
ML4(bi)	10.49	✗	✓	✗	✗	✓	26.40	0.8036
ML5(bi)	10.32	✗	✗	✓	✗	✓	26.51	0.8068

dataset. From Table I, we can learn that our Bi-RSTU network achieves the best performance in terms of T-MOVIE, MOVIE and VMAF with the least parameters compared with the other 15 STVSR methods. In detail, compared to the SOTA one-stage method: TMNet, our Bi-RSTU network obtains 0.75 gains in terms of VMAF with fewer parameters, which proves that our network can restore HR video frames with better perceptual quality. Besides, the value of MOVIE about DAIN+EDVR on the Vid4 dataset is 2.78. In contrast, our model with only a quarter of its model size obtains a smaller value in MOVIE, which demonstrates the bidirectional RNN structure used in our approach is beneficial to generating higher video consistency quality.

Qualitative results of various methods are illustrated in Fig. 5. It is observed that DAIN+EDVR acquires the best visual effect among the 6 two-stage STVSR approaches. Compared with DAIN+EDVR and one-stage model: Zooming Slow-Mo, our Bi-RSTU network reconstructs more structures and texture details and achieves better perceptual quality, such as *veni5_015*, Bi-RSTU restores more clear patterns on the clothes. The visual results of Zooming Slow-Mo and Bi-RSTU demonstrate that the information from the previous time step is conducive to the space-time super-resolution (STSR) reconstruction at the current time step. Specifically, the architecture of Zooming Slow-Mo is limited in exploiting the auxiliary information between adjacent time steps. Thanks to the bidirectional RNN structure, our method can fully capture the auxiliary information to assist the STSR process. For this reason, our model recovers better STVSR results with the auxiliary information.

E. Ablation Study

We conduct ablation studies to investigate the effect of our proposed feature interpolation module and reconstruction module. First, we develop a baseline model called *ML1*. As shown in Table II, the *ML1*, constructed in a unidirectional manner, adopts the plain residual (PR) blocks [41] as basic blocks of the feature interpolation module H_{fi} and reconstruction module H_{rec}. Compared to the baseline model, the *ML2* employs residual channel attention (RCA) blocks [5] to constitute H_{fi} for feature interpolation and achieves an improvement of 0.01dB. With our developed second-order attention spindle (SOAS) blocks, the *ML3* outperforms the *ML2* by 0.19dB on Vid4, which demonstrates the effectiveness of SOAS blocks in feature reconstruction. In addition, we can see that the *ML4* follows the same setting as the *ML3* but in a bidirectional manner. Surprisingly, the *ML4* obtains a 1.48dB performance gain compared to the *ML3*, which confirms that the network, in a bidirectional manner, can excavate more LR and LFR video information to reconstruct the corresponding HR and HFR video. The *ML5* adopts residual efficient channel attention (RECA) blocks to constitute the interpolation module, and it achieves 0.11dB gain compared to the *ML4*, which proves that the RECA blocks are better than RCA blocks for feature interpolation. In addition, we also list the parameters of various models in Table II for a more detailed comparison.

The visual comparisons of the five models are illustrated in Fig. 6. Note that the *ML1*, *ML2* and *ML3* only utilize the unidirectional information, which leads to inaccurate content of the reconstructed intermediate frame. In contrast, the *ML4* and the *ML5* with a bidirectional manner can synthesize the intermediate frame close to the ground truth. With the RECA blocks in the feature interpolation module, the *ML5* restores more image details than the *ML4*, which demonstrates that RECA blocks used in our network are more suitable for the STVSR task than RCA blocks.

V. Future Work

In this work, we only use the conventional RNN structure to establish the Bi-RSTU network for STVSR. Although our Bi-RSTU can effectively utilize the auxiliary information from various time steps, it may demonstrate unsatisfactory performance in dealing with a very long sequence of LR frames in a real-world application due to the limited temporal correction abilities of the conventional RNN. Thanks to the special gate units mechanism, the LSTM [26] performs better than conventional RNN in handling long video sequences. Therefore, it is worthwhile to construct a new STVSR framework using the LSTM for processing long video sequences.

In Bi-RSTU, we adopt RECA blocks to perform feature interpolation. Advanced modules like deformable attention block [29] for dynamically leveraging two neighboring frame features can be considered to enhance the synthesis of their intermediate frame feature.

VI. Conclusion

In this paper, we propose a one-stage network called Bi-RSTU for space-time video super-resolution (STVSR) based on a bidirectional RNN structure. Unlike previous one-stage STVSR methods that cannot effectively explore the auxiliary information among various time steps, our Bi-RSTU fully exploits the auxiliary information to assist the STVSR process at the current time step. Specifically, we devise the efficient channel attention feature interpolation module to synthesize the intermediate frame feature. The feature reconstruction module composed of second-order attention spindle modules can map the low-resolution features fused information from the previous time step to high-resolution features. Experimental results show that our method acquires competitive performance compared with two-stage and one-stage state-of-the-art STVSR methods.

VII. Acknowledgment

We would like to thank the anonymous reviewers for their constructive suggestions, which have significantly improved our manuscript.

References

[1] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, pp. 261–270. 1, 2, 7

[2] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 9000–9008. 1, 7

[3] 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 transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 933–948, 2021. 1, 4

[4] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019, pp. 3698–3707. 3, 7

[5] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich, Germany: Springer, 2018, pp. 294–310. 1, 2, 3, 5, 8

[6] X. Yang, H. Mei, J. Zhang, K. Xu, B. Yin, Q. Zhang, and X. W, "DRFN: Deep Recurrent Fusion Network for Single-Image Super-Resolution With Large Factors," in *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 328–337, 2019. 2

[7] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Long Beach, CA, USA: IEEE, 2019, pp. 11 057–11 066. 1

[8] Y. Jo, S. Wug Oh, J. Kang, and S. Joo Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 3224–3232. 1, 2

[9] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Long Beach, CA, USA: IEEE, 2019, pp. 1954–1963. 1, 2, 6, 7

[10] E. Shechtman, Y. Caspi, and M. Irani, "Increasing space-time resolution in video," in *European Conference on Computer Vision*. Copenhagen, Denmark: Springer, 2002, pp. 753–768. 1, 3

[11] O. Shahar, A. Faktor, and M. Irani, *Space-time super-resolution from a single video*. Colorado Springs, CO, USA: IEEE, 2011. 1, 3

[12] T. Li, X. He, Q. Teng, Z. Wang, and C. Ren, "Space–time super-resolution with patch group cuts prior," *Signal Processing: Image Communication*, vol. 30, pp. 147–165, 2015. 1, 3

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015. 2

[14] C. Ding and D. Tao, "Robust Face Recognition via Multimodal Deep Face Representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015. 2

[15] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive Contexts for Object Detection," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 944–954, 2017. 2

[16] X. Xiang, Y. Tian, Y. Zhang, Y. Fu, J. P. Allebach, and C. Xu, "Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020, pp. 3367–3376. 2, 3, 6, 7

[17] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Las Vegas, NV, USA: IEEE, 2016, pp. 1646–1654. 2

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, HI, USA: IEEE, 2017, pp. 105–114. 2

[19] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proceedings of the IEEE International Conference on Computer Vision*. Venice, Italy: IEEE, 2017, pp. 4809–4817. 2

[20] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019. 2, 6

[21] L. Wang, Y. Guo, Z. Lin, X. Deng, and W. An, "Learning for video super-resolution through hr optical flow estimation," in *Asian Conference on Computer Vision*. Perth, Australia: Springer, 2018, pp. 514–529. 2, 6

[22] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA: IEEE, 2019, pp. 3892–3901. 2

[23] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020, pp. 3357–3366. 2

[24] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773. 2, 3

[25] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316. 2, 3

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 9

[27] M. Haris, G. Shakhnarovich, and N. Ukita, "Space-time-aware multi-resolution video enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2859–2868. 3, 7

[28] G. Xu, J. Xu, Z. Li, L. Wang, X. Sun, and M.-M. Cheng, "Temporal modulation network for controllable space-time video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6388–6397. 3, 7

[29] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision transformer with deformable attention," *arXiv preprint arXiv:2201.00520*, 2022. 9

[30] 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 *Proceedings of the IEEE International Conference on Computer Vision*. Seoul, Korea (South): IEEE, 2019, pp. 3106–3115. 2

[31] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive convolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu, HI, USA: IEEE, 2017, pp. 2270–2279.

[32] S. Niklaus and F. Liu, "Context-aware synthesis for video frame interpolation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 1701–1710. 3

[33] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *AAAI*. New York City, NY, USA: AAAI Press, 2020, pp. 10 663–10 671. 3, 4, 7

[34] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, WA, USA: IEEE, 2020, pp. 11 534–11 542. 4, 5

[35] W. Yang, W. Wang, X. Zhang, S. Sun, and Q. Liao, "Lightweight feature fusion network for single image super-resolution," *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 538–542, 2019. 5

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 7

[37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 7132–7141.

[38] C. Liu and D. Sun, "On bayesian adaptive video super resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 346–360, 2013. 6, 7, 8

[39] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *2018 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy: IEEE, 2017, pp. 4482–4490. 6, 7

[40] P. Li, Z. Wang, K. Jiang, J. Jiang, and J. Ma, "Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea(South): IEEE, 2019, pp. 3106–3115. 6, 7

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 770–778. 8

[42] Y. Chen, Y. Chen, J.-H. Xue, W. Yang, and Q. Liao, "Lightweight single image super-resolution through efficient second-order attention spindle network," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. London, UK: IEEE, 2020, pp. 1–6. 5

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *2015 IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, 2015, pp. 1026–1034. 4, 5

[44] R. Timofte, R. Rothe, and L. V. Gool, "Seven ways to improve example-based single image super resolution," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 1865–1873. 7

[45] Y. Huang, W. Wang, and L. Wang, "Video super-resolution via bidirectional recurrent convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 1015–1028, 2018. 2, 3

[46] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010. 7

[47] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 6

[48] T. Isobe, F. Zhu, and S. Wang, "Revisiting temporal modeling for video super-resolution," in *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. Virtual Event, UK: BMVA Press, 2020. 2

[49] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016, pp. 1874–1883. 3

[50] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, "Closed-loop matters: Dual regression networks for single image super-resolution," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, pp. 5406–5415. 2, 7

[51] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward A Practical Perceptual Video Quality Metric," in *Netflix TechBlog*, June, 2016. 7