Store and Fetch Immediately: Everything Is All You Need for Space-Time Video Super-resolution

Mengshun Hu[1]*, Kui Jiang[2]*, Zhixiang Nie[1], Jiahuan Zhou[3], Zheng Wang[1]†

[1]School of Computer Science, Wuhan University
[2] Huawei Technologies, Cloud BU
[3]Wangxuan Institute of Computer Technology, Peking University
{shunmh, niezhixiang, wangzwhu}@whu.edu.cn, jiangkui5@huawei.com, jiahuanzhou@pku.edu.cn

Abstract

Existing space-time video super-resolution (ST-VSR) methods fail to achieve high-quality reconstruction since they fail to fully explore the spatial-temporal correlations, long-range components in particular. Although the recurrent structure for ST-VSR adopts bidirectional propagation to aggregate information from the entire video, collecting the temporal information between the past and future via one-stage representations inevitably loses the long-range relations. To alleviate the limitation, this paper proposes an immediate store-and-fetch network to promote long-range correlation learning, where the stored information from the past and future can be refetched to help the representation of the current frame. Specifically, the proposed network consists of two modules: a backward recurrent module (BRM) and a forward recurrent module (FRM). The former first performs backward inference from future to past, while storing future super-resolution (SR) information for each frame. Following that, the latter performs forward inference from past to future to super-resolve all frames, while storing past SR information for each frame. Since FRM inherits SR information from BRM, therefore, spatial and temporal information from the entire video sequence is immediately stored and fetched, which allows drastic improvement for ST-VSR. Extensive experiments both on ST-VSR and space video super-resolution (S-VSR) as well as time video super-resolution (T-VSR) have demonstrated the effectiveness of our proposed method over other state-of-the-art methods on public datasets.

Introduction

Space-time video super-resolution (ST-VSR), aiming to generate the high-resolution (HR) and high-frame-rate (HFR) photo-realistic video sequences from the given low resolution (LR) and low-frame-rate (LFR) inputs, gradually becomes the research hotspot in computer vision and machine learning communities (Kim, Oh, and Kim 2020). In practice, ST-VSR technologies are widely applied to the movie production (Xu et al. 2021; Haris, Shakhnarovich, and Ukita 2020), high-definition television upgrades (Kang et al. 2020) and video compression (Xiang et al. 2020), *etc.*

Figure 1: input frames from Vid4 (Liu and Sun 2011) have different characteristics for supplement each other.

To tackle ST-VSR task, as illustrated in Figure 2, the advanced ST-VSR methods are roughly divided into three categories: two-stage, one-stage and compact one-stage based methods. The first category tries to decompose the ST-VSR task into two sub-tasks: space video super-resolution (S-VSR) and time video super-resolution (T-VSR), which independently and sequentially performs on LR and LFR videos, to increase the spatial resolution (Zhang et al. 2018; Haris, Shakhnarovich, and Ukita 2018, 2019; Wang et al. 2019) and frame rates (Jiang et al. 2018; Niklaus, Mai, and Liu 2017; Bao et al. 2019; Choi et al. 2020) on image space and vice versa. However, these methods barely consider the inherent spatial-temporal correlations (Haris, Shakhnarovich, and Ukita 2020; Hu et al. 2022a) due to the individual processing of these two sub-tasks. Consequently, they are far from producing satisfactory results, shown in Figure 2(a).

To alleviate this issue, as shown in Figure 2(b), one-stage based methods explore spatial-temporal correlations by integrating S-VSR and T-VSR tasks into a unified framework for the joint optimization (Xiang et al. 2020; Xu et al. 2021). However, researchers in this line still separate out S-VSR and T-VSR tasks on feature space, which fails to fully utilize spatial-temporal correlations of these two sub-tasks for ST-VSR (Hu et al. 2022a; Zhou et al. 2021). Unlike the one-stage based methods, compact one-stage based methods shown in Figure 2(c), propose to simultaneously increase

*These authors contributed equally.
†Corresponding author.

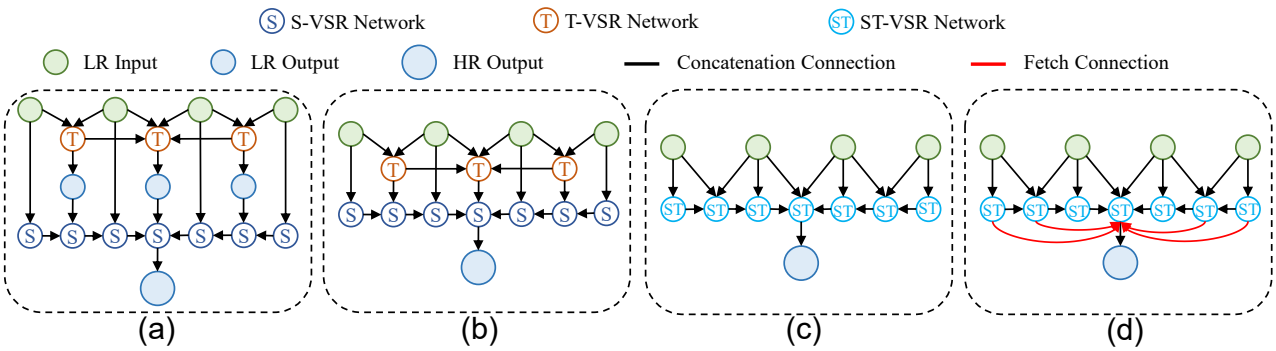

Figure 2: Different schemes for ST-VSR. (a): Two-stage based methods: They perform ST-VSR by independently and sequentially using advanced S-VSR and T-VSR on image space. (b) One-stage based methods: They unify S-VSR and T-VSR into a single stage for ST-VSR on feature space. (c) Compact one-stage based methods: They directly employ ST-VSR network to explore spatial-temporal correlations for ST-VSR on feature space. (d) Our proposed method: Our method proposes ST-VSR network with immediate store-and-fetch property to sufficiently utilize long-range spatial-temporal correlations on feature space from the entire sequences for ST-VSR.

the spatial and temporal resolution via a recurrent framework with a holistic design. Specifically, it performs spatial-temporal feature aggregation to endow the model with the ability to optimize temporal consistency and spatial texture details each other (Hu et al. 2022b). Unfortunately, the above methods almost focus on short-range spatial-temporal correlations, while resorting to the hidden states from adjacent frames with the one-stage representations scheme to collect the temporal information from the past and the future. However, some complementary spatial and temporal information from distant frames also matter for ST-VSR under large motion and occlusion scenarios, which are ignored in these methods.

In this paper, the intuition is that long-range spatial-temporal correlations are provided for supplement each other, where occlusion or large motion regions from adjacent frames for interpolation and super-resolution would probably be compensated in a region from other frames. As shown in Figure 1, the face from frame 40 is obscured in the adjacent frames (frame 39, frame 41, frame 42), but reappears in the distant frames (frame 43, frame 44, frame 45). In this way, we propose to further optimize the compact ST-VSR framework to fully utilize long-range spatial-temporal correlations via immediate store-and-fetch strategy. Specifically, we design a forward recurrent module (FRM) and a backward recurrent module (BRM), which are interactive to fetch past, current and future SR information for super-resolving current frames, while storing all updated SR information. Therefore, as shown in Figure 2(d), the proposed network can aggregate and fuse spatial and temporal information from all frames in the video sequence. Technically, FRM and BRM share similar structures, involving an adjacent fetch block (AF) and a distant fetch block (DF) to aggregate adjacent and distant spatial information, respectively. In addition, considering the aggregated spatial information from different frames mixes irrelevant or redundant components, instead of directly fusing spatial information from different frames for temporal aggregation (Xu et al. 2021; Chan et al. 2021), we further adopt a dynamic selection block (DS) to adaptively guide aggregated spatial information for temporal aggregation. Experiments on public datasets demonstrate that our method produces higher quality videos than the state-of-the-art methods in ST-VSR, S-VSR and T-VSR tasks.

Our contributions are summarized as follows:

- We propose a novel store-and-fetch framework for ST-VSR, where long-range spatial-temporal correlations from all frames in the video sequence are fully mined.
- We devise an immediate store-and-fetch scheme to fetch spatial and temporal SR information from past, current and future frames for the super-resolving current frame, while storing the current updated SR information.
- We conduct extensive experiments to compare our network on ST-VSR, S-VSR and T-VSR tasks, which demonstrates our method performs well against the state-of-the-art methods on public datasets.

Related Work

Space-Time Video Super-Resolution

ST-VSR aims to increase the spatial and temporal resolution of video sequences (Kim, Oh, and Kim 2020; Geng et al. 2022; Wang et al. 2022). The key challenge lies in sufficiently utilizing spatial-temporal information from video sequences. The traditional methods (Shechtman, Caspi, and Irani 2005) attempt to adopt hand-crafted regularization, prior knowledge and specific assumptions to optimize the model for ST-VSR, but these constraints limit the model to build spatial-temporal correlations on complex patterns. Recently, some studies (Haris, Shakhnarovich, and Ukita 2020; Zhou et al. 2021) propose to mutually learn S-VSR and T-VSR for ST-VSR. Xiang *et al.* propose to explore local temporal contexts for features interpolation, and then exploit global temporal contexts for super-resolution (Xiang et al. 2020). Inspired by (Xiang et al. 2020), Xu *et al.* further utilize the locally temporal feature comparison module to

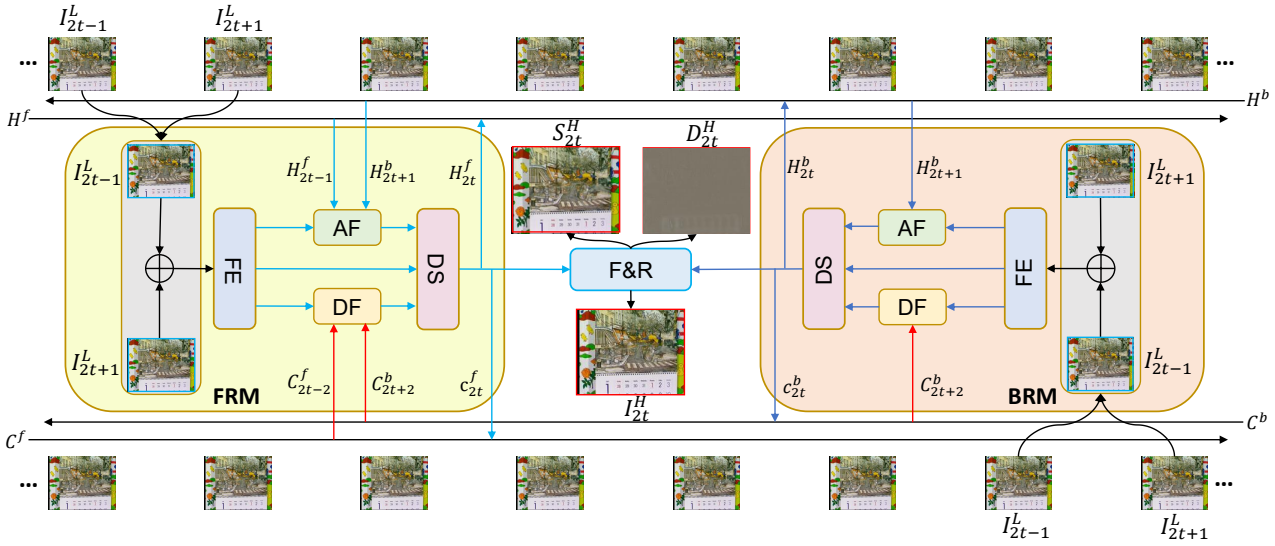

Figure 3: Architecture of the proposed our network. Given a video sequence with low-resolution (LR) and low-frame-rate (LFR), we adopt a backward recurrent module (BRM) and a forward recurrent module (FRM) with bidirectional immediate store-and-fetch containers (C^b,C^f) to align and aggregate the past, current and future representations. Then we assign backward and forward inferences to learn structures and details components from temporal information, and progressively fuse and reconstruct (F&R) the final high-resolution (HR) ($\times 4$) and high-frame-rate (HFR) ($\times 2$) videos, structures and details.

explore local motion cues for refinement (Xu et al. 2021). However, the above methods fail to model the accurate and complete spatial-temporal correlations among the video sequences.

Store-and-Fetch Network.

Store-and-Fetch network aims to store the potentially informative components for further fusion and refinement (Ji and Yao 2022), which has been widely used in natural language processing (Sukhbaatar et al. 2015) and video object segmentation (Oh et al. 2019). From their wisdom, we devise an immediate store-and-fetch scheme to fully explore the spatial-temporal correlations. Specifically, the past, present and future SR features are distilled and stored during the bidirectional interactive propagation process, which can be immediately fetched for a comprehensive fusion to promote the current representation.

Proposed Approach

Overview of Network Architecture

Figure 3 details the overall network architecture. Given the LR and LFR inputs $[I^L_{2t-1}]^{n+1}_{t=1}$, our proposed method aims to reconstruct HR and HFR video sequence $[I^H_t]^{2n+1}_{t=1}$. Technically, the shared feature extractor (FE) is firstly introduced to project the input frames to feature space to obtain the initial representations ($[F^L_t]^{2n+1}_{t=1}$), expressed as:

$$\begin{aligned} F^L_{2t-1} &= FE(I^L_{2t-1}) \quad t = 1, 2, ..., n+1, \\ F^L_{2t} &= FE(\frac{1}{2}(I^L_{2t-1} + I^L_{2t+1})) \quad t = 1, 2, ..., n. \end{aligned} \tag{1}$$

Following that, the forward recurrent module (FRM) and backward recurrent module (BRM) with bidirectional store-and-fetch containers (C^f, C^b) are used to aggregate the spatial-temporal information from all frames in video sequences. Concretely, we first perform backward inference to learn temporal correlations from the future to past frame while storing the updated hidden state via backward store-and-fetch container, defined as:

$$\begin{aligned} H^b_{2t}, c^b_{2t} =& DS(AF(F^L_{2t}, H^b_{2t+1}), F^L_{2t}, DF(F^L_{2t}, C^b_{2t+2})), \\ C^b_{2t} =& U(C^b_{2t+1}, c^b_{2t}), \end{aligned} \tag{2}$$

where $AF(\cdot)$ and $DF(\cdot)$ denote the fetch blocks, whose primary duty is to extract the correlated components from adjacent frames and distant frames, respectively. $DS(\cdot)$ refers to the dynamic selection block to distill the informative features for refinement. H^b and c^b are the hidden states generated by BRM, while C^b represents the backward store-and-fetch container. The main role of $U(\cdot)$ is to store the hidden states c^b_{2t} to the container C^b_{2t+1} for updating.

The procedure of forward recurrent module (FRM) is similar to BRM, which can be described as:

$$\begin{aligned} H^f_{2t}, c^f_{2t} =& DS(AF(F^L_{2t}, H^b_{2t+1}, H^f_{2t-1}), F^L_{2t}, \\ & DF(F^L_{2t}, C^b_{2t+2}, C^f_{2t-2})), \\ C^f_{2t} =& U(C^f_{2t-1}, c^f_{2t}), \end{aligned} \tag{3}$$

where H^f_{2t} and c^f_{2t} are the hidden states generated by FRM. C^f represents the forward store-and-fetch container for storing the past hidden states. Since FRM inherits the hidden states from BRM, and thus it can fully utilize the spatial-

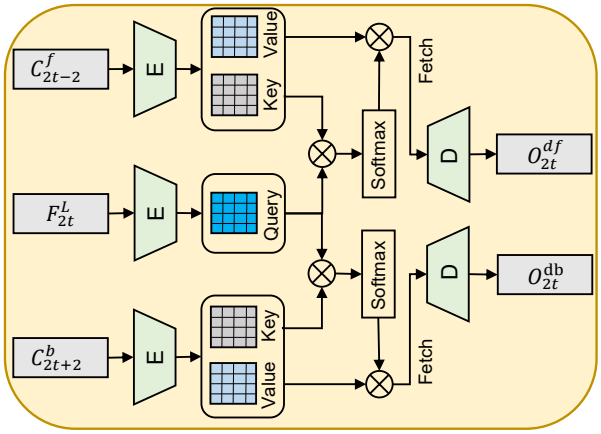

Figure 4: Architecture of the proposed distant fetch block (DF).

temporal information from all frames. Note that the initialization of bidirectional store-and-fetch containers and adjacent hidden states is empty array and zero, respectively. To reconstruct video sequences with more textures and details, inspired by (Yi et al. 2021; Hu et al. 2022b), the progressive fusion and reconstruction (F&R) is introduced to produce the final HR and HFR videos. More specifically, the structures and details are respectively refined via the backward and forward inferences in a progressive manner.

Adjacent Fetch Block

To effectively explore spatial information from adjacent frames (*e.g.,* H^f_{2t-1} and H^b_{2t+1}), inspired by (Xiang et al. 2020), we propose an adjacent fetch block (AF). It can be achieved by a temporal-deformable network (Tian et al. 2020), to implicitly aggregate informative components from adjacent frames. Taking FRM as an example, technically, we feed the current feature F^L_{2t} and hidden states H^f_{2t-1} and H^b_{2t+1} from adjacent frames into AF, and obtain aligned hidden states H^{af}_{2t} and H^{ab}_{2t} by a series of motion offsets prediction and deformable convolution operations (Please refer to more details in the supplementary material).

Distant Fetch Block

AF has shown its effectiveness in capturing inter-frame motions between adjacent frames (Xu et al. 2021). However, since there exist large motions between hidden states from bidirectional containers and the super-resolved feature, the training of deformable alignment of AF becomes unstable, and the overflow of motion offsets severely degrades the model performance (Wang et al. 2019). To sufficiently mine these distant spatial information, we propose a distant fetch block (DF) to fetch and gather hidden states from bidirectional containers. The principle is that the stored hidden states and the super-resolved feature are served as the key-value format and query components for representation. In this way, we compare query features with hidden states in the key space and retrieve the associated values. During the

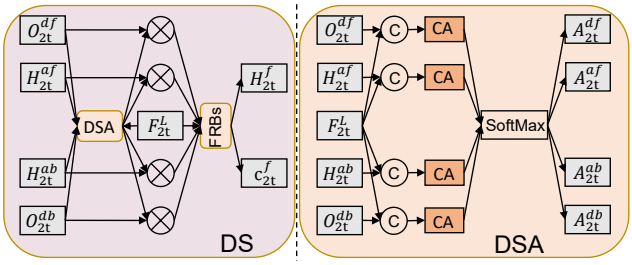

Figure 5: Architecture of the proposed dynamic selection block (DS) and dynamic selection attention (DSA).

Encoding (E) in Figure 4, to handle large motions, we first downsample all stored hidden states from bidirectional containers and the current feature to obtain the key, value and query at 1/8 resolution by the encoder, expressed as:

$$K^f_{2t-2}, K^b_{2t+2}, V^f_{2t-2}, V^b_{2t+2}, Q^L_{2t} = E(C(C^f_{2t-2}), C(C^b_{2t+2}), F^L_{2t}) \tag{4}$$

where $E(\cdot)$ denotes the encoder and $C(\cdot)$ refers to the channel concatenation. Fetch: We compute the affinity matrix between keys and query for comparison, which can be used to weight summation for retrieving the associated values and as follows:

$$\begin{aligned} O^f_{2t} &= V^f_{2t-2} \cdot Softmax(K^f_{2t-2} \cdot Q^L_{2t}/\delta), \\ O^b_{2t} &= V^f_{2t+2} \cdot Softmax(K^b_{2t+2} \cdot Q^L_{2t}/\delta), \end{aligned} \tag{5}$$

where δ is a learnable scaling parameter to control the magnitude of the dot product of the key and query. O^f_{2t} and O^b_{2t} are the corresponding outputs. During the Decoding (D) in Figure 4, the fetched components (O^f_{2t} and O^b_{2t}) from forward and backward containers are decoded to the original resolution:

$$O^{df}_{2t}, O^{db}_{2t} = D(O^f_{2t}, O^b_{2t}), \tag{6}$$

where $D(\cdot)$ denotes the decoder.

Dynamic Selection Block

Since the importance of the spatial information fetched by AF and DF varies with different frames, the direct fusion consequently causes the feature to be redundant and confused. Thus, we design a dynamic selective block (DS) to guide the fusion process. As shown in Figure 5, given the aligned hidden states from adjacent frames and distant frames H^{af}_{2t}, H^{ab}_{2t}, O^{df}_{2t} and O^{db}_{2t}, and the current feature F^L_{2t}, we separately feed the feature pair $\{H^{af}_{2t}, F^L_{2t}\}$, $\{H^{ab}_{2t}, F^L_{2t}\}$, $\{O^{df}_{2t}, F^L_{2t}\}$ and $\{O^{db}_{2t}, F^L_{2t}\}$ into dynamic selection attention block (DSA). Through evaluating the importance of different hidden states, DSA can obtain the corresponding attention map A^{af}_{2t}, A^{ab}_{2t}, A^{df}_{2t} and A^{db}_{2t} for selecting the required features via cross attention (CA) and Softmax operations. Then, the distilled features and current feature are packed into a series of fusion residual blocks (FRBs) (Yi et al. 2019) for obtaining present hidden states H^f_{2t} and c^f_{2t}, which further explores intra-frame spatial correlations and inter-frame temporal correlations for further redundant optimization.

VFI Method	VSR Method	Vid4 PSNR(Y)↑	Vid4 SSIM(Y)↑	Fast PSNR(Y)↑	Fast SSIM(Y)↑	Medium PSNR(Y)↑	Medium SSIM(Y)↑	Slow PSNR(Y)↑	Slow SSIM(Y)↑	Parameters (Million)
SuperSloMo	Bicubic	22.84	0.577	31.88	0.879	29.94	0.848	28.37	0.810	19.8
SuperSloMo	RCAN	23.80	0.640	34.52	0.908	32.50	0.888	30.69	0.862	19.8+16.0
SuperSloMo	RBPN	23.76	0.636	34.73	0.911	32.79	0.893	30.48	0.858	19.8+12.7
SuperSloMo	EDVR	24.40	0.677	35.05	0.914	33.85	0.897	30.99	0.867	19.8+20.7
SepConv	Bicubic	23.51	0.627	32.27	0.889	30.61	0.863	29.04	0.829	21.7
SepConv	RCAN	24.92	0.724	34.97	0.920	33.59	0.913	32.13	0.897	21.7+16.0
SepConv	RBPN	26.08	0.775	35.07	0.924	34.09	0.923	32.77	0.909	21.7+12.7
SepConv	EDVR	25.93	0.779	35.23	0.925	34.22	0.924	32.96	0.911	21.7+20.7
DAIN	Bicubic	23.55	0.627	32.41	0.891	30.67	0.864	29.06	0.829	24.0
DAIN	RCAN	25.03	0.726	35.27	0.924	33.82	0.915	32.26	0.897	24.0+16.0
DAIN	RBPN	25.96	0.778	35.55	0.930	34.45	0.926	32.92	0.910	24.0+12.7
DAIN	EDVR	26.12	0.784	35.81	0.932	34.76	0.928	33.11	0.912	24.0+20.7
STARnet		26.06	0.805	36.19	0.937	34.86	0.936	33.10	0.916	111.6
Zooming SlowMo		26.31	0.798	36.81	0.942	35.41	0.936	33.36	0.914	11.1
TMNet		26.43	0.802	37.04	0.944	35.60	0.938	33.51	0.916	12.3
RSTT		26.43	0.799	36.80	0.940	35.66	0.938	33.50	0.915	7.7
YOGO		26.34	0.802	36.93	0.942	35.55	0.937	33.44	0.915	12.1
Ours		26.43	0.811	37.12	0.946	35.63	0.940	33.49	0.918	16.4

Table 1: Quantitative comparisons of our results and the SOTA ST-VSR methods on Vid4 and Vimeo90K (Seven) datasets. Note we input four LR image with the resolution of 112 × 64 to test Times and FLOPs on a RTX 3090Ti GPU.

Model	Bicubic	RCAN	TOFlow	DUF	RBPN	EDVR-L	BasicVSR	IconVSR	Ours
PSNR(Y)↑	31.32	35.35	34.83	36.37	37.07	37.61	37.18	37.47	37.22
SSIM(Y)↑	0.8684	0.925	0.9220	0.939	0.9435	0.949	0.9450	0.948	0.947
Parameters(Million)↓	—	16.0	1.4	5.8	12.7	20.6	6.3	8.7	14.1

Table 2: Quantitative comparisons of our results and the SOTA S-VSR methods on Vimeo90K (Seven) dataset, which is tested with 4× downsampling using Bicubic (BI).

Model	SepConv	TOFlow	CyclicGen	DAIN	CAIN	BMBC	AdaCoF	EDSC	XVFI	Ours
PSNR↑	33.79	33.73	32.09	34.71	34.65	35.01	34.47	34.84	35.07	35.23
SSIM↑	0.970	0.968	0.949	0.976	0.973	0.976	0.973	0.975	0.976	0.977
Parameters(Million)↓	21.7	1.4	19.8	24.0	42.8	11.0	21.8	9.8	5.5	10.1

Table 3: Quantitative comparisons of our results and the SOTA T-VSR methods on Vimeo90K (Triplets) dataset.

Fusion and Reconstruction Network

Following the BRM and FRM, the backward inference H_{2t}^{b} and forward inference H_{2t}^{f} are respectively assigned to learn structures and details components via a hybrid fusion module (Hu et al. 2022b). Then, a reconstruction module, which consists of two pixel-shuffle layers (Shi et al. 2016) and a sequence of "Conv-LeakyReLU-Conv" operations, is designed to produce the corresponding HR (4×) and HFR (2×) videos, structures and details. To optimize the whole network, we use a reconstruction loss function, expressed as:

$$L_r = \sum_{t=1}^{2n+1} (\rho(I_t^H - I_t^{GT}) + \rho(D_t^H - D_t^{GT}) + \rho(S_t^H - S_t^{GT})), \tag{7}$$

where I_t^{GT}, D_t^{GT} and S_t^{GT} denote the corresponding ground-truth video frames, details and structures components, where the detail components are the residuals between the bicubic sampling (the structural components) and the ground-truth video frames I_t^{GT}. $\rho = \sqrt{(x^2 + w^2)}$ is the Charbonnier penalty function with the constant w set to 10^{-3} (Charbonnier et al. 1994).

Experiments and Analysis

Datasets and Metrics

The Vimeo90K trainset (Xue et al. 2019) is used to train our network for a fair comparison of other methods (Xu et al. 2021; Xiang et al. 2020), which consists of more than 60,000 7-frame training video sequences with the resolution of 448×256. Vid4 (Liu and Sun 2011) and Vimeo90K testsets are used to evaluate ST-VSR methods, where Vimeo90K testsets are divided into Vimeo90K-Fast, Vimeo90K-Medium and Vimeo90K-Slow subsets for testing according to the degree of motion. In addition, we adopt Peak Signal-to-Noise (PSNR) and Structural Similarity Index (SSIM) (Wang et al. 2004) to quantitatively compare different ST-VSR methods.

Implementation Details

We take out the odd-indexed 4 frames as LR and LFR inputs, and the corresponding consecutive HR 7-frame for supervision. Specifically, we train our network using Pytorch 1.9 with four NVIDIA Tesla V100 by AdaMax optimizer (Kingma and Ba 2014) for 600,000 iterations with batch size 24. The learning rate is initially set to 1e-3, and gradually decays to 1e-7 following a cosine attenuation

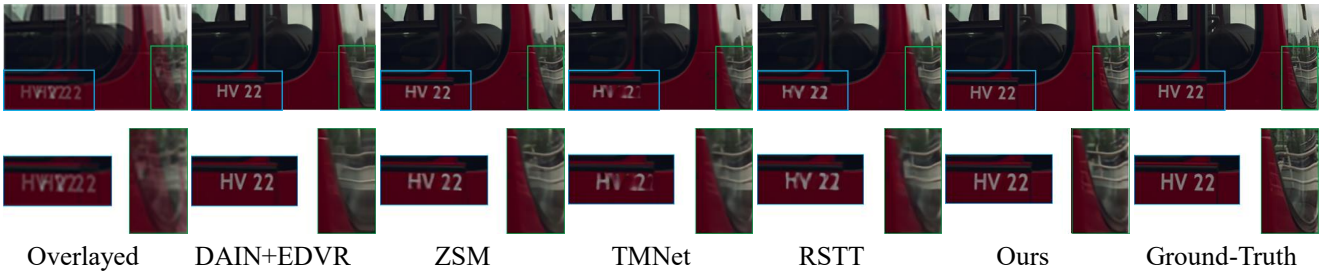

Figure 6: Visual comparisons with state-of-the-art ST-VSR methods on the Vimeo90K-Fast dataset.

schedule. During training, the images from Vimeo90K trainset are randomly rotated and flipped, and randomly cropped into small patches with a resolution of 32×32 for training and 64×64 for finetune.

Comparison with State-of-the-Art Methods

ST-VSR. We compare our framework with SOTA (compact) one-stage based ST-VSR methods, including STARnet (Haris, Shakhnarovich, and Ukita 2020), Zooming SlowMo (ZSM) (Xiang et al. 2020), TMNet (Xu et al. 2021), RSTT (Geng et al. 2022) and YOGO (Hu et al. 2022b). In addition, we also compare two-stage based methods, which perform T-VSR by SuperSloMo (Jiang et al. 2018), SepConv (Niklaus, Mai, and Liu 2017) and DAIN (Bao et al. 2019), then perform S-VSR by Bicubic interpolation, RCAN (Zhang et al. 2018), RBPN (Haris, Shakhnarovich, and Ukita 2019) and EDVR (Wang et al. 2019).

Quantitative Results. As shown in Table 1, most of (compact) one-stage based methods show significant superiority than the two-stage based methods. The main reason, as we have analyzed, lies that the two-stage based methods fail to utilize the inherent spatial-temporal correlations. Moreover, compact one-stage based methods (RSTT and YOGO) have fewer parameters but perform almost as well as one-stage based methods (Zooming SlowMo and TMNet). This also proves the effectiveness and efficiency of simultaneously learning S-VSR and T-VSR tasks. In addition, our method aims to further optimize the compact one-stage based methods and achieves the SOTA performance on all datasets. This is attributed to the fact that long-range spatial-temporal correlations from video sequences can help model the temporal motion flow of ST-VSR tasks, large motion and occlusion scenarios in particular (Vimeo90K-Fast).

Qualitative Results. Figure 6 further provides the visual comparison of ST-VSR. We can see that the results produced by the two-stage based methods suffer more from temporal artifacts degradation due to the rare exploration spatial-temporal correlations between two tasks (See the front view mirror). Although one-stage based methods try to explore the unilateral correlations, and utilize more temporal information to reconstruct spatial information by sequentially performing T-VSR and T-VSR on feature space, the generated result is still ambiguous (See the text message). On the contrary, the compact one-stage based methods simultaneously learn T-VSR and S-VSR to implicitly explore mutual correlations, and gain significant improvement of reconstruction performance. However, due to the only dependence on short-range spatial-temporal correlations, the super-resolved results fail to infer rich details on the motion regions. Compared to the above methods, our proposed method can sufficiently mine long-range spatial-temporal correlations from all frames, producing more natural and realistic results.

S-VSR. We retrain our proposed method on Vimeo90K (Seven) dataset, and compare it with the advanced S-VSR methods, including RCAN (Zhang et al. 2018), TOFlow (Xue et al. 2019), RBPN (Haris, Shakhnarovich, and Ukita 2019), EDVR (Wang et al. 2019), DUF (Jo et al. 2018) and BasicVSR/IconVSR (Chan et al. 2021),

Quantitative results on Vimeo90K (Seven) dataset are shown in Table 2. It is obvious that our method surpasses S-VSR methods based sliding-window framework TOFlow (Xue et al. 2019) and DUF (Jo et al. 2018), since these methods only explore local temporal correlations among video sequences. Moreover, our proposed approach also outperforms some S-VSR methods based recurrent framework RBPN (Haris, Shakhnarovich, and Ukita 2019) and BasicVSR (Chan et al. 2021) by 0.15dB and 0.04dB in terms of PSNR. The main reason is that these methods fail to fully explore the spatial-temporal complementary information while compressing all past and future information into one representation. By contrast, our method can reconstruct better results since all spatial and temporal information is aggregated and fused to sufficiently explore spatial-temporal correlations. In addition, we can see that our method has no significant advantages over the advanced sliding window-based method EDVR (Wang et al. 2019) and recurrent method IconVSR (Chan et al. 2021), The reason is that for S-VSR task, the inputs are a short sequence of 7 frames on Vimeo90K dataset, which may cut down on our ability to build long-range temporal correlations with store-and-fetch strategies. Therefore, short video processing tasks are not suitable for our proposed framework and do not match the real scenrio assumptions.

T-VSR. We retrain our method on Vimeo90K (Triplets) dataset to interpolate single frame according to two consecutive input frames, and compare it with advanced T-VSR methods, including SepConv (Niklaus, Mai, and Liu 2017), SuperSloMo (Jiang et al. 2018), TOFlow (Xue et al.

Setting		(a)	(b)	(c)	(d)	(e)
SA	AF	✗	✓	✗	✓	✓
	DF	✗	✗	✓	✓	✓
TA	DS	✗	✗	✗	✗	✓
Fast		36.20	36.50	36.34	36.64	36.86
Medium		35.05	35.18	35.06	35.25	35.47
Slow		33.16	33.27	33.19	33.30	33.43

Table 4: Quantitative comparisons in PSNR from Vimeo90K datasets on different models. SA denotes spatial aggregation and TA denotes temporal aggregation.

2019), CyclicGen (Liu et al. 2019), DAIN (Bao et al. 2019), CAIN (Choi et al. 2020), AdaCoF (Lee et al. 2020), BMBC (Park et al. 2020), EDSC (Cheng and Chen 2021) and XVFI (Sim, Oh, and Kim 2021).

Quantitative results are provided in Table 3, our method achieves better scores on the Vimeo90K dataset, while enjoying fewer parameters. Specifically, our method outperforms the deformable convolution method (EDSC (Cheng and Chen 2021)) and optical flow method (BMBC (Park et al. 2020)) by 0.31 dB and 0.14 dB in term of PSNR. In addition, we can see that our method has the advantage of 0.15dB over the advanced method (XVFI (Sim, Oh, and Kim 2021)) even if the inputs are two frames.

Model Analysis

Note for an efficient and fair comparison to verify the effectiveness of each module, we train all models with different variants on Vimeo90K dataset for 300,000 iterations.

Ablation Study. To verify the individual effectiveness of proposed modules, we conduct a comprehensive ablation study on different variants.

Model (a): We directly concatenate adjacent and distant hidden states for temporal aggregation, and then directly fuse aggregated spatial information for temporal aggregation.

Model (b): We utilize AF to aggregate adjacent hidden states and directly concatenate distant hidden states for spatial aggregation, following a direct fusion of aggregated spatial information for temporal aggregation.

Model (c): We utilize DF to aggregate distant hidden states and directly concatenate adjacent hidden states for spatial aggregation, following a direct fusion of aggregated spatial information for temporal aggregation.

Model (d): We utilize AF and DF to aggregate adjacent and distant hidden states for spatial aggregation, and then directly fuse spatial information for temporal aggregation.

Model (e): The complete version of our method to use all strategies.

The numerical comparisons are shown in Table 4, showing that Model (b) and Model (c) benefit from the accurate adjacent spatial information fusion via AF or distant spatial information aggregation via DF, outperforming Model (a) by 0.30 dB and 0.14 dB, respectively. Compared to Model (b) and Model (c), Model (d) can aggregate all spatial information, and gains further improvement by 0.14 dB, but it lacks the ability to perform dynamic fusion of spatial information for temporal aggregation. In contrast, Model (e) adopts DS to guide all aggregated spatial information for temporal aggregation and achieves better performance. All these results on testsets validate the effectiveness of the proposed AF, DF and DS for the final reconstruction performance.

Setting	w/o BC	only FWC	only BWC	w/BC
Fast	36.47	36.56	36.69	36.86
Medium	35.15	35.23	35.35	35.47
Slow	33.19	33.25	33.32	33.43

Table 5: Comparison of different variants of store-and-fetch containers in PSNR. w/o BC denotes without bidirectional containers. FWC denotes forward container. BWC denotes backward container. w/ BC denotes with bidirectional container.

Effects of Immediate Store-and-Fetch Strategy. To verify the importance of our immediate store-and-fetch strategy, we thus conduct an ablation study on different container variants. As shown in Table 5, we remove bidirectional containers from our framework as baseline (w/o BC), which only utilizes adjacent hidden states via bidirectional interactive propagation. Compared to the baseline, we can find that adopting a single directional container (only FWC or only BWC) has a great improvement about 0.09dB and 0.22dB, respectively. This is because the adjacent frames are locally similar, and while encountering large motions or occlusions, the distant frames from the past or future can provide useful complementary information for supplement via DF. This motivates us to design bidirectional containers (w/BC) to simultaneous capture the past, current and future information for ST-VSR, and achieves the best results.

Conclusion

This paper proposes a framework for ST-VSR, involving a forward recurrent module (FRM) and a backward recurrent module (BRM) with bidirectional store-and-fetch containers. Specifically, it allows the network to fetch the past, current and future SR information to sufficiently explore long-range spatial-temporal correlations, while storing all updated SR information for the next refinement. In FRM and BRM, an adjacent fetch block (AF) and a distant fetch block (DF) are designed to explore the adjacent and distant information for spatial aggregation. Furthermore, we further propose a dynamic selection block (DS) to guide aggregated spatial information for temporal aggregation. We also conduct extensive experiments on three tasks, demonstrating our method significantly outperforms SOTA methods.

Acknowledgments

This work was supported by National Key R&D Project (2021YFC3320301), National Natural Science Foundation of China (62171325), Hubei Key R&D (2022BAA033) and CAAI-Huawei MindSpore Open Fund. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Bao, W.; Lai, W.-S.; Ma, C.; Zhang, X.; Gao, Z.; and Yang, M.-H. 2019. Depth-aware video frame interpolation. In *CVPR*, 3703–3712.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*, 4947–4956.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, volume 2, 168–172. IEEE.

Cheng, X.; and Chen, Z. 2021. Multiple video frame interpolation via enhanced deformable separable convolution. *TPAMI*, 44(10): 7029–7045.

Choi, M.; Kim, H.; Han, B.; Xu, N.; and Lee, K. M. 2020. Channel Attention Is All You Need for Video Frame Interpolation. In *AAAI*, 10663–10671.

Geng, Z.; Liang, L.; Ding, T.; and Zharkov, I. 2022. RSTT: Real-time Spatial Temporal Transformer for Space-Time Video Super-Resolution. In *CVPR*, 17441–17451.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *CVPR*, 1664–1673.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2019. Recurrent back-projection network for video super-resolution. In *CVPR*, 3897–3906.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2020. Space-time-aware multi-resolution video enhancement. In *CVPR*, 2859–2868.

Hu, M.; Jiang, K.; Liao, L.; Xiao, J.; Jiang, J.; and Wang, Z. 2022a. Spatial-Temporal Space Hand-in-Hand: Spatial-Temporal Video Super-Resolution via Cycle-Projected Mutual Learning. In *CVPR*, 3574–3583.

Hu, M.; Jiang, K.; Nie, Z.; and Wang, Z. 2022b. You Only Align Once: Bidirectional Interaction for Spatial-Temporal Video Super-Resolution. *ACM MM*.

Ji, B.; and Yao, A. 2022. Multi-Scale Memory-Based Video Deblurring. In *CVPR*, 1919–1928.

Jiang, H.; Sun, D.; Jampani, V.; Yang, M.-H.; Learned-Miller, E.; and Kautz, J. 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 9000–9008.

Jo, Y.; Oh, S. W.; Kang, J.; and Kim, S. J. 2018. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, 3224–3232.

Kang, J.; Jo, Y.; Oh, S. W.; Vajda, P.; and Kim, S. J. 2020. Deep space-time video upsampling networks. In *ECCV*, 701–717. Springer.

Kim, S. Y.; Oh, J.; and Kim, M. 2020. Fisr: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *AAAI*, volume 34, 11278–11286.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, H.; Kim, T.; Chung, T.-y.; Pak, D.; Ban, Y.; and Lee, S. 2020. AdaCoF: Adaptive Collaboration of Flows for Video Frame Interpolation. In *CVPR*, 5316–5325.

Liu, C.; and Sun, D. 2011. A bayesian approach to adaptive video super resolution. In *CVPR*, 209–216. IEEE.

Liu, Y.-L.; Liao, Y.-T.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Deep video frame interpolation using cyclic frame generation. In *AAAI*, volume 33, 8794–8802.

Niklaus, S.; Mai, L.; and Liu, F. 2017. Video frame interpolation via adaptive separable convolution. In *ICCV*, 261–270.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *ICCV*, 9226–9235.

Park, J.; Ko, K.; Lee, C.; and Kim, C.-S. 2020. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. *arXiv preprint arXiv:2007.12622*.

Shechtman, E.; Caspi, Y.; and Irani, M. 2005. Space-time super-resolution. *TPAMI*, 27(4): 531–545.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 1874–1883.

Sim, H.; Oh, J.; and Kim, M. 2021. Xvfi: Extreme video frame interpolation. In *ICCV*, 14489–14498.

Sukhbaatar, S.; Weston, J.; Fergus, R.; et al. 2015. End-to-end memory networks. *NeurIPS*, 28.

Tian, Y.; Zhang, Y.; Fu, Y.; and Xu, C. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 3360–3369.

Wang, H.; Xiang, X.; Tian, Y.; Yang, W.; and Liao, Q. 2022. Stdan: deformable attention network for space-time video super-resolution. *arXiv preprint arXiv:2203.06841*.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR*, 0–0.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J. P.; and Xu, C. 2020. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *CVPR*, 3370–3379.

Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; and Cheng, M.-M. 2021. Temporal Modulation Network for Controllable Space-Time Video Super-Resolution. In *CVPR*, 6388–6397.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *IJCV*, 127(8): 1106–1125.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; Lu, T.; Tian, X.; and Ma, J. 2021. Omniscient video super-resolution. In *ICCV*, 4429–4438.

Yi, P.; Wang, Z.; Jiang, K.; Jiang, J.; and Ma, J. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*, 3106–3115.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 286–301.

Zhou, C.; Lu, Z.; Li, L.; Yan, Q.; and Xue, J.-H. 2021. How Video Super-Resolution and Frame Interpolation Mutually Benefit. In *ACM MM*, 5445–5453.