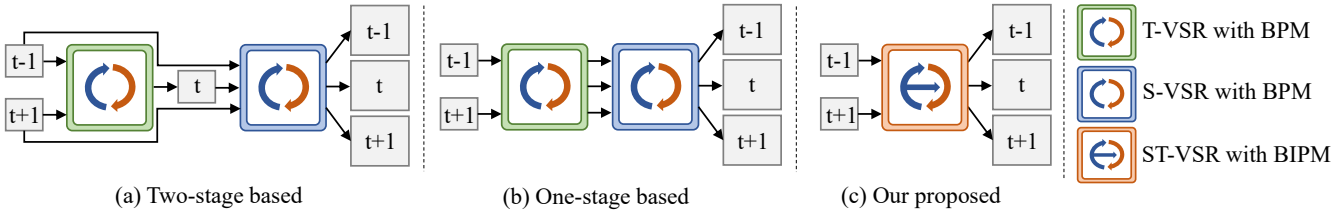
# You Only Align Once: Bidirectional Interaction for Spatial-Temporal Video Super-Resolution

Mengshun Hu
Kui Jiang*
Zhixiang Nie
Zheng Wang†
National Engineering Research Center for Multimedia Software, Hubei Key Laboratory of Multimedia and Network Communication Engineering, School of Computer Science, Wuhan University




(a) Two-stage based (b) One-stage based (c) Our proposed

**Figure 1: (a): Two-stage based methods: They perform ST-VSR by independently and sequentially using advanced S-VSR and T-VSR without merging common pipelines (*i.e.* feature extraction, alignment, fusion and reconstruction), both involving bidirectional propagation module (BPM) for alignment and fusion on image space. (b) One-stage based methods: They unify S-VSR and T-VSR into a single stage for ST-VSR with sharing feature extraction and reconstruction network, both involving bidirectional propagation module (BPM) for alignment and fusion on feature space. (c) Our proposed method: We can accurately and fast super-resolve videos by efficiently merging common pipelines from S-VSR and T-VSR, only involving bidirectional interactive propagation module (BIPM) for once alignment and fusion on feature space.**

## ABSTRACT

Spatial-Temporal Video Super-Resolution (ST-VSR) technology generates high-quality videos with higher resolution and higher frame rates. Existing advanced methods accomplish ST-VSR tasks through the association of Spatial and Temporal video super-resolution (S-VSR and T-VSR). These methods require two alignments and fusions in S-VSR and T-VSR, which is obviously redundant and fails to sufficiently explore the information flow of consecutive spatial LR frames. Although bidirectional learning (future-to-past and past-to-future) was introduced to cover all input frames, the direct fusion of final predictions fails to sufficiently exploit intrinsic correlations of bidirectional motion learning and spatial information from all frames. We propose an effective yet efficient recurrent network with bidirectional interaction for ST-VSR, where only one alignment and fusion is needed. Specifically, it first performs backward inference from future to past, and then follows forward inference to super-resolve intermediate frames. The backward and forward inferences are assigned to learn structures and details to simplify the learning task with joint optimizations. Furthermore, a Hybrid Fusion Module (HFM) is designed to aggregate and distill information to refine spatial information and reconstruct high-quality video frames. Extensive experiments on two public datasets demonstrate that our method outperforms state-of-the-art methods in efficiency, and reduces calculation cost by about 22%.

## CCS CONCEPTS

• **Human-centered computing** → **Visualization**; **Displays and imagers**; • **Computing methodologies** → **Computer vision**.

## KEYWORDS

Spatial-temporal, Video super-resolution, Bidirectional interaction

**ACM Reference Format:**
Mengshun Hu, Kui Jiang, Zhixiang Nie, and Zheng Wang. 2022. You Only Align Once: Bidirectional Interaction for Spatial-Temporal Video Super-Resolution. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547874

## 1 INTRODUCTION

Spatial-temporal video super-resolution (ST-VSR) refers to the task of generating the high-resolution (HR) and high-frame-rate (HFR) photo-realistic video sequences from the given low-resolution (LR) and low-frame-rate (LFR) input. This task has drawn increasing

*Equal contribution
†Corresponding author


*MM '22, October 10–14, 2022, Lisboa, Portugal*

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547874

attention and become a research hotspot in the field of multimedia [12, 48, 49, 54, 55], because of the broad range of applications such as movie production [23], high-definition television upgrades [22] and video compression [48], *etc.*

1) *Two-stage based*: To tackle ST-VSR, quite intuitively, a direct combination of temporal video super-resolution (T-VSR) [1] and spatial video super-resolution [21] (S-VSR) [1] can cope with some simple cases with small motions and scene changes. Sequential T-VSR and S-VSR barely considers the intrinsic relationship between these two tasks when generating HR and HFR videos, where rich spatial information can help temporal prediction and sequential information is crucial for texture inference (see Figure 1(a)). Therefore, this sort of two-stage based methods are far from producing satisfactory reconstruction results, especially for large motion and magnification factors. Moreover, separate T-VSR and S-VSR require repetitive operations such as feature extraction, alignment, fusion, and reconstruction, which is structurally redundant and inefficient.

2) *One-stage based*: By contrast, another sort of methods tackle ST-VSR on feature space [47–49, 54] by integrating T-VSR and S-VSR into a unified framework for joint optimization with shareable feature extraction and reconstruction modules (see Figure 1 (b)). Although spatial-temporal correlation learning is improved with a compact ST-VSR framework, these methods still require separate alignment and fusion operations for spatial and temporal information representations. The motions with bidirectional propagation module (BPM) in T-VSR and S-VSR are independent and lack interactions. The reconstruction of the current frame does not sufficiently explore spatial-temporal correlations from past, current and future frames, and such long-term relationships are critical for large motion estimation.

We propose to further compact ST-VSR framework and sufficiently explore spatial-temporal correlations by learning the interaction of bidirectional inference. The proposed framework introduces a novel bidirectional inferences scheme and a Bidirectional Interactive Propagation Module (BIPM) to implicitly align and mine spatial-temporal information. In BIPM, two Recurrent Cells (RCs) are equipped to enable bidirectional interaction so that the current representation can absorb knowledge from the past, current and future frames. In this way, our proposed framework only requires one alignment operation to capture the spatial-temporal correlations from all frames, dubbed You Only aliGn Once (YOGO).

In addition, since spatial information may gradually vanish during the propagation process, instead of directly reconstructing HR frame from the output of two recurrent cells [4, 47, 49], we propose a Hybrid Fusion Module (HFM) to further progressively refine spatial information via the outputs of two recurrent cells.

Our contributions are summarized as follows:

- We propose a novel yet high-efficiency framework for ST-VSR, namely YOGO. In YOGO, T-VSR and S-VSR are integrated into a unified network to promote the compact.
- We devise a bidirectional interactive propagation scheme to explore spatial-temporal correlations, where past, current and future knowledge from all frames are aggregated by updating the hidden states.
- We conduct extensive experiments to compare our YOGO on ST-VSR, which demonstrates our method performs well against the state-of-the-art ST-VSR methods in efficiency.

## 2 RELATED WORK

### 2.1 Spatial Video Super-resolution

S-VSR aims to super-resolve LR videos to HR videos by sufficiently utilizing temporal information. Thus, the key to this task lies in making full use of temporal correlations among multiple frames. The existing S-VSR method can be mainly divided into two frameworks: sliding-window and recurrent frameworks. The former firstly conducts motion estimations between low-resolution frames from a sliding window, and then perform spatial alignments based on predicted motions and fusion for HR reconstruction [2, 3, 26, 43, 45]. However, these methods are time-consuming and each input frame is processed and aligned multiple times. Moreover, Since they cannot build long-range temporal correlations from input videos, they tend to generate temporally consistent results.

Unlike S-VSR methods based on sliding window framework, due to the recurrent property, S-VSR methods based on recurrent framework [4, 5, 11, 18, 27, 36] are able to explore and utilize long-range temporal correlations without multiple alignments for each frame. For example, RSDN [19] proposes a recurrent dual-branch network to learn the structures and details of frames for S-VSR. But it fails to leverage the information from subsequent LR frames while conducting single-direction recurrent propagation. To alleviate this issue, Yi *et al.* design a novel recurrent network with bidirectional coupled propagation for ST-VSR, which can make full use of long-range temporal correlations by implicitly aligning and fusing the past, current, future information from video sequences [52].

### 2.2 Temporal Video Super-resolution

T-VSR (*i.e.*, video frame interpolation) aims to generate non-existent intermediate frames between consecutive input frames. Thus, the key to this task lies in finding pixel correspondences between consecutive frames. The common methods for T-VSR are mainly divided into two categories: flow-based methods [1, 16, 20, 34, 35, 42] and kernel-based methods [8, 32, 41]. flow-based methods [10] mainly consist of the following steps: feature extraction, forward and backward alignment, aligned intermediate representations fusion. intermediate frame reconstruction. To further refine reconstructed results, additional contextual information [2, 17, 29, 30] is introduced to a post-processing module to predict residual information for compensation. However, These methods rely heavily on optical flow estimation accuracy. As for kernel-based methods, some methods estimate dynamic convolution kernels to resample the input frames for intermediate frame interpolation [7, 25, 31–33, 41]. But most of these methods only consider resampling of local neighborhood patches, leading to blurry results.

### 2.3 Spatial-temporal Video Super-resolution

ST-VSR is to increase the spatial and temporal resolution of low-resolution (LR) and low-frame-rate (LFR) videos [37]. The key challenge lies in sufficiently exploiting the spatial-temporal information

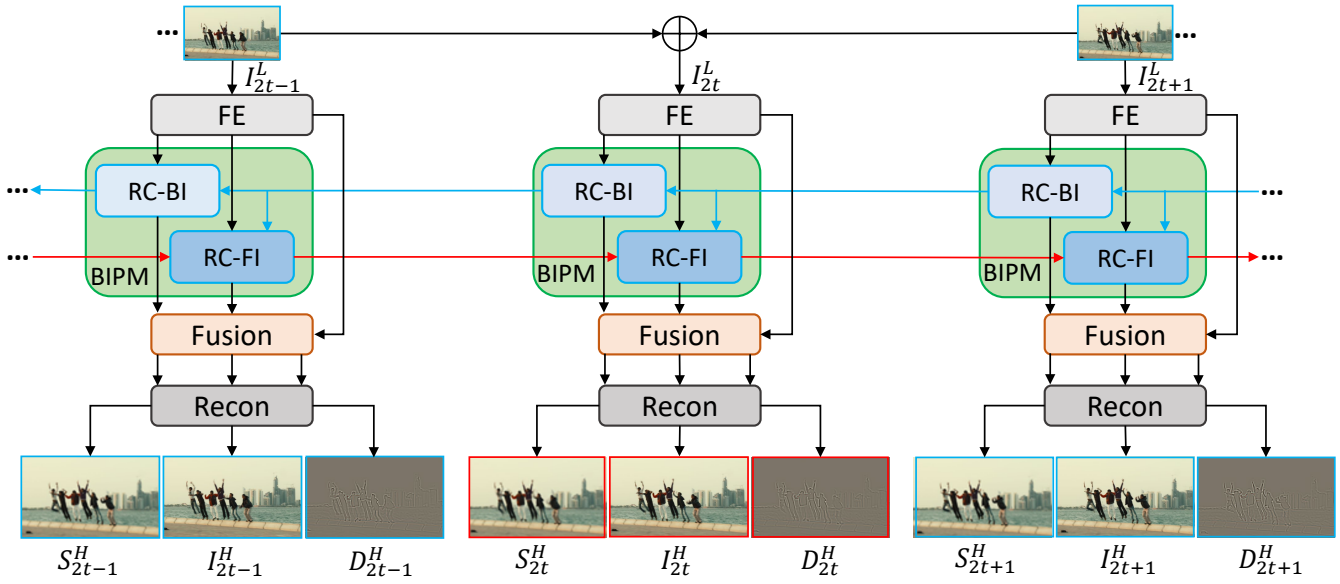



**Figure 2: Architecture of the proposed YOGO. Given multiple low-resolution (LR) and low-frame-rate (LFR) input frames, we firstly extract representations from input frames by feature extractor (FE). We then adopt a bidirectional interactive propagation module (BIPM) containing two recurrent cells (*e.g.* recurrent cell with backward inference (RC-BI) and recurrent cell with forward inference (RC-FI)) to implicitly align corresponding hidden states. Then, we assign backward and forward inferences to learn structures and details components from temporal information, and further progressively aggregate and distill structures and details components for spatial reconstruction via a hybrid fusion Module (HFM) in the Fusion process. Finally, we employ the reconstruction module (Recon) to output the high-resolution (HR) (x4) and high-frame-rate (HFR) (x2) videos, structures and details components.**

of input frames. For instance, Shechtman *et al.* [38] adopt a directional spatial-temporal smoothness regularization to constrain high spatial-temporal resolution. However, this constrain makes it difficult model spatial-temporal correlations on complex or diverse visual patterns. Lately, learning-based methods [9] attempt to decompose ST-VSR into two sub-tasks that are achieved on image space sequentially: spatial video super-resolution (S-VSR) and temporal video super-resolution (T-VSR). However, they fail to utilize intrinsic relations between S-VSR and T-VSR. Recently, some studies [48, 49] show one-stage based ST-VSR methods are significantly better than two-stage based ST-VSR methods on effectiveness and efficiency. STARnet [12], MBnet [55] and CycMu-Net [15] propose a mutual learning strategy for ST-VSR, which makes full use of spatial-temporal information by jointly learning S-VSR and T-VSR via iterations. Zooming Slow-Mo [47] propose to firstly capture local temporal contexts for intermediate feature interpolation by deformable convolution [56], then explore global temporal contexts to further build temporal correlations by bidirectional deformable ConvLSTM [40], and finally reconstructs high-resolution (HR) and high-frame-rate (HFR) spatial-temporal videos by a reconstruction network. Inspired by [47], Xu *et al.* [49] further propose a locally temporal feature comparison module to extract local motion cues for refinement, achieving better performances on two public datasets. However, these one-stage based methods are procedure redundant due to ignorance of effective combination of common pipelines.

# 3 PROPOSED METHOD

## 3.1 Framework Overview

Spatial-temporal video super-resolution (ST-VSR) aims to reconstruct high-resolution (HR) and high-frame-rate (HFR) video sequences $[I^H_t]^{2n+1}_{t=1}$ from given low-resolution (LR) and low-frame-rate (LFR) inputs $[I^L_{2t-1}]^{n+1}_{t=1}$. Figure 2 shows the pipeline of our proposed YOGO (You Only aliGn Once) algorithm, which mainly consists of four seamless components: Feature Extraction, Alignment with Bidirectional Interaction Propagation Module (BIPM), Fusion, Reconstruction.

Given LR and LFR video sequences $[I^L_{2t-1}]^{n+1}_{t=1}$, a feature extraction (FE) module, involving one convolution layer and five residual blocks [13], is used to project input frames into feature space to generate the initial representation ($[F^L_t]^{2n+1}_{t=1}$). Followed by a Bidirectional Interaction Propagation Module (BIPM), the temporal information between these frames is sufficiently exploited to estimate the motion. In particular, BIPM allows the network to aggregate the feature representation from the past, current and future frames. More specifically, recurrent cell with backward inference (RC-BI) learns the temporal relations from future to past frame, and packages them into a hidden unit while recurrent cell with forward inference (RC-FI) can exploit the spatial-temporal information from all frames by implicitly aligning and updating the packaged hidden unit. To further aggregate the spatial and temporal information, a hybrid fusion module (HFM) containing multiple hybrid fusion

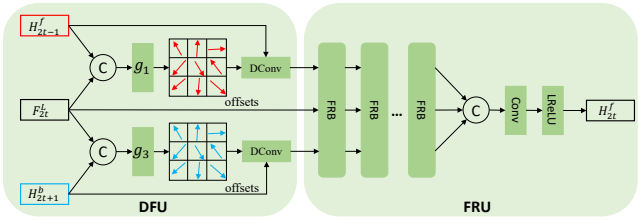



**Figure 3: Architecture of the proposed recurrent cell with forward inference (RC-FI). Note that the backward inference (RC-BI) shares the similar framework to RC-FI, but takes $H^b_{2t+1}$ and $F^L_{2t}$ as inputs.**

blocks (HFBs) is designed to combine the outcomes of these two inferences, and two pixel-shuffle layers [39] are used to yield the final HR and HFR solution $[I^H_t]^{2n+1}_{t=1}$.

## 3.2 Bidirectional Interactive Propagation Module

Advanced spatial-temporal video super-resolution (ST-VSR) methods sequentially perform spatial video super-resolution (S-VSR) and temporal video super-resolution (T-VSR) on image or feature spaces. However, individual alignment and fusion operations are adopted in these methods to fuse spatial and temporal information. Besides the structure redundancy, motion estimation and alignment in T-VSR and S-VSR are independent and lack interactions, which hinders accurate spatial-temporal correlation learning. To alleviate this issue, we propose a novel bidirectional interactive propagation module (BIPM), where the backward and forward inferences share the alignment pipeline to implicitly learn the spatial-temporal correlations from all frames. More specifically, a recurrent cell with backward inference (RC-BI) learns the temporal relations from future to past frame, and packages them into a hidden unit while a recurrent cell with forward inference (RC-FI) can exploit the spatial-temporal information from all frames by implicitly aligning and updating the packaged hidden unit. For convenience, RC-BI and RC-FI have the similar framework, composed of a dynamic filter unit (DFU) and a fusion residual unit (FRU). The proposed DFU explores short-term temporal correlations by aligning hidden states, while FRU further explores long-term variations of the whole video by utilizing the current input representation and aligned hidden states. This two-stage temporal feature alignment scheme accurately and sufficiently aggregates the spatial-temporal features.

**DFU:** Considering the feature redundancy and alignment errors between the different frames and backward-forward inference, we introduce a dynamic filter unit (DFU) to first learn offsets among the current input and hidden states. It provides the guidance to distill and fuse the informative components while mitigating the accumulated errors. Taking RC-FI as an example, as shown in Figure 3, the proposed DFU takes past hidden state $H^f_{2t-1}$ and input representation $F^L_{2t}$ , as well as the future hidden state $H^b_{2t+1}$ from backward inference as input, and predicts the offsets via the offset estimator, which guide the deformable convolutions to capture the most related components. The implicit alignment in DFU can be

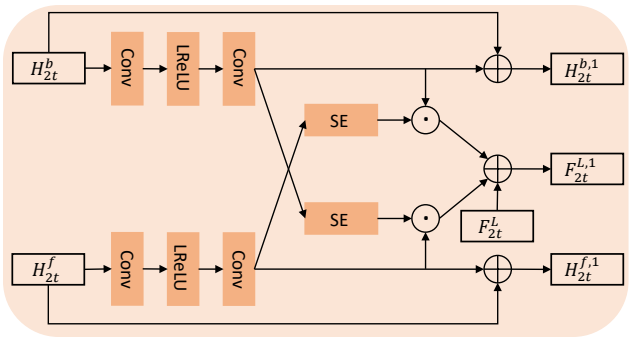



**Figure 4: Architecture of the proposed hybrid fusion block (HFB). SE denotes squeeze-and-excitation networks [14].**

expressed as

$$\begin{aligned} \Delta P_1 &= g_1([H^f_{2t-1}, F^L_{2t}]), \\ \Delta P_2 &= g_3([H^b_{2t+1}, F^L_{2t}]), \\ \hat{H}^f_{2t-1} &= DConv(H^f_{2t-1}, \Delta P_1), \\ \hat{H}^b_{2t+1} &= DConv(H^b_{2t+1}, \Delta P_2), \end{aligned} \tag{1}$$

where $\Delta P_1$ and $\Delta P_2$ are the learnable offsets via the offset estimator ($g_1(\cdot)$ and $g_3(\cdot)$) [45]. $DConv(\cdot)$ denotes the deformable convolution.

**FRU:** DFU focuses on the short-term temporal correlations between the adjacent frames, rendering itself unfit for the occluded scenes or large motion, which may benefit from the long-term correlation learning. To this end, a fusion residual unit (FRU) is introduced to aggregate the spatial-temporal information from all frames. Specifically, FRU first takes aligned hidden states ($\hat{H}^f_{2t-1}$ and $\hat{H}^b_{2t+1}$) and current representation ($F^L_{2t}$) as inputs, where a series of fusion residual blocks (FRB) are introduced to learn the individual representation of current frame while eliminating the spatial-temporal redundancy through $1 \times 1$ channel fusion [51, 53]. The outcomes of all branches can be combined to yield a final solution ($H^f_{2t}$) to the predicted frame.

To simplify the learning procedure, inspired by the divide-and-conquer strategy, the backward and forward inferences are assigned to learn specific tasks, where RC-BI focuses on the global structure while RC-FI aims to refine the textures. It has been verified that refining the details with past, current and future representation in RC-FI can gain better performance than that of RC-BI. (More details are included in Section 3.3.)

## 3.3 Hybrid Fusion Module

To sufficiently aggregate the spatial-temporal representation, the outputs ($H^b_{2t}$ and $H^f_{2t}$) of backward and forward inferences as well as the initial features ($F^L_{2t}$) are packed into a hybrid fusion module (HFM), where the individual components are progressively fused via multiple hybrid fusion blocks (HFBs) to eliminate the redundancy and yield the final prediction. Taking the first HFB as an example in Figure 4, HFB takes the outputs of RC-BI and RC-FI as inputs, and applies two branches to deeply characterize the specific features,

**Table 1: Quantitative comparisons of our results and two-stage ST-VSR methods on Vid4 and Vimeo90K datasets. The best and second best results are highlighted in red and blue, respectively. "Ours (56)" denotes YOGO with the filter number as 56.**

| VFI Method | VSR Method | Vid4 | | Vimeo90K-Fast | | Vimeo90K-Medium | | Vimeo90K-Slow | | Parameters (millions)↓ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | |
| SuperSloMo | Bicubic | 22.84 | 0.5772 | 31.88 | 0.8793 | 29.94 | 0.8477 | 28.37 | 0.8102 | 19.8 |
| SuperSloMo | RCAN | 23.80 | 0.6397 | 34.52 | 0.9076 | 32.50 | 0.8884 | 30.69 | 0.8624 | 19.8+16.0 |
| SuperSloMo | RBPN | 23.76 | 0.6362 | 34.73 | 0.9108 | 32.79 | 0.8930 | 30.48 | 0.8584 | 19.8+12.7 |
| SuperSloMo | EDVR | 24.40 | 0.6706 | 35.05 | 0.9136 | 33.85 | 0.8967 | 30.99 | 0.8673 | 19.8+20.7 |
| SepConv | Bicubic | 23.51 | 0.6273 | 32.27 | 0.8890 | 30.61 | 0.8633 | 29.04 | 0.8290 | 21.7 |
| SepConv | RCAN | 24.92 | 0.7236 | 34.97 | 0.9195 | 33.59 | 0.9125 | 32.13 | 0.8967 | 21.7+16.0 |
| SepConv | RBPN | 26.08 | 0.7751 | 35.07 | 0.9238 | 34.09 | 0.9229 | 32.77 | 0.9090 | 21.7+12.7 |
| SepConv | EDVR | 25.93 | 0.7792 | 35.23 | 0.9252 | 34.22 | 0.9240 | 32.96 | 0.9112 | 21.7+20.7 |
| DAIN | Bicubic | 23.55 | 0.6268 | 32.41 | 0.8910 | 30.67 | 0.8636 | 29.06 | 0.8289 | 24.0 |
| DAIN | RCAN | 25.03 | 0.7261 | 35.27 | 0.9242 | 33.82 | 0.9146 | 32.26 | 0.8974 | 24.0+16.0 |
| DAIN | RBPN | 25.96 | 0.7784 | 35.55 | 0.9300 | 34.45 | 0.9262 | 32.92 | 0.9097 | 24.0+12.7 |
| DAIN | EDVR | 26.12 | 0.7836 | 35.81 | 0.9323 | 34.76 | 0.9281 | 33.11 | 0.9119 | 24.0+20.7 |
| Ours (56) | | 26.28 | 0.7996 | 36.76 | 0.9397 | 35.32 | 0.9349 | 33.33 | 0.9134 | 9.5 |

**Table 2: Quantitative comparisons of our results and one-stage ST-VSR methods on Vid4 and Vimeo90K datasets. "Ours (56)" and "Ours (64)" denote YOGO with the filter number as 56 and 64, respectively. The total runtime is measured on the entire Vid4 dataset [28], Note we input four LR image with the resolution of 180×144 to test FLOPS.**

| ST-VSR Method | Vid4 | | Vimeo90K-Fast | | Vimeo90K-Medium | | Vimeo90K-Slow | | Speed (fps)↑ | FLOPs (T)↓ | Parameters (millions)↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | | | |
| STARnet | 26.06 | 0.8046 | 36.19 | 0.9368 | 34.86 | 0.9356 | 33.10 | 0.9164 | 14.92 | 27.926 | 111.6 |
| Zooming Slow-Mo | 26.31 | 0.7976 | 36.81 | 0.9415 | 35.41 | 0.9361 | 33.36 | 0.9138 | 17.34 | 1.766 | 11.1 |
| TMNet | 26.43 | 0.8016 | 37.04 | 0.9435 | 35.60 | 0.9380 | 33.51 | 0.9159 | 15.62 | 1.874 | 12.3 |
| Ours (56) | 26.28 | 0.7996 | 36.76 | 0.9397 | 35.32 | 0.9349 | 33.33 | 0.9134 | 16.72 | 1.148 | 9.5 |
| Ours (64) | 26.34 | 0.8022 | 36.93 | 0.9416 | 35.55 | 0.9365 | 33.44 | 0.9150 | 15.87 | 1.470 | 12.1 |

expressed as

$$H_{2t}^{b,1} = H_{2t}^{b} + R_1(H_{2t}^{b}),$$
$$H_{2t}^{f,1} = H_{2t}^{f} + R_2(H_{2t}^{f}), \tag{2}$$

where $R_1(\cdot)$ and $R_2(\cdot)$ denote residual blocks [13].

The middle branch is to exploit the correlations between the backward and forward inferences via cross-attention, the feature map from one modality can be used to enhance another modality. In this way, the spatial-temporal information from all frames is sufficiently aggregated for a better spatial reconstruction. Thus, this process is described as

$$F_{2t}^{L,1} = F_{2t}^{L} + SE_1(R_1(H_{2t}^{b})) \odot R_2(H_{2t}^{f}) + SE_2(R_2(H_{2t}^{f})) \odot R_1(H_{2t}^{b}), \tag{3}$$

where $SE_1(\cdot)$ and $SE_2(\cdot)$ denote squeeze-and-excitation networks [14].

## 3.4 Reconstruction and Optimization

Finally, a reconstruction module (Recon) is designed to output the HR (4× ) and HFR (2× ) video $[I_t^H]_{t=1}^{2n+1}$, involving two pixel-shuffle layers [39] and a sequence of "Conv-LeakyReLU-Conv" operations. Meanwhile, specific representations of backward and forward inferences are projected into the image space to generate the corresponding structures $([S_t^H]_{t=1}^{2n+1})$ and details $([D_t^H]_{t=1}^{2n+1})$. Then, the loss functions on the predicted frames, structure and detail images are expressed as

$$L_r = \sum_{t=1}^{2n+1} (\rho(I_t^H - I_t^{GT}) + \rho(D_t^H - D_t^{GT}) + \rho(S_t^H - S_t^{GT})), \tag{4}$$

where $I_t^{GT}$, $D_t^{GT}$, $S_t^{GT}$ refer to the corresponding ground-truth video frames, details and structures components, where the detail components denote the residue 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 where constant $w$ is set to $10^{-3}$ [6].

## 3.5 Implementation Details

In our study, the training images are randomly cropped into small patches with a fixed size of 112×64 and randomly rotated and flipped. We take out the odd-indexed 4 frames as LR and LFR inputs,

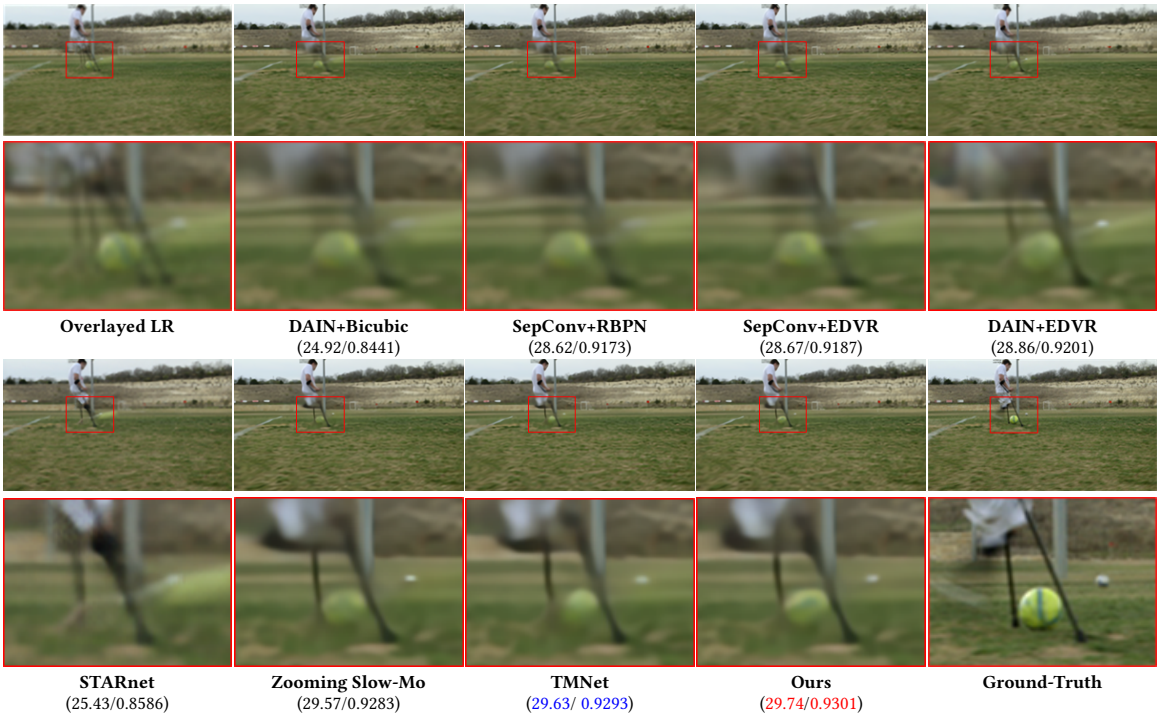



**Figure 5: Visual comparisons with state-of-the-art two-stage and one-stage based methods on Vimeo90K dataset.**

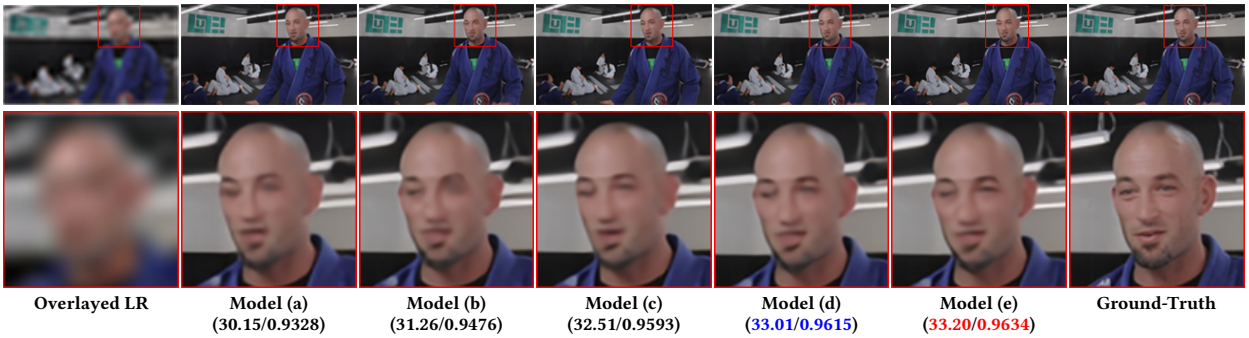



**Figure 6: Visual comparisons of five variants for the ablation studies Vimeo90K dataset.**

and the corresponding consecutive HR 7-frame for supervision. Specifically, the batch size is set to 10. We trained our proposed YOGO using Pytorch 1.9 with four NVIDIA Tesla V100 and adopt AdaMax optimizer [24], where the initial learning is set to $10^{-4}$ with the decay rate of 0.1 at every 30 epochs till 70 epochs.

## 4 EXPERIMENTS AND ANALYSIS

### 4.1 Datasets and Metrics

Similar to [48], the **Vimeo90K trainset** is used to train our YOGO and other compared methods for fairness. This dataset consists of more than 60,000 7-frame training video sequences with the resolution of 448×256 [50]. In addition, **Vid4** [28] and **Vimeo90K** testsets are used as the evaluation datasets. Following [47], **Vimeo90K** testsets are split into three subsets of fast, medium and slow motion, including 1225, 4977 and 1613 video sequences, respectively. In this study, before passed into the network, the odd-indexed LR frames are sampled via bicubic to generate the degraded inputs. Two commonly used evaluation metrics, such as Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) [46], as well as the Frame Per Second (fps) are employed for comparison. Higher fps indicates faster speed.

### 4.2 Comparison with State-of-the-Art Methods

We compare our proposed network with state-of-the-art two-stage based ST-VSR methods. We perform T-VSR by SuperSloMo [20], SepConv [32] and DAIN [1], and perform S-VSR by Bicubic Interpolation, RCAN [44], RBPN [11], and EDVR [45]. In addition,


**Figure 7: Feature maps of DFU and FRU outputs from RC-FI have been visualized using same grayscale colormap. The first rows indicates overlayed the second and third LR frames, aligned feature of DFU from the second frame, and aligned feature of FRU from the second frame, respectively. The second rows indicates overlayed the third and fourth LR frames, aligned feature of DFU from the fourth frame, and aligned feature of FRU from the fourth frame, respectively.**

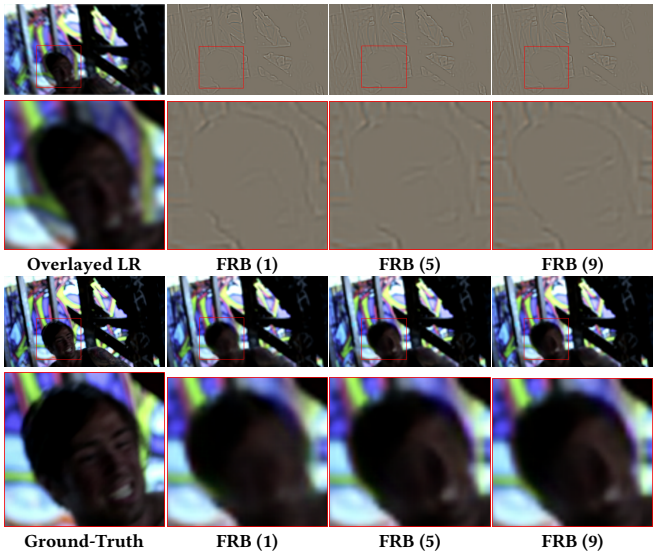



**Figure 8: Visual comparisons of details maps from different numbers of hybrid fusion blocks (HFBs) in hybrid fusion module (HFM). The third and fourth rows represent visual comparisons of structures maps from different numbers of HFBs in HFM.**

we also compare our proposed network with state-of-the-art one-stage based ST-VSR methods, including STARnet [12], Zooming SlowMo [47] and TMNet [49]. we test these methods on the all testsets based on public codes provided by authors.

**Quantitative results.** Quantitative results are tabulated in Table 1 and Table 2. It is obvious that two-stage based methods require more parameters while one-stage based methods achieve better performances with only half or even fewer parameters. This is mainly attributed to the fact that two-stage based methods contain multiple redundant pipelines, like the individual feature extraction and reconstruction in T-VSR and T-VSR. Furthermore, the best two-stage methods is 0.95dB lower than our method YOGO (56) on Vimeo90k-Fast dataset. In addition, While the accuracy of YOGO (64) is marginally worse than TMNet [49], it is faster by more than 0.25fps with only about 78% of calculation cost. Compared to Zooming Slow-Mo [47], YOGO (64) achieves better results on all testsets with only about 83% of calculation cost. Furthermore, YOGO (56) also achieve competitive accuracy with less parameters and calculation cost. All these results validate the effectiveness of our proposed method for ST-VSR in efficiency.

**Qualitative results.** Figure 5 provides the visual comparison of three mainstream one-stage and two-stage based ST-VSR baselines, while the quantitative results (PSNR and SSIM) are also compared. It is observed that two-stage based ST-VSR methods tend to produce blurry results with more artifacts (see the stick in the red boxes). The main reason is that the T-VSR and S-VSR are performed independently, where the spatial-temporal correlations between two tasks are under-explored [22, 47]. Compared to two-stage based methods, one-stage based methods can generate sharper results. However, these methods cannot sufficiently learn the intrinsic correlations of bidirectional motion and ignore the difficulties of recovering structures and details information for spatial reconstruction, producing blurry results (see the ball and stick in the red boxes). On the contrary, our proposed method can sufficiently aggregate the spatial and temporal information from all frames via bidirectional interactive learning, which is helpful for texture reconstruction. Moreover, the divide-and-conquer learning strategy allows the network to focus more on the specific representation, which is followed by a fusion and reconstruction to generate more nature and pleasant results. (More comparisons are included in our supplementary document)

## 4.3 Model Analysis

In model analysis, we train all models for 50 epochs on the **Vimeo90k** dataset with $32 \times 32$ cropped small patches, and test them on **Vid4** and **Vimeo90K** testsets.

**Ablation study.** To further verify the effectiveness of different modules in our proposed network, we conduct a comprehensive ablation study on different variants.

**Model (a):** We utilize a recurrent cell with only backward inference scheme to implicit align hidden states from temporal information, Then directly fuse and refine spatial information for reconstruction.

**Model (b):** We utilize two recurrent cells with bidirectional inferences scheme to implicit align hidden states from temporal information, Then directly fuse and refine spatial information for reconstruction.

**Model (c):** We utilize two recurrent cells with bidirectional inferences scheme to implicit align hidden states from temporal information, Then fuse and distill structures and details to refine spatial information for reconstruction via multiple hybrid fusion blocks (HFBs).

**Model (d):** We utilize two recurrent cells with bidirectional interactive inferences scheme to implicit align hidden states from temporal information, Then directly fuse and refine spatial information for reconstruction.

**Model (e):** The complete version of YOGO.

**Table 3: Quantitative comparisons in PSNR on the performances of different modules. Direct Fusion denotes the direct processing of temporal information in fusion process, SD Fusion denotes processing of structure and detail components from temporal information in fusion process.**

| Setting | Model (a) | Model (b) | Model (c) | Model (d) | Model (e) |
|---|---|---|---|---|---|
| Single Direction | ✓ | ✗ | ✗ | ✗ | ✗ |
| Bidirection | ✗ | ✓ | ✓ | ✗ | ✗ |
| Bidirectional Interaction | ✗ | ✗ | ✗ | ✓ | ✓ |
| Direct Fusion | ✓ | ✓ | ✗ | ✓ | ✗ |
| SD Fusion | ✗ | ✗ | ✓ | ✗ | ✓ |
| Vimeo90K-Fast | 35.36 | 35.51 | 35.71 | 36.08 | 36.18 |
| Vimeo90K-Medium | 34.21 | 34.36 | 34.58 | 34.86 | 34.96 |
| Vimeo90K-Slow | 32.42 | 32.56 | 32.74 | 32.93 | 33.02 |

The numerical comparisons and visual comparisons are shown in Table 3 and Figure 6, we can see that **Model (b)** outperforms **Model (a)** by 0.15dB, 0.15dB and 0.14dB on **Vimeo90K-Fast**, **Vimeo90K-Medium** and **Vimeo90K-Slow** dataset, respectively, and produce clearer results (the left eye of the person), since information from past and future frame can be utilized via bidirectional inferences scheme. In addition, **Model (c)** is significantly better than **Model (b)** (see eyes of the person in red boxes). This is mainly due to the fact that structure-detail fusion can focus more on the specific representation for spatial refinement. Furthermore, we observe that **Model (e)** contains more textures and further improve ST-VSR performance over **Model (c)** on three testsets, which also demonstrate that bidirectional interactive inferences scheme is more conducive to sufficiently utilize temporal information, contributing to spatial structure and detail reconstruction.

**Different numbers of FRB in FRU from RC-BI and RC-FI:** Our proposed two recurrent cells (RCs) consists of recurrent cell with backward inference (RC-BI) and recurrent cell with forward inference (RC-FI), which individually explore structures and details from temporal information. In this section, we conduct the detailed experiments to adjust fusion residual block (FRB) in FRU from RC-BI and RC-FI to explore the optimal proportion for ST-VSR. Thus, we keep the total number of FRB fixed (10), and adjust the FRB in RC-BI (0,2,4,6,8,10) and RC-FI (10,8,6,4,2,0). As shown in Table 4, we find neither of "0+10" and "10+0" cannot produce optimal results. This also proves that only exploring long-term temporal information from details in RC-FI or structures in RC-BI fails to mutually learn two components and cannot sufficiently and accurately utilize temporal information, affecting the recovery of spatial structures and details. On the contrary, "4+6" outperforms all other proportions in structures and details components reconstruction. This also proves that it is important to simultaneously explore the long-term structures and details from temporal information for spatial recovery, especially details information.

**The Impact of DFU and FRU:** To further analyze the specific role of DFU and FRU, as shown in Figure 7 and Table 5, we visualize bidirectional output feature maps from DFU and FRU in a gray level. We can find the former is more concerned with local motion information when exploring short-term temporal correlations, while the latter is more effective in exploring the global motion information by utilizing the long-term temporal correlations of the whole sequence, especially motion boundaries and background information.

**Table 4: : PSNR (dB) evaluated by adjusting FRB in FRU. "6+4" denotes setting 6 FRBs in RC-BI and 4 in RC-FI.**

| Setting | 0+10 | 2+8 | 4+6 | 6+4 | 8+2 | 10+0 |
|---|---|---|---|---|---|---|
| Vimeo90K-Fast | 36.10 | 36.13 | 36.18 | 36.12 | 36.04 | 36.00 |
| Vimeo90K-Medium | 34.86 | 34.90 | 34.96 | 34.88 | 34.82 | 34.79 |
| Vimeo90K-Slow | 32.94 | 32.98 | 33.02 | 32.96 | 32.92 | 32.88 |

**Table 5: PSNR (dB) evaluated by ablating DFU and FRU.**

| Setting | DFU | FRU | FRU+DFU | DFU+FRU |
|---|---|---|---|---|
| Vimeo90K-Fast | 35.88 | 35.68 | 36.14 | 36.18 |
| Vimeo90K-Medium | 34.62 | 34.63 | 34.89 | 34.96 |
| Vimeo90K-Slow | 32.77 | 32.79 | 32.98 | 33.02 |

This is also in line with our expectation that DFU and FRU perform two-stage feature alignment scheme accurately aggregate the spatial-temporal information. In addition, "DFU+FRU" outperforms "FRU+DFU" on three testsets. The main reason is that performing FRU before DFU brings noisy long-term irrelevant information, confusing the short-term temporal correlations learning of DFU. This indicates that it is essential that we firstly utilize DFU and then FRU to sufficiently explore and utilize temporal information.

**Different Numbers of HFB in HFM:** HFB is mainly to aggregate and distill structures and details components from temporal information to further refine spatial information. In this section, we conduct the ablation study to verify the impact of different numbers of HFB. As shown in Figure 8, We can see that as the number of HFBs increases, the proposed network can generate more high-frequency information for texture restoration and clearer structure. Considering the trade-off between efficacy and efficiency, we set the numbers of HFB as 9 in the fusion process. The above experiments also show that more HFBs are helpful for spatial reconstruction by assigning learning specific tasks based on the divide-and-conquer strategy.

## 5 CONCLUSIONS

We proposed a YOGO method, which introduces a novel bidirectional interactive propagation module (BIPM) to sufficiently aggregate the spatial-temporal information. Only one alignment and fusion are required in YOGO where feature representations can benefit from the past, current and future information via the bidirectional interaction. Furthermore, a Hybrid Fusion Module (HFM) is designed to aggregate and distill information to refine spatial information and reconstruct high-quality video frames. Extensive quantitative and qualitative evaluations demonstrate our proposed method performs well against the state-of-the-art methods in ST-VSR tasks.

**Acknowledgements**. This work was supported by National Key R&D Project (2021YFC3320301) and National Natural Science Foundation of China (62171325). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

# REFERENCES


[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Depth-aware video frame interpolation. In *CVPR*. 3703–3712.

[2] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE TPAMI* (2019).

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*. 4778–4787.

[4] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. BasicVSR: The search for essential components in video super-resolution and beyond. In *CVPR*. 4947–4956.

[5] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2021. BasicVSR++: Improving Video Super-Resolution with Enhanced Propagation and Alignment. *arXiv preprint arXiv:2104.13371* (2021).

[6] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, Vol. 2. IEEE, 168–172.

[7] Xianhang Cheng and Zhenzhong Chen. 2020. Video frame interpolation via deformable separable convolution. In *AAAI*, Vol. 34. 10607–10614.

[8] Xianhang Cheng and Zhenzhong Chen. 2021. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE TPAMI* (2021).

[9] Saikat Dutta, Nisarg A Shah, and Anurag Mittal. 2021. Efficient space-time video super resolution using low-resolution flow and mask upsampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 314–323.

[10] Shurui Gui, Chaoyue Wang, Qihua Chen, and Dacheng Tao. 2020. Featureflow: Robust video interpolation via structure-to-texture generation. In *CVPR*. 14004–14013.

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. 2019. Recurrent back-projection network for video super-resolution. In *CVPR*. 3897–3906.

[12] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. 2020. Space-time-aware multi-resolution video enhancement. In *CVPR*. 2859–2868.

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

[14] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[15] Mengshun Hu, Kui Jiang, Liang Liao, Jing Xiao, Junjun Jiang, and Zheng Wang. 2022. Spatial-Temporal Space Hand-in-Hand: Spatial-Temporal Video Super-Resolution via Cycle-Projected Mutual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3574–3583.

[16] Mengshun Hu, Liang Liao, Jing Xiao, Lin Gu, and Shin'ichi Satoh. 2020. Motion Feedback Design for Video Frame Interpolation. In *ICASSP*. IEEE, 4347–4351.

[17] Mengshun Hu, Jing Xiao, Liang Liao, Zheng Wang, Chia-Wen Lin, Mi Wang, and Shin'ichi Satoh. 2021. Capturing Small, Fast-Moving Objects: Frame Interpolation via Recurrent Motion Enhancement. *IEEE TCSVT* (2021).

[18] Yan Huang, Wei Wang, and Liang Wang. 2017. Video super-resolution via bidirectional recurrent convolutional networks. *IEEE TPAMI* 40, 4 (2017), 1015–1028.

[19] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. 2020. Video super-resolution with recurrent structure-detail network. In *European Conference on Computer Vision*. Springer, 645–660.

[20] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*. 9000–9008.

[21] Kui Jiang, Zhongyuan Wang, Peng Yi, and Junjun Jiang. 2020. Hierarchical dense recursive network for image super-resolution. *Pattern Recognition* 107 (2020), 107475.

[22] Jaeyeon Kang, Younghyun Jo, Seoung Wug Oh, Peter Vajda, and Seon Joo Kim. 2020. Deep space-time video upsampling networks. In *ECCV*. Springer, 701–717.

[23] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2020. FISR: deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *AAAI*, Vol. 34. 11278–11286.

[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[25] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. 2020. Adacof: Adaptive collaboration of flows for video frame interpolation. In *CVPR*. 5316–5325.

[26] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. 2020. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*. Springer, 335–351.

[27] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. 2021. COMISR: Compression-Informed Video Super-Resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2543–2552.

[28] Ce Liu and Deqing Sun. 2011. A bayesian approach to adaptive video super resolution. In *CVPR*. IEEE, 209–216.

[29] Simon Niklaus and Feng Liu. 2018. Context-aware synthesis for video frame interpolation. In *CVPR*. 1701–1710.

[30] Simon Niklaus and Feng Liu. 2020. Softmax splatting for video frame interpolation. In *CVPR*. 5437–5446.

[31] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive convolution. In *CVPR*. 670–679.

[32] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive separable convolution. In *ICCV*. 261–270.

[33] Simon Niklaus, Long Mai, and Oliver Wang. 2021. Revisiting adaptive convolutions for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1099–1109.

[34] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. 2020. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *European Conference on Computer Vision*. Springer, 109–125.

[35] Junheum Park, Chul Lee, and Chang-Su Kim. 2021. Asymmetric bilateral motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14539–14548.

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

[37] Oded Shahar, Alon Faktor, and Michal Irani. 2011. *Space-time super-resolution from a single video*. IEEE.

[38] Eli Shechtman, Yaron Caspi, and Michal Irani. 2005. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 4 (2005), 531–545.

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

[40] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems* 28 (2015).

[41] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. 2021. Video Frame Interpolation via Generalized Deformable Convolution. *IEEE TMM* (2021).

[42] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. 2021. XVFI: eXtreme Video Frame Interpolation. *arXiv preprint arXiv:2103.16206* (2021).

[43] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. 2017. Detail-revealing deep video super-resolution. In *CVPR*. 4472–4480.

[44] Longguang Wang, Yulan Guo, Zaiping Lin, Xinpu Deng, and Wei An. 2018. Learning for video super-resolution through HR optical flow estimation. In *ACCV*. Springer, 514–529.

[45] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*. 0–0.

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

[47] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. 2020. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *CVPR*. 3370–3379.

[48] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. 2021. Zooming SlowMo: An Efficient One-Stage Framework for Space-Time Video Super-Resolution. *arXiv preprint arXiv:2104.07473* (2021).

[49] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. 2021. Temporal Modulation Network for Controllable Space-Time Video Super-Resolution. In *CVPR*. 6388–6397.

[50] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *IJCV* 127, 8 (2019), 1106–1125.

[51] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, and Jiayi Ma. 2020. A progressive fusion generative adversarial network for realistic and consistent video super-resolution. *TPAMI* (2020).

[52] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, Tao Lu, Xin Tian, and Jiayi Ma. 2021. Omniscient Video Super-Resolution. *arXiv preprint arXiv:2103.15683* (2021).

[53] Peng Yi, Zhongyuan Wang, Kui Jiang, Junjun Jiang, and Jiayi Ma. 2019. Progressive fusion video super-resolution network via exploiting non-local spatio-temporal correlations. In *ICCV*. 3106–3115.

[54] Chenyu You, Lianyi Han, Aosong Feng, Ruihan Zhao, Hui Tang, and Wei Fan. 2022. MEGAN: Memory Enhanced Graph Attention Network for Space-Time Video Super-Resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1401–1411.

[55] Chengcheng Zhou, Zongqing Lu, Linge Li, Qiangyu Yan, and Jing-Hao Xue. 2021. How Video Super-Resolution and Frame Interpolation Mutually Benefit. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5445–5453.

[56] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable convnets v2: More deformable, better results. In *CVPR*. 9308–9316.