Towards Interpretable Video Super-Resolution via Alternating Optimization

Jiezhang Cao[1], Jingyun Liang[1], Kai Zhang[1⋆], Wenguan Wang[1], Qin Wang[1], Yulun Zhang[1], Hao Tang[1], and Luc Van Gool[1,2]

[1]Computer Vision Lab, ETH Zürich, Switzerland [2]KU Leuven, Belgium
{jiezhang.cao, jingyun.liang, kai.zhang, wenguan.wang, qin.wang, yulun.zhang, hao.tang, vangool}@vision.ee.ethz.ch
https://github.com/caojiezhang/DAVSR

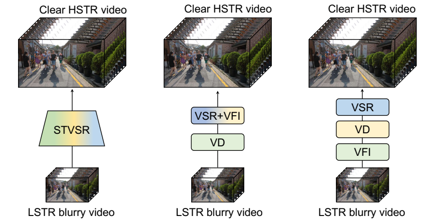
Abstract. In this paper, we study a practical space-time video super-resolution (STVSR) problem which aims at generating a high-framerate high-resolution sharp video from a low-framerate low-resolution blurry video. Such problem often occurs when recording a fast dynamic event with a low-framerate and low-resolution camera, and the captured video would suffer from three typical issues: i) motion blur occurs due to object/camera motions during exposure time; ii) motion aliasing is unavoidable when the event temporal frequency exceeds the Nyquist limit of temporal sampling; iii) high-frequency details are lost because of the low spatial sampling rate. These issues can be alleviated by a cascade of three separate sub-tasks, including video deblurring, frame interpolation, and super-resolution, which, however, would fail to capture the spatial and temporal correlations among video sequences. To address this, we propose an interpretable STVSR framework by leveraging both model-based and learning-based methods. Specifically, we formulate STVSR as a joint video deblurring, frame interpolation, and super-resolution problem, and solve it as two sub-problems in an alternate way. For the first sub-problem, we derive an interpretable analytical solution and use it as a Fourier data transform layer. Then, we propose a recurrent video enhancement layer for the second sub-problem to further recover high-frequency details. Extensive experiments demonstrate the superiority of our method in terms of quantitative metrics and visual quality.

Keywords: Video Super-Resolution, Motion Blur, Motion Aliasing

1 Introduction

Compared with existing space-time video super-resolution (STVSR) methods [55,56], we mainly focus on the more practical STVSR problem which aims at synthesizing a high space-time resolution (HSTR) clear video from a low space-time resolution (LSTR) blurry video. Although great progress has been made in existing STVSR methods, these methods mainly solve the video frame interpolation and video super-resolution tasks jointly, neglecting the motion blur

[⋆] Corresponding author.



(a) One-stage method (b) Two-stage method (c) Three-stage method

Fig. 1: Illustration of one-stage, two-stage and three-stage methods.

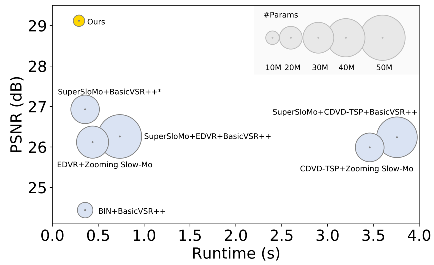


Fig. 2: Comparison on performance, runtime and parameter number.

and motion aliasing artifacts [39] that often occur in many real-world scenarios due to limited shutter speed and non-negligible exposure time. Different from exiting methods, we take above two temporal motion degradations into consideration and formulate it as a joint video restoration problem of the video frame interpolation, video deblurring and spatial video super-resolution.

To address the video restoration problem, one straightforward way is to directly combine a video frame interpolation (VFI) method (*e.g.*, SuperSloMo [15] and DAIN [3]), a video deblurring (VD) method (*e.g.*, EDVR [53] and CDVD-TSP [38]), and a video super-resolution (VSR) method (*e.g.*, IconVSR [6] and BasicVSR++ [7]) in a two-stage or three-stage manner, as shown in Fig. 1. For example, one can firstly interpolate missing intermediate video frames with VFI methods, then deblur the frames with VD methods, and finally super-resolve them with VSR methods. In this case, the VFI methods cannot eliminate motion blur nor motion aliasing because they only synthesize new blurry frames, while the VD methods cannot increase the framerate and resolve motion aliasing.

Alternatively, there are other combinations of VFI, VD and VSR methods, but solving these sub-tasks separately with multi-stage methods may suffer from the following limitations. First, ignoring the correlations among VFI, VD and VSR may lead to limited performance since these video restoration tasks are highly intra-related. As will be discussed in Sec. 3.1, the degradation from an HSTR clear video to an LSTR blurry video can be well-modelled by a single joint model. A "divide-and-conquer" strategy may not be able to benefit from the natural intra-relatedness and suffer from accumulated reconstruction errors from the first stage to the last stage. Second, as shown in Fig. 2, the composition of different methods may lead to expensive computational cost and a large number of parameters. This is because the overall runtime and parameter number are the summation of different standalone methods. Therefore, developing a one-stage unified model for the STVSR problem may a better choice.

To solve the above problem, in this paper, we propose a novel STVSR framework that exploits the correlation among different sub-tasks and boosts the overall efficiency significantly. We first reformulate this space-time video super-resolution problem as two sub-problems according to the half quadratic splitting algorithm, and solve them by an analytical solution and a deep learning-based neural network, respectively. More specifically, for the first sub-problem, we solve it based on the fast Fourier transform and propose a Fourier data transform layer to alleviate the motion blur and motion aliasing. For the second sub-problem, we propose a recurrent video enhancement layer with a multi-scale recurrent neural network to enhance the quality of the restored videos. Based on the alternating optimization, our end-to-end training method is able to jointly handle video frame interpolation, video deblurring and video super-resolution in STVSR.

The main contributions of this paper are summarized as follows:

- We formulate a more practical space-time video super-resolution (STVSR) problem by exploring the camera's intrinsic properties related to motion blur, motion aliasing and other spatial degradation.
- We make the first attempt to provide an analytical solution for the STVSR problem by leveraging the model-based methods and learning-based methods. With the help of the analytical solution, we develop a deep alternative video super-resolution network (DAVSR) to improve STVSR performance.
- We propose a new one-stage framework that can address video frame interpolation, video deblurring and video super-resolution simultaneously. By exploiting the correlation among the three sub-problems, our method is more effective than two-stage and three-stage methods.
- Our method achieves the state-of-the-art performance on both the REDS4 and Vid4 datasets. It is able to restore high-resolution and high-framerate videos even when the input videos have severe motion blur. Moreover, it only has a small number of parameters and has fast inference time.

2 Related Work

In this section, we discuss the related literature for video frame interpolation (VFI), video deblurring, video super-resolution (VSR) and other joint tasks as they are closely related to our practical STVSR problem.

Video frame interpolation (VFI) aims to synthesize intermediate frames between adjacent frames of the original frames. Recent VFI methods [28] propose to learn image matching or take local convolution over two input frames with a learned adaptive convolution kernel. Meyer et al. [30] propose a phase-based frame interpolation method which represents motion in the phase shift of individual pixels. In addition, flow-based video interpolation methods [3,4,15,27,35] propose to handle motions by estimating the optical flow. However, directly using VFI cannot reduce motion blur and motion aliasing [39]. This can become an potential issue in our STVSR probelm setup as the input videos are blurry.

Video deblurring aims at removing the blur artifacts from the input videos. Depending on the number of required input frames, there are multiple-frame [46,16,18,33,13,21,23] and single-frame [49,46,20] deblurring methods. EDVR [53] restores high-quality deblurred frames by first extracting features of multiple inputs and then conducting feature alignment and fusion. To further exploit the temporal information, some recurrent mechanisms based video deblurring methods have been proposed [14,60,58,34]. Zhou et al. [62] propose a deblurring network based on filter adaptive convolutional layers. Recently, CDVD-TSP [38], a CNN-based video deblurring method, approaches the problem by optical flow estimation and latent frame restoration steps. To solve the STVSR problem, one can combine video deblurring methods with the traditional STVSR methods.

Video super-resolution reconstructs HR video frames from the corresponding LR frames. There are several VSR methods [5,48,41,52,57,23] that use optical flow for explicit temporal alignment. Recently, RBPN [12] combines ideas from single- and multiple-frame SR for VSR, and estimates inter-frame motion to generate SR frames. However, the estimated flow is often inaccurate, resulting in poor performance. To address this, DUF [17] synthesizes SR frames by generating dynamic upsampling filters and a residual image based on the local spatio-temporal neighborhood of each pixel without explicit motion estimation. TDAN [51] proposes deformable alignment at the feature level without computing optical flow. Based on TDAN, EDVR [53] aligns frames at the feature level using deformable convolution networks (DCN) in a coarse-to-fine manner, and proposes an attention module to fuse different frames both temporally and spatially. However, most of the above methods are computationally inefficient due to many-to-one frameworks. To ease this, recurrent neural networks (RNN) are adopted in VSR methods [6,7] for leveraging temporal information. BasicVSR and its extension (*i.e.*, IconVSR) [6] propose to improve the performance of feature alignment using bidirectional propagation. To improve the performance of BasicVSR, BasicVSR++[7] proposes second-order grid propagation and flow guided deformable alignment. To address the STVSR problem, one can combine the above VFI, VD and VSR methods in a three-stage manner.

Space-time video super-resolution (STVSR) aims at synthesizing a high-resolution slow-motion video from a low-framerate and low-resolution video. Shechtman et al. [43] are among the first to extend SR methods to the space-time domain. Further STVSR methods based on Markov random field [31] and motion assisted steering kernel regression [47] are then proposed. In addition, Shahar et al. [42] explore the degree of the recurrences within natural videos. However, these methods are computationally expensive in practice. To address this, Xiang et al. [55] propose a one-stage STVSR network to directly learn the mapping from partial LR frames to HR frames. Different from our problem setting, traditional STVSR methods ignore the blur degradation from the camera, which can be critical in real-world applications.

Video frame interpolation and deblurring are a joint video restoration problem [11,2,36] that generates high-framerate clear videos from low-framerate blurry inputs. Recently, Shen et al. [44] propose a pyramid module and an inter-pyramid recurrent module to enhance the restoration quality and exploits the spatio-temporal information, respectively. Although these methods deal with motion blur and aliasing, spatial degradation is not considered as in STVSR. In addition, Pollak et al. [39] propose a deep internal learning approach by exploiting the recurrence within and across different spatio-temporal scales of the video. However, it is a zero-shot temporal-SR and is difficult to improve the SR performance without supervised information.

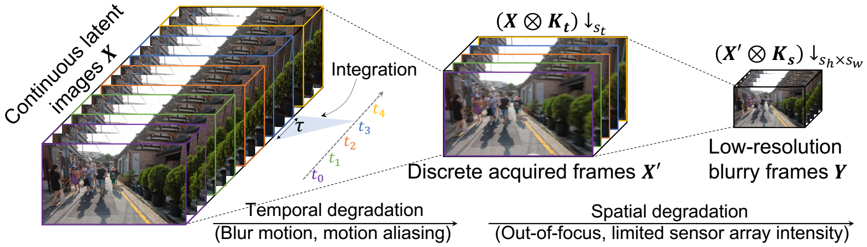

Continuous latent images X
Integration
τ
t_0
t_1
t_2
t_3
t_4
$(X \otimes K_t) \downarrow_{s_t}$
Discrete acquired frames X'
$(X' \otimes K_s) \downarrow_{s_h \times s_w}$
Low-resolution blurry frames Y
Temporal degradation
(Blur motion, motion aliasing)
Spatial degradation
(Out-of-focus, limited sensor array intensity)

Fig. 3: Illustrations of our video degradation. Note that we show temporal and spatial degradation separately for clarity, although they occur simultaneously. Camera sensors capture discrete frames at the time step $t_i (i \geq 0)$ by integrating the continuous latent images within an exposure time interval τ, leading to temporal degradation. Then, non-ideal imaging factors such as out-of-focus and limited sensor array intensity result in spatial degradation as well. These two degradations can be implemented by using a temporal kernel $\boldsymbol{K}_t$ and a spatial kernel $\boldsymbol{K}_s$ with the downsampling $\downarrow_{s_t}$ and $\downarrow_{s_h \times s_w}$.

3 Proposed Method

3.1 Video Degradation Model

We present the video degradation model for the practical STVSR task, as shown in Fig. 3. In general, a sequence of video frames is captured by a camera with a periodically on-and-off shutter [50,44]. When the shutter is open, the camera sensors collect reflected photons and convert them into electrical signals. This can be formulated as an integration of luminous intensity over the exposure time, during which the motion blur may occur if the object moves or the camera shakes. Besides, due to limited shutter on-and-off frequency (framerate), motion aliasing may also occur when the temporal dynamic event frequency is beyond the Nyquist limit of framerate. In addition to above temporal degradation, video capturing also suffers from similar spatial degradation to single image capturing as a result of non-ideal imaging factors such as out-of-focus and limited sensor array intensity [24]. Formally, given a high spatio-temporal resolution (HSTR) video $\boldsymbol{X} \in \mathbb{R}^{T_h \times H_h \times W_h \times 3}$, a 3D blur kernel $\boldsymbol{K}$, a low spatio-temporal resolution (LSTR) video $\boldsymbol{Y} \in \mathbb{R}^{T_l \times H_l \times W_l \times 3}$ can be formulated as

$$\boldsymbol{Y} = (\boldsymbol{X} \otimes \boldsymbol{K}) \downarrow_{s_t \times s_h \times s_w} + \boldsymbol{N}, \tag{1}$$

where $\otimes$ represents the 3D convolution, and $\downarrow_{s_t \times s_h \times s_w}$ (abbreviated as $\downarrow_s$ in the rest of the paper for clarity) denotes the standard s-fold downsampling in three directions: temporal, vertical and horizontal directions. $\boldsymbol{N}$ is often assumed to be the additive white Gaussian noise with a noise level of σ. In addition, the sizes of $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy $T_h = s_t T_l, H_h = s_h H_l$ and $W_h = s_w W_l$. In fact, some popular video restoration tasks, including spatial VSR, VFI and VD, *etc*, can be seen as special cases of the above degradation model. Note that Eq. (1) can also be used for bicubic degradation since it can be approximated by blur and downsampling with a center shift [59].

3.2 Problem Setting and Optimization Difficulty

Problem formulation. The goal of practical STVSR is to solve the joint video restoration tasks, including video frame interpolation, video deblurring and super-resolution. Specifically, given a low-framerate and low-resolution (low spatio-temporal resolution) blurry video $\boldsymbol{Y}$, a 3D blur kernel $\boldsymbol{K}$ and downsampling $\downarrow_s$, we propose to restore a high-framerate and high-resolution (high spatio-temporal resolution) video $\boldsymbol{X}$. According to the Maximum A Posteriori (MAP) framework, we solve the problem by minimizing the energy function $E(\boldsymbol{X})$,

$$\widehat{\boldsymbol{X}} = \arg\min_{\boldsymbol{X}} E(\boldsymbol{X}) := \frac{1}{2\sigma^2} \underbrace{\|\boldsymbol{Y} - (\boldsymbol{X} \otimes \boldsymbol{K}) \downarrow_s \|^2}_{\text{data fidelity term}} + \lambda \underbrace{\Phi(\boldsymbol{X})}_{\text{prior term}} , \qquad (2)$$

where λ is a trade-off parameter, the data fidelity term is associated with the model likelihood for reconstruction, and the prior term is a regularization which is related to the prior information of the high spatio-temporal resolution video. However, the prior term is often unknown in practice, and thus it is intractable to directly compute an analytical solution to Problem (2). Compared with classic image SR problem, our task is more challenging since the high spatio-temporal resolution video lose high-frequency details in space and time.

Alternating optimization. Based on the Half-Quadratic Splitting (HQS) algorithm [1,59], we introduce an auxiliary variable $\boldsymbol{Z}$ that is close to $\boldsymbol{X}$, and we have a regularization $\|\boldsymbol{Z} - \boldsymbol{X}\|^2$ with a penalty parameter μ. Then, we reformulate Problem (2) as the following optimization problem:

$$E(\boldsymbol{X}, \boldsymbol{Z}) = \frac{1}{2\sigma^2} \|\boldsymbol{Y} - (\boldsymbol{Z} \otimes \boldsymbol{K}) \downarrow_s \|^2 + \lambda \Phi(\boldsymbol{X}) + \frac{\mu}{2} \|\boldsymbol{Z} - \boldsymbol{X}\|^2. \qquad (3)$$

Then, Problem (3) can be solved by alternately optimizing two sub-problems Eq. (4) (for $\boldsymbol{Z}$) and Eq. (5) (for $\boldsymbol{X}$) as follows

$$\begin{cases} \boldsymbol{Z}_k = \arg\min_{\boldsymbol{Z}} \|\boldsymbol{Y} - (\boldsymbol{Z} \otimes \boldsymbol{K}) \downarrow_s \|^2 + \mu\sigma^2 \|\boldsymbol{Z} - \boldsymbol{X}_{k-1}\|^2, & (4) \\ \boldsymbol{X}_k = \arg\min_{\boldsymbol{X}} \frac{\mu}{2} \|\boldsymbol{Z}_k - \boldsymbol{X}\|^2 + \lambda \Phi(\boldsymbol{X}). & (5) \end{cases}$$

With the help of the alternating optimization, we can calculate the closed-form solution to the sub-problem (4), and solve the sub-problem (5) as a video denoising problem. However, directly finding an analytic solution is time-consuming since it requires the inversion of a high dimensional matrix, whose computational complexity is $O(T_h^3 W_h^3 H_h^3)$. One can use simulation-based methods (*e.g.*, Markov Chain Monte Carlo [10]) to solve the problem, but it would still be computationally expensive for large videos. Inspired by existing image super-resolution methods [59,9,37], we take two additional challenging video problems, *i.e.*, dynamic blur removal and frame interpolation, into consideration, and derive a new analytical solution for the practical STVSR. Note that it is not a trivial issue for these methods to handle video problems because the practical STVSR suffers from more complex degradation such as motion blur and aliasing.

3.3 Analytical Solution

To solve the sub-problem (4), we propose to derive a theorem to compute an analytical solution for the practical STVSR problem. To develop the theorem, we introduce the Fourier transform to efficiently exploit intrinsic properties of the downsampling and the blur kernel in the frequency domain.

Theorem 1 *Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ be the fast Fourier transform (FFT) and inverse FFT, and $\overline{\mathcal{F}}$ be the complex conjugate of $\mathcal{F}$. Assume the blur kernel $\boldsymbol{K}$ and the downsampling $\downarrow_s$ satisfy some properties [61]. Given a video $\boldsymbol{X}_{k-1}$ at the k-th iteration and a low-resolution video $\boldsymbol{Y}$, the solution to Eq. (4) can be computed using the following closed-form expression [1], i.e.,*

$$\boldsymbol{Z}_k = \mathcal{F}^{-1}\left(\frac{1}{\alpha_k}\left(\mathcal{F}(\boldsymbol{R}_{k-1}) - \overline{\mathcal{F}(\boldsymbol{K})}\left(\frac{(\mathcal{F}(\boldsymbol{K})\mathcal{F}(\boldsymbol{R}_{k-1}))\downarrow_s^a}{\left(s\alpha_k\boldsymbol{I} + \left(\mathcal{F}(\boldsymbol{K})\overline{\mathcal{F}(\boldsymbol{K})}\right)\downarrow_s^a\right)}\right)\Big\uparrow_s^r\right)\right), \quad (6)$$

where $\boldsymbol{R}_{k-1} = \overline{\mathcal{F}(\boldsymbol{K})}\mathcal{F}(\boldsymbol{Y}\uparrow_s) - \alpha_k\boldsymbol{X}_{k-1}$ with $\alpha_k = \mu_k\sigma^2$, $\uparrow_s$ is a standard s-fold upsampler , i.e., upsampling the spatial size by filling the new entries with zeros, $\uparrow_s^r$ is an upsampler by repeating the tensor the desired dimension, and $\downarrow_s^a$ is a distinct block downsampler, i.e., averaging the $s_t \times s_h \times s_w$ distinct blocks.

In Theorem 1, we are able to derive an analytical solution to the sub-problem (4). Note that the assumptions are not strong and they are widely used in existing studies [40,45]. For example, the assumption of the blur kernel does not depend on the shape and it can be used in different kinds of blurring, such as motion blur and out-of-focus blur [61]. Different from USRNet [59], our theorem is more general and can be applied to more image or video restoration tasks, including classic image super-resolution, video frame interpolation, video deblurring and the space-time video super-resolution. Based on the analytical solution, we are able to further solve the next sub-problem (5).

Complexity analysis. With the help of Theorem 1, we further analyze the complexity of calculating the analytical solution (6) and show that we are able to improve the effectiveness of computing the analytical solution to sub-problem (4). In the theorem, Eq. (6) requires three FFT computations and one inverse FFT computation, which are the most expensive parts in the implementation. Considering the computation complexities of FFT and inverse FFT, the computation complexity of Eq. (6) is $\mathcal{O}(T_hW_hH_h\log(T_hW_hH_h))$, which is much smaller than the computation complexity of directly solving Eq. (2) (*i.e.*, $\mathcal{O}(T_h^3W_h^3H_h^3)$) and can be computed efficiently on the modern GPU devices. More analyses can be found in the supplementary materials. With the efficient calculation of $\boldsymbol{Z}_k$, we can deal with space-time blur (including motion and spatial blur, encoded in $\boldsymbol{K}$) and space-time downsampling (including temporal and spatial downsampling, encoded in $\downarrow_s$) in a joint and analytical way. The blur is reduced and the details are restored gradually.

[1] Please see the detailed proof in the supplementary materials.

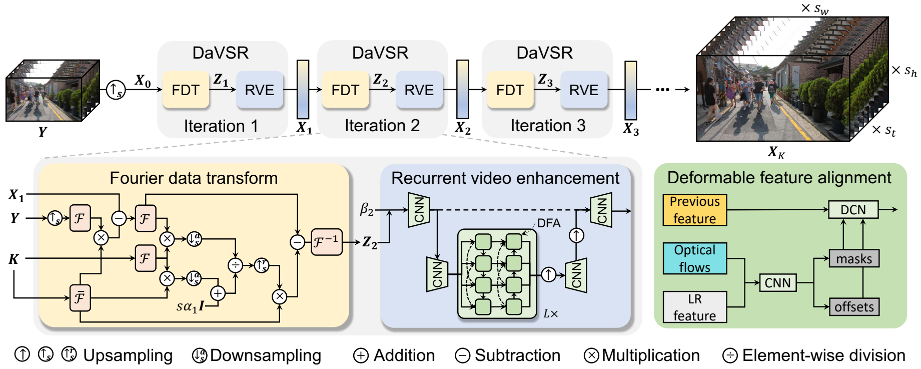


Fig. 4: The overall architecture of the proposed method with K iterations. Our one-stage model is able to handle different video degradation (*i.e.*, Eq. (1)) in a joint way and synthesize HSTR frames by taking an LSTR blur video $\boldsymbol{Y}$, scale factor s and blur kernel $\boldsymbol{K}$ as inputs. Specifically, the architecture consists of two main modules, including the FDT layer that reduces the blur degradation, and the RVE layer that makes HR synthesized videos cleaner.

3.4 Deep Alternating Video Super-Resolution Network

In this paper, we propose to design a new video super-resolution network based on the alternating optimization, called DAVSR, which solves the sub-problems (4) and (5) alternately. To this end, we propose a Fourier data transform layer $\mathcal{T}$ and a recurrent video enhancement layer $\mathcal{R}$ to address the above two sub-problems, respectively. The overall architecture of our method is shown in Fig. 4. Specifically, the Fourier data transform layer aims to alleviate the video degradation, while the recurrent video enhancement layer aims to enhance synthesized videos by adding more high-frequency details.

Fourier data transform layer. With the help of the analytical solution (4), we aim to reduce the video degradation from the LSTR blurry video. During the optimization, we propose to find a clearer HSTR video such that it minimizes a weighted combination of the data fidelity term and the quadratic regularization term. To this end, we propose a Fourier data transform (FDT) layer, as shown in Fig. 4. Specifically, given an LSTR blurry video $\boldsymbol{Y}$, the scale factor $\boldsymbol{s}{=}[s_t, s_h, s_w]$, blur kernel $\boldsymbol{K}$ and the parameters α_k, we calculate the video by using the analytical solution Eq. (6), *i.e.*,

$$\boldsymbol{Z}_k = \mathcal{T}(\boldsymbol{X}_{k-1}|\boldsymbol{Y}, \boldsymbol{K}, \boldsymbol{s}, \alpha_k). \tag{7}$$

Noted that the Fourier data transform is a model-based method and it has no trainable parameters. In this sense, this module has good generalization and it is able to generate meaningful data. In addition, this layer is differentiable since every sub-operation is differentiable. Compared with USRNet [59], this layer helps simultaneously synthesize high-frequency information in space and time.

Algorithm 1: Deep Alternating Video Super-Resolution

```
Input: LSTR video $\boldsymbol{Y}$, blur kernel $\boldsymbol{K}$, scale factor $\boldsymbol{s}$, parameters $\alpha_k, \beta_k$
Initialize number of iterations $K$, and initialize $\boldsymbol{X}_0$ on $\boldsymbol{Y}$ in space and time;
while not convergent do
    for $k \leftarrow 1$ to $K$ do
        Update $\boldsymbol{Z}_k$ by computing $\boldsymbol{Z}_k = \mathcal{T}(\boldsymbol{X}_{k-1}|\boldsymbol{Y}, \boldsymbol{K}, \boldsymbol{s}, \alpha_k)$;
        Update $\boldsymbol{X}_k$ by computing $\boldsymbol{X}_k = \mathcal{R}(\boldsymbol{Z}_k|\beta_k)$;
    Update the RVE model $\mathcal{R}$ by minimize the training loss (11).
```

Recurrent video enhancement layer. For the sub-problem (5), we propose a recurrent video enhancement (RVE) layer to enhance the quality of videos and restore high-frequency sequential textures. Such sequential information is important in the continuous video frames, which, however, is neglected in the FDT layer and used in feature alignment. To address this, the RVE layer aims to model the sequential dependency and align the features of video frames. Specifically, given a video $\boldsymbol{Z}_k$ transformed by the FDT layer and the noise level β_k, the RVE model $\mathcal{R}$ restores a cleaner HSTR video, *i.e.*,

$$\boldsymbol{X}_k = \mathcal{R}(\boldsymbol{Z}_k|\beta_k), \quad \text{where } \beta_k = \sqrt{\lambda/\mu_k}. \tag{8}$$

Note that RVE is a learning-based model, and it is implemented by using deformable feature alignment (DFA) module, motivated by [7]. Specifically, we first use convolutional layers to extract low-level features $\{\boldsymbol{g}_k^1, \ldots, \boldsymbol{g}_k^{T_h}\}$ from the concatenation of $\boldsymbol{Z}_k$ and β_k. Then, we use DFA $\mathcal{G}$ to propagate and align the features $\widetilde{\boldsymbol{z}}_{k,j}^i$, and further use residual blocks $\mathcal{C}$ to fuse the features $\widehat{\boldsymbol{z}}_{k,j}^i$, *i.e.*,

$$\widetilde{\boldsymbol{z}}_{k,j}^i = \mathcal{G}\left(\boldsymbol{g}_k^i, \widehat{\boldsymbol{z}}_{k,j}^{i-1}, \widehat{\boldsymbol{z}}_{k,j}^{i-2}, \boldsymbol{f}_k^{i\to i-1}, \boldsymbol{f}_k^{i\to i-2}\right), \tag{9}$$

$$\widehat{\boldsymbol{z}}_{k,j}^i = \widetilde{\boldsymbol{z}}_{k,j}^i + \mathcal{C}\left(\left[\widehat{\boldsymbol{z}}_{k,j-1}^i; \widetilde{\boldsymbol{z}}_{k,j}^i\right]\right), \tag{10}$$

where $\widehat{\boldsymbol{z}}_{k,j}^i$ is the feature at the i-th timestep in the j-th propagation branch at the k-th iteration, $\widehat{\boldsymbol{z}}_{k,0}^i = \boldsymbol{g}_k^i$, and $\boldsymbol{f}_k^{i_1 \to i_2}$ is the optical flow from i_1-th frame to the i_2-th frame at the k-th iteration, $[\cdot;\cdot]$ is a concatenation along the channel dimension. More details of $\mathcal{G}$ can be found in the supplementary. Last, we use convolutional layers to reconstruct $\boldsymbol{x}_k^i$ which is the i-th element of $\boldsymbol{X}_k$.

Loss function and algorithm. In the training, our goal is to train the model to minimize the distance between the HSTR video $\boldsymbol{X}_K$ at the last iteration and the ground-truth video $\boldsymbol{X}$. To this end, we use the following Charbonnier loss [8] because it can handle outliers, *i.e.*,

$$\mathcal{L} = \sqrt{\|\boldsymbol{X}_K - \boldsymbol{X}\|^2 + \epsilon^2}, \quad \text{where } \epsilon = 1\times 10^{-3}. \tag{11}$$

Algorithm 1 shows the detailed model optimization process. We alternately optimize two sub-problems (4) and (5), and update the RVE model by minimize the loss (11). Note that the FDT and RVE modules are differentiable during training. At the K-th iteration, the algorithm outputs the HSTR video.

4 Experiments

Implementation details. We use Adam optimizer [19] with β_1=0.9, β_2=0.99, and use Cosine Annealing [29] to decay the learning rate of the main network from 1×10^{-4} to 10^{-7}. In the FAT layer, we use `torch.fft.fftn` and `torch.fft.ifftn` to calculate FFT and inverse FFT operators, respectively. The total number of iterations for training is 300K, and the number of iterations K for updating $\boldsymbol{Z}$ and $\boldsymbol{X}$ is 3. The batch size is 4 and the patch size of input LR frames is 64×64. The number of residual blocks for each branch is 7, and the number of feature channels is 64.

Datasets and evaluation metrics. REDS [32] is a realistic and dynamic scenes dataset, whose the version of "blur_bicubic" can be used for the practical time-space video super-resolution task. The dataset is synthesized by averaging five subsequent frames, and then downsampling these frames with 4× bicubic kernel. REDS contains 266 training clips (each with 100 LR frames and 500 GT frames) and 4 testing clips (000, 011, 015 and 020, denoted as REDS4) that have diverse scenes and motions. In addition, another widely used dataset Vid4 [25] is also used in our experiments. The same degradation model as REDS is used for Vid4 experiments. For the evaluation metrics, we use PSNR and SSIM [54] to measure the performance of VSR methods.

4.1 Comparison with State-of-the-art Methods

To achieve STVSR, we compare with the following SOTA methods, including single-image SR model (*i.e.*, SwinIR [22]), two VFI approaches (*i.e.*, SuperSloMo [15] and DAIN [3]), two deblurring methods (*i.e.*, EDVR [53] and CDVD-TSP [38]), and four recent VSR models, EDVR [53], BasicVSR [6], IconVSR [6], and BasicVSR++ [7]. In addition, we also consider joint video restoration methods, including Zooming Slow-Mo [55] and BIN [44].

Quantitative comparison. From Tables 1 and 2, we have the following observations. (1) Our method outperforms the two-stage and three-stage methods by a large margin on REDS4 and Vid4. Especially for fast motion videos (*e.g.*, clips 011 and 020), our method has the more significant improvements than these methods. It suggests that our one-stage network is able to exploit diverse spatio-temporal patterns and simultaneously handle VFI, VD and VSR. (2) The SOTA image SR method (*i.e.*, SwinIR [22]) performs worse than other VSR methods since it cannot handle sequential information of videos and cannot remove the motion blur. (3) The two-stage framework (*e.g.*, SuperSloMo+BasicVSR++*) is better than the three-stage framework because the reconstruction error propagates severely along the stages. (4) The performance of two-stage and three-stage methods are influenced by different motions. REDS4 has diverse scenes and motions, *e.g.*, Clip_000 has small motions and other video clips has more complex motions. Moreover, the VFI methods are sensitive to large motions. Thus, the performance of SuperSloMo+BasicVSR++* is better than that of our method on Clip_000, but is worse on other clips due to large motions.

Table 1: Quantitative comparison of our results and two-stage/three-stage on REDS4. **Red** and **blue** indicates the best and the second best performance. The superscript * means that the model is trained on the "blur_bicubic" version of REDS, and the superscript † means the pre-trained video deblurring model.

Methods	Clip_000		Clip_011		Clip_015		Clip_020		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic+linear interpolation	23.57	0.606	21.75	0.587	25.59	0.739	20.19	0.564	22.78	0.624
BIN+EDVR	25.02	0.709	22.00	0.603	25.89	0.765	21.15	0.572	23.52	0.662
BIN+BasicVSR	25.69	0.716	22.49	0.622	26.33	0.789	21.74	0.583	24.06	0.678
BIN+IconVSR	25.80	0.735	22.65	0.632	26.52	0.792	21.99	0.590	24.24	0.687
BIN+BasicVSR++	26.01	0.752	22.82	0.638	26.69	0.799	22.23	0.654	24.44	0.711
CDVD-TSP+Zooming Slow-Mo	26.74	0.786	24.68	0.700	29.52	0.859	23.02	0.689	25.99	0.759
EDVR+Zooming Slow-Mo	26.89	0.791	24.72	0.710	29.68	0.861	23.19	0.695	26.12	0.764
DAIN+CDVD-TSP+SwinIR	25.00	0.705	24.23	0.689	28.22	0.812	22.59	0.672	25.01	0.720
DAIN+CDVD-TSP+EDVR	25.95	0.751	24.14	0.686	28.84	0.839	22.61	0.673	25.38	0.737
DAIN+CDVD-TSP+BasicVSR	26.32	0.777	24.23	0.688	29.11	0.849	22.79	0.681	25.61	0.748
DAIN+CDVD-TSP+IconVSR	26.40	0.782	24.37	0.691	29.32	0.855	22.86	0.684	25.74	0.753
DAIN+CDVD-TSP+BasicVSR++	26.58	0.793	24.45	0.691	29.57	0.860	23.02	0.689	25.90	0.758
DAIN+EDVR†+SwinIR	25.56	0.719	24.35	0.688	28.98	0.839	22.83	0.672	25.43	0.730
DAIN+EDVR†+EDVR	26.50	0.774	24.25	0.685	29.51	0.855	22.85	0.674	25.78	0.747
DAIN+EDVR†+BasicVSR	26.66	0.783	24.21	0.683	29.61	0.858	22.76	0.670	25.81	0.748
DAIN+EDVR†+IconVSR	26.75	0.787	24.25	0.685	29.77	0.863	22.80	0.672	25.89	0.752
DAIN+EDVR†+BasicVSR++	26.91	0.795	24.25	0.685	29.94	0.868	22.82	0.673	25.98	0.755
DAIN+EDVR*	26.29	0.760	24.53	0.699	29.32	0.848	22.96	0.682	25.77	0.747
DAIN+BasicVSR*	26.82	0.790	24.71	0.709	29.65	0.861	23.15	0.693	26.08	0.763
DAIN+IconVSR*	26.95	0.796	24.89	0.713	29.89	0.867	23.25	0.695	26.24	0.768
DAIN+BasicVSR++*	27.34	0.814	25.11	0.719	30.22	0.874	23.46	0.703	26.53	0.778
SuperSloMo+CDVD-TSP+SwinIR	25.26	0.709	24.20	0.686	28.82	0.835	22.76	0.675	25.26	0.726
SuperSloMo+CDVD-TSP+EDVR	26.22	0.753	24.18	0.686	29.32	0.845	22.75	0.674	25.62	0.739
SuperSloMo+CDVD-TSP+BasicVSR	26.70	0.779	24.29	0.689	29.57	0.855	22.97	0.683	25.88	0.752
SuperSloMo+CDVD-TSP+IconVSR	26.81	0.785	24.42	0.691	29.85	0.862	23.06	0.686	26.04	0.756
SuperSloMo+CDVD-TSP+BasicVSR++	27.06	0.797	24.53	0.692	30.12	0.868	23.24	0.691	26.24	0.762
SuperSloMo+EDVR†+SwinIR	25.86	0.726	24.32	0.687	29.36	0.844	22.87	0.672	25.60	0.732
SuperSloMo+EDVR†+EDVR	26.98	0.784	24.27	0.685	30.04	0.862	22.92	0.674	26.05	0.751
SuperSloMo+EDVR†+BasicVSR	27.12	0.793	24.24	0.683	30.07	0.864	22.82	0.670	26.06	0.753
SuperSloMo+EDVR†+IconVSR	27.24	0.798	24.27	0.685	30.28	0.870	22.86	0.672	26.16	0.756
SuperSloMo+EDVR†+BasicVSR++	27.41	0.805	24.28	0.685	30.46	0.874	22.88	0.673	26.26	0.759
SuperSloMo+EDVR*	26.66	0.769	24.60	0.700	29.83	0.855	23.11	0.684	26.05	0.752
SuperSloMo+BasicVSR*	27.35	0.803	24.78	0.710	30.16	0.868	23.35	0.696	26.41	0.769
SuperSloMo+IconVSR*	**27.52**	0.810	24.95	0.714	30.48	0.876	23.46	0.699	26.60	0.775
SuperSloMo+BasicVSR++*	**28.00**	**0.829**	**25.19**	**0.720**	**30.83**	**0.883**	**23.70**	**0.707**	**26.93**	**0.785**
DAVSR (Ours)	27.23	**0.820**	**30.15**	**0.865**	**31.05**	**0.894**	**28.06**	**0.858**	**29.12**	**0.859**

Table 2: Quantitative comparison of our results and two-stage/three-stage methods on Vid4. **Red** and **blue** indicates the best and the second best performance.

Methods	Calendar		City		Foliage		Walk		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DAIN+CDCD-TSP+BasicVSR++	17.05	0.480	21.64	0.421	19.01	0.340	19.68	0.668	19.35	0.477
DAIN+EDVR†+BasicVSR++	17.35	0.497	21.80	0.433	19.22	0.350	19.91	0.687	19.57	0.492
DAIN+BasicVSR++*	18.05	0.522	**22.22**	**0.439**	20.08	**0.400**	**20.33**	**0.692**	20.17	0.513
SuperSloMo+CDCD-TSP+BasicVSR++	17.11	0.488	21.70	0.427	19.09	0.344	19.78	0.672	19.42	0.483
SuperSloMo+EDVR†+BasicVSR++	17.59	0.505	21.78	0.432	19.22	0.346	19.86	0.687	19.61	0.493
SuperSloMo+BasicVSR++*	**18.30**	**0.530**	22.21	0.439	**20.09**	0.398	20.30	0.692	**20.22**	**0.515**
DAVSR (Ours)	**22.09**	**0.747**	**25.18**	**0.731**	**23.96**	**0.680**	**24.18**	**0.827**	**23.85**	**0.746**

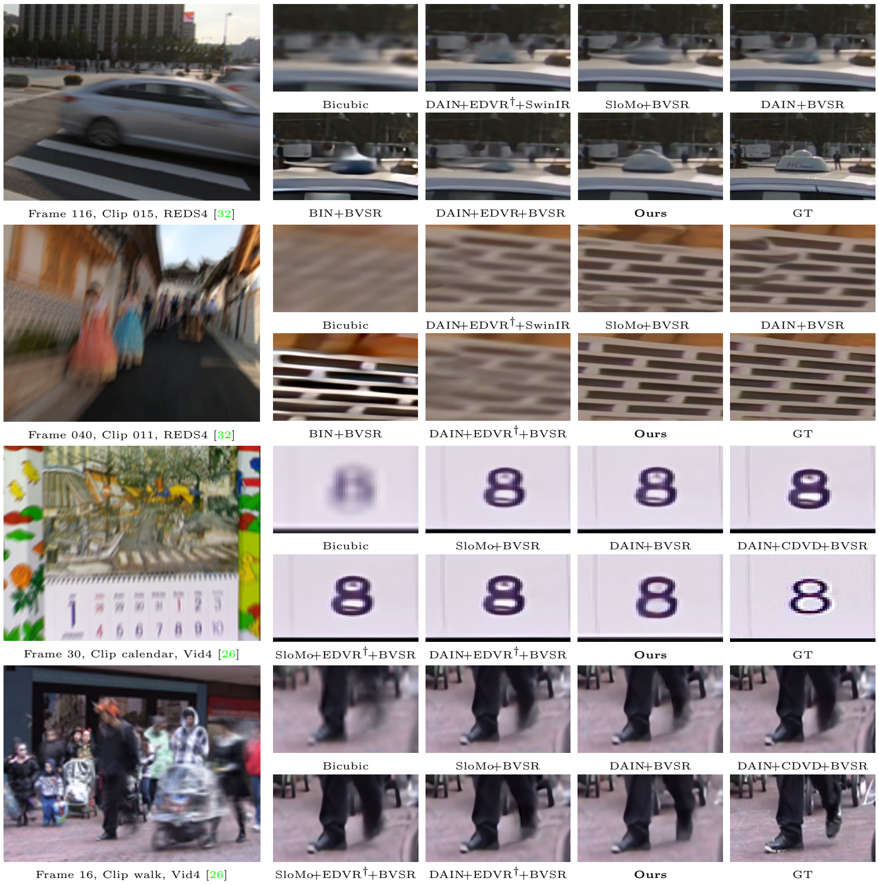


Fig. 5: Visual results of different methods on REDS4 and Vid4. Due to the space limitations of this figure, CDCD-TSP, SuperSloMo and BasicVSR++ are abbreviated as CDCD, SloMo and BVSR, respectively.

Qualitative comparison. Visual results of different methods are shown in Figure 5. In this figure, our proposed method achieves significant visual improvements over two-stage and three-stage methods. This suggests that our one-stage framework is able to learn spatio-temporal information by exploring the natural intra-relatedness among the video interpolation, deblurring and super-resolution tasks. Compared with multi-stage methods, our method is able to synthesize high-quality HR video frames with clearer details and fewer blurring artifacts even for fast motion video sequences. Taking the second line (Frame 040, Clip 011) as an example, our method is able to restore textures which are very close to GT although Clip_011 has large motion. In contrast, multi-stage networks based on VFI methods (*i.e.*, Super-SloMo [15] and DAIN [3]) tend to suffer from severe motion blur in the synthesized HR video, because it is difficult for the VFI methods to handle large motions in the videos. More visual comparison results can be found in the supplementary materials.

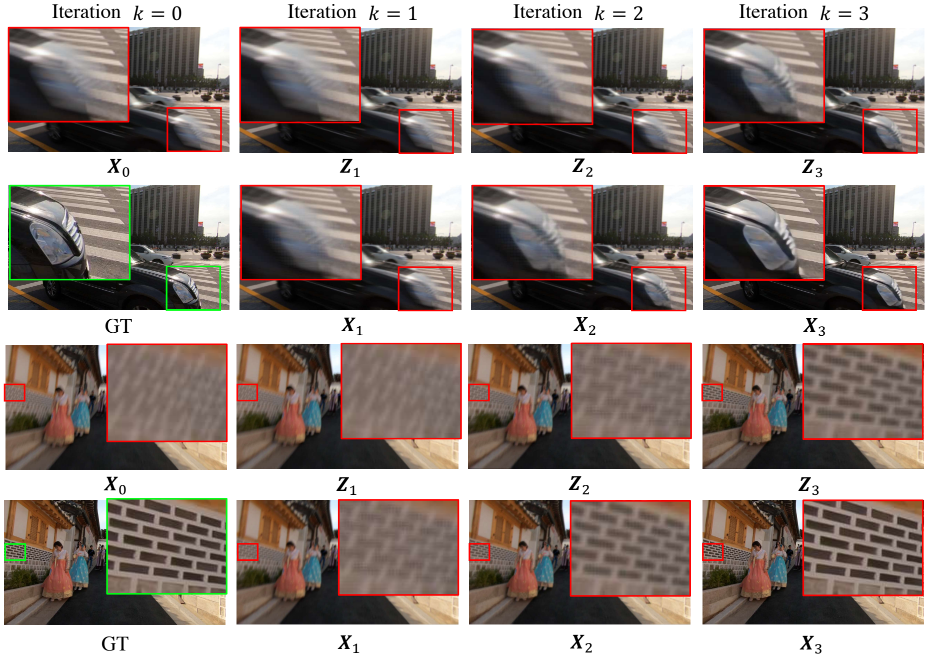


Fig. 6: Visualization of FDT and RVE at different iterations on REDS4.

4.2 Visualization on Different Iterations

It is very interesting to investigate the synthesized outputs of the Fourier data transform (FDT) layer and the recurrent video enhancement (RVE) layer at different iterations. The visualization results of our method at different iterations are provided in Fig. 6. As one can see, the quality of synthesized video frames continuously improves as the iteration number increases. This shows that, given an LSTR blurry video frame, the FDT and RVE layers are able to cooperatively and alternately deblur and recover high-frequency details of video frames. Specifically, the FDT layer eliminates blur kernel induced degradation and recovers tiny structures and fine textures, then the RVE layer restores the high-frequency information in videos. The visualization results demonstrate that our proposed DAVSR provides an interpretable way to understand STVSR.

4.3 Model Size and Inference Time

We investigate model sizes and runtime of different networks, and the results are shown in Fig. 2. To synthesize HSTR clear frames, the composed two-stage or three-stage methods can lead to very large model sizes for frame reconstruction. In contrast, our one-stage model has much fewer parameters than the SOTA two-stage and three-stage networks. For example, it is more than 3× smaller than SuperSloMo+BasicVSR++*. With the help of the smaller model size, our model also achieves a fast inference time. More comparison results of model size and inference time can be found in the supplementary materials.

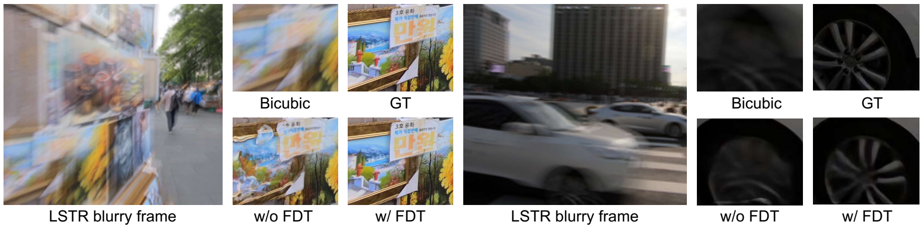


Fig. 7: Visual results of the RDT layer on the REDS4 dataset.

4.4 Ablation Study

We have already verified the superiority of our one-stage framework over two-stage and three-stage frameworks. To further demonstrate the effectiveness of our network, we propose to conduct a comprehensive ablation study on the Fourier data transform (FDT) layer and the recurrent video enhancement (RVE) layer. As shown in Table 3, the model without the FDT or RVE layer is worse than original model. Next, we show the visual results of the ablation study in Fig. 7. Here, we only conduct an ablation study on FDT because it contains no trainable parameters. If only the RVE layer is adopted without FDT, the synthesized SR video frames may suffer from blur artifacts. On the other hand, if the data module is used without the RVE layer, the network cannot recover the high-frequency textures. This suggests that the FDT and EVD layers are complementary to each other and both are important to the proposed alternating optimization-based model.

Table 3: Ablation study on the FDT and EVD layers. Here we use PSNR as the evaluation metric.

FDT	✗	✓	✓
EVD	✓	✗	✓
PSNR	27.25	23.05	**29.12**

5 Conclusions

In this paper, we propose an alternating optimization for the practical space-time video super-resolution (STVSR) task by leveraging the model-based methods and learning-based methods. From an interpretable point of view, we first formulate the STVSR problem as two sub-problems related to the motion blur and motion aliasing. Specifically, we provide an analytical solution and propose a Fourier data transform layer to reduce the motion blur and motion aliasing for the first sub-problem. Then, we propose a recurrent video enhancement layer in the second sub-problem to enhance the quality of the synthesized video. By the alternating optimization, our method is able to jointly handle the video interpolation, video deblurring and video super-resolution. Extensive experiments demonstrate that our framework achieves the state-of-the-art performance and it is more effective yet efficient than existing multi-stage networks.

Acknowledgements. This work was partly supported by Huawei Fund and the ETH Zürich Fund (OK).

References

1. Afonso, M.V., Bioucas-Dias, J.M., Figueiredo, M.A.: Fast image recovery using variable splitting and constrained optimization. IEEE Transactions on Image Processing **19**(9), 2345–2356 (2010)
2. Argaw, D.M., Kim, J., Rameau, F., Kweon, I.S.: Motion-blurred video interpolation and extrapolation. In: AAAI Conference on Artificial Intelligence. pp. 901–910 (2021)
3. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3703–3712 (2019)
4. Bao, W., Lai, W.S., Zhang, X., Gao, Z., Yang, M.H.: Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(3), 933–948 (2019)
5. Caballero, J., Ledig, C., Aitken, A., Acosta, A., Totz, J., Wang, Z., Shi, W.: Real-time video super-resolution with spatio-temporal networks and motion compensation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4778–4787 (2017)
6. Chan, K.C., Wang, X., Yu, K., Dong, C., Loy, C.C.: Basicvsr: The search for essential components in video super-resolution and beyond. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4947–4956 (2021)
7. Chan, K.C., Zhou, S., Xu, X., Loy, C.C.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5972–5981 (2022)
8. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: International Conference on Image Processing. vol. 2, pp. 168–172 (1994)
9. Chiche, B.N., Frontera-Pons, J., Woiselle, A., Starck, J.L.: Deep unrolled network for video super-resolution. In: International Conference on Image Processing Theory, Tools and Applications. pp. 1–6 (2020)
10. Gilavert, C., Moussaoui, S., Idier, J.: Efficient gaussian sampling for solving large-scale inverse problems using mcmc. IEEE Transactions on Signal Processing **63**(1), 70–80 (2014)
11. Gupta, A., Aich, A., Roy-Chowdhury, A.K.: Alanet: Adaptive latent attention network forjoint video deblurring and interpolation. arXiv preprint arXiv:2009.01005 (2020)
12. Haris, M., Shakhnarovich, G., Ukita, N.: Recurrent back-projection network for video super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3897–3906 (2019)
13. Hyun Kim, T., Mu Lee, K.: Generalized video deblurring for dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5426–5434 (2015)
14. Hyun Kim, T., Mu Lee, K., Scholkopf, B., Hirsch, M.: Online video deblurring via dynamic temporal blending network. In: IEEE International Conference on Computer Vision. pp. 4038–4047 (2017)
15. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9000–9008 (2018)

16. Jin, M., Meishvili, G., Favaro, P.: Learning to extract a video sequence from a single motion-blurred image. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6334–6342 (2018)
17. Jo, Y., Oh, S.W., Kang, J., Kim, S.J.: Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3224–3232 (2018)
18. Kim, T.H., Nah, S., Lee, K.M.: Dynamic video deblurring using a locally adaptive blur model. IEEE Transactions on Pattern analysis and Machine Intelligence **40**(10), 2374–2387 (2017)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
20. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8183–8192 (2018)
21. Liang, J., Cao, J., Fan, Y., Zhang, K., Ranjan, R., Li, Y., Timofte, R., Van Gool, L.: Vrt: A video restoration transformer. arXiv preprint arXiv:2201.12288 (2022)
22. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: IEEE International Conference on Computer Vision. pp. 1833–1844 (2021)
23. Liang, J., Fan, Y., Xiang, X., Ranjan, R., Ilg, E., Green, S., Cao, J., Zhang, K., Timofte, R., Van Gool, L.: Recurrent video restoration transformer with guided deformable attention. arXiv preprint arXiv:2206.02146 (2022)
24. Liang, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Mutual affine network for spatially variant kernel estimation in blind image super-resolution. In: IEEE Conference on International Conference on Computer Vision. pp. 4096–4105 (2021)
25. Liu, C., Sun, D.: On bayesian adaptive video super resolution. IEEE Transactions on Pattern analysis and Machine Intelligence **36**(2), 346–360 (2013)
26. Liu, C., Sun, D.: On bayesian adaptive video super resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(2), 346–360 (2013)
27. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: IEEE International Conference on Computer Vision. pp. 4463–4471 (2017)
28. Long, G., Kneip, L., Alvarez, J.M., Li, H., Zhang, X., Yu, Q.: Learning image matching by simply watching video. In: European Conference on Computer Vision. pp. 434–450 (2016)
29. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
30. Meyer, S., Wang, O., Zimmer, H., Grosse, M., Sorkine-Hornung, A.: Phase-based frame interpolation for video. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1410–1418 (2015)
31. Mudenagudi, U., Banerjee, S., Kalra, P.K.: Space-time super-resolution using graph-cut optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 995–1008 (2010)
32. Nah, S., Baik, S., Hong, S., Moon, G., Son, S., Timofte, R., Mu Lee, K.: Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
33. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3883–3891 (2017)

34. Nah, S., Son, S., Lee, K.M.: Recurrent neural networks with intra-frame iterations for video deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8102–8111 (2019)
35. Niklaus, S., Liu, F.: Context-aware synthesis for video frame interpolation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1701–1710 (2018)
36. Oh, J., Kim, M.: Demfi: Deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. arXiv preprint arXiv:2111.09985 (2021)
37. Pan, J., Bai, H., Dong, J., Zhang, J., Tang, J.: Deep blind video super-resolution. In: IEEE International Conference on Computer Vision. pp. 4811–4820 (2021)
38. Pan, J., Bai, H., Tang, J.: Cascaded deep video deblurring using temporal sharpness prior. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3043–3051 (2020)
39. Pollak Zuckerman, L., Naor, E., Pisha, G., Bagon, S., Irani, M.: Across scales & across dimensions: Temporal super-resolution using deep internal learning. arXiv e-prints pp. arXiv–2003 (2020)
40. Robinson, M.D., Toth, C.A., Lo, J.Y., Farsiu, S.: Efficient fourier-wavelet super-resolution. IEEE Transactions on Image Processing **19**(10), 2669–2681 (2010)
41. Sajjadi, M.S., Vemulapalli, R., Brown, M.: Frame-recurrent video super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6626–6634 (2018)
42. Shahar, O., Faktor, A., Irani, M.: Space-time super-resolution from a single video. IEEE (2011)
43. Shechtman, E., Caspi, Y., Irani, M.: Increasing space-time resolution in video. In: European Conference on Computer Vision. pp. 753–768 (2002)
44. Shen, W., Bao, W., Zhai, G., Chen, L., Min, X., Gao, Z.: Blurry video frame interpolation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5114–5123 (2020)
45. Šroubek, F., Kamenickỳ, J., Milanfar, P.: Superfast superresolution. In: IEEE International Conference on Image Processing. pp. 1153–1156 (2011)
46. Su, S., Delbracio, M., Wang, J., Sapiro, G., Heidrich, W., Wang, O.: Deep video deblurring for hand-held cameras. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1279–1288 (2017)
47. Takeda, H., Beek, P.v., Milanfar, P.: Spatiotemporal video upscaling using motion-assisted steering kernel (mask) regression. In: High-Quality Visual Experience, pp. 245–274. Springer (2010)
48. Tao, X., Gao, H., Liao, R., Wang, J., Jia, J.: Detail-revealing deep video super-resolution. In: IEEE International Conference on Computer Vision. pp. 4472–4480 (2017)
49. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8174–8182 (2018)
50. Telleen, J., Sullivan, A., Yee, J., Wang, O., Gunawardane, P., Collins, I., Davis, J.: Synthetic shutter speed imaging. In: Computer Graphics Forum. vol. 26, pp. 591–598 (2007)
51. Tian, Y., Zhang, Y., Fu, Y., Xu, C.: Tdan: Temporally-deformable alignment network for video super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3360–3369 (2020)

52. Wang, L., Guo, Y., Lin, Z., Deng, X., An, W.: Learning for video super-resolution through hr optical flow estimation. In: Asian Conference on Computer Vision. pp. 514–529 (2018)
53. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
54. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4) (2004)
55. Xiang, X., Tian, Y., Zhang, Y., Fu, Y., Allebach, J.P., Xu, C.: Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3370–3379 (2020)
56. Xiao, Z., Xiong, Z., Fu, X., Liu, D., Zha, Z.J.: Space-time video super-resolution using temporal profiles. In: ACM International Conference on Multimedia. pp. 664–672 (2020)
57. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**(8), 1106–1125 (2019)
58. Zamir, A.R., Wu, T.L., Sun, L., Shen, W.B., Shi, B.E., Malik, J., Savarese, S.: Feedback networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1308–1317 (2017)
59. Zhang, K., Gool, L.V., Timofte, R.: Deep unfolding network for image super-resolution. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3217–3226 (2020)
60. Zhang, K., Luo, W., Zhong, Y., Ma, L., Liu, W., Li, H.: Adversarial spatio-temporal learning for video deblurring. IEEE Transactions on Image Processing **28**(1), 291–301 (2018)
61. Zhao, N., Wei, Q., Basarab, A., Dobigeon, N., Kouamé, D., Tourneret, J.Y.: Fast single image super-resolution using a new analytical solution for l2-l2 problems. IEEE Transactions on Image Processing **25**(8), 3683–3697 (2016)
62. Zhou, S., Zhang, J., Pan, J., Xie, H., Zuo, W., Ren, J.: Spatio-temporal filter adaptive network for video deblurring. In: IEEE International Conference on Computer Vision. pp. 2482–2491 (2019)