Dual-Stream Fusion Network for Spatiotemporal Video Super-Resolution

Min-Yuan Tseng[†] Yen-Chung Chen[†] Yi-Lun Lee[†]
Wei-Sheng Lai[‡] Yi-Hsuan Tsai[§] Wei-Chen Chiu[†]
[†]National Chiao Tung University [‡]Google [§]NEC Labs America

Abstract

Visual data upsampling has been an important research topic for improving the perceptual quality and benefiting various computer vision applications. In recent years, we have witnessed remarkable progresses brought by the renaissance of deep learning techniques for video or image super-resolution. However, most existing methods focus on advancing super-resolution at either spatial or temporal direction, i.e, to increase the spatial resolution or the video frame rate. In this paper, we instead turn to discuss both directions jointly and tackle the spatiotemporal upsampling problem. Our method is based on an important observation that: even the direct cascade of prior research in spatial and temporal super-resolution can achieve the spatiotemporal upsampling, changing orders for combining them would lead to results with a complementary property. Thus, we propose a dual-stream fusion network to adaptively fuse the intermediate results produced by two spatiotemporal upsampling streams, where the first stream applies the spatial super-resolution followed by the temporal super-resolution, while the second one is with the reverse order of cascade. Extensive experiments verify the efficacy of the proposed method against several baselines. Moreover, we investigate various spatial and temporal upsampling methods as the basis in our two-stream model and demonstrate the flexibility with wide applicability of the proposed framework.

1. Introduction

Videos have been widely used to record memorable moments and entertainment in our daily life. Along with the advance of optical sensors and camera technology, the sensor resolution and video frame rate have become higher and higher to provide a better visual quality. However, when watching old film footage or videos made several years ago, one may easily experience unpleasant artifacts, such as blurry and low-resolution blocks, on contemporary displays. Hence, it is desired to increase the spatial resolution and the frame rate to achieve better viewing experience.

There are dozens of studies aiming at improving the visual quality of a video through increasing the spatial or temporal frequency. For example, video frame interpolation methods increase the frame rate (i.e., temporal frequency) of a video by synthesizing intermediate frames between two consecutive frames. On the other hand, image super-resolution (SR) methods increase the spatial resolution (i.e. spatial frequency) of an image by reconstructing a high-resolution (HR) version of its low-resolution (LR) counterpart, such that the resultant image looks sharper and more visually pleasing. Although image super-resolution methods can be applied to a video sequence in a frame-by-frame manner, the temporal coherence is left unexploited. Therefore, video super-resolution approaches take multiple LR frames into account to generate temporally consistent HR video frames. Nevertheless, both video frame interpolation and image/video super-resolution methods target to increase the frequency of videos along one of the directions (e.g., either temporal or spatial).

In this paper, we take one step further to address the spatiotemporal upsampling problem, where the goal is to simultaneously upsample a video in both the spatial and temporal domains. For simplicity, we consider upscaling both the spatial and temporal resolutions by $2\times$. Given a video sequence with N LR frames, our goal is to generate a $2\times$ spatial resolution HR video with $2N-1$ frames. The spatiotemporal upsampling can be achieved through a cascade of spatial upsampling and temporal upsampling, and vice versa. In this work, we analyze these two approaches (i.e., spatial upsampling followed by temporal upsampling, and temporal upsampling followed by spatial upsampling) and discover their complementary property on complex motion area. We then propose a dual-stream fusion framework to adaptively merge and refine the results from the two spatiotemporal upsampling streams. Our method takes advantage from both streams to reconstruct intermediate HR frames with better visual quality. In particular, the proposed method can be easily integrated with any off-the-shelf CNN-based spatial and temporal upsampling models. Finally, we demonstrate that the proposed method performs favorably against the baselines and its variants.

2. Related work

The spatial and temporal upsampling approaches have been widely studied for several decades. Here we focus our discussion on recent learning-based algorithms.

2.1. Spatial Upsampling

Several single-image super-resolution methods based on deep CNNs [6] have been proposed in recent years. A large amount of effort focuses on learning effective deep features by exploring advanced network architectures, including the residual learning [14, 22], recursive layers [15], progressive upsampling [17, 18], dense connections [40], channel attention [47, 5], and non-local module [23]. Recent methods explore orthogonal directions on improving the perceptual quality [20, 41], handling multiple degradation in a single model [46], and unsupervised learning [3, 45, 48].

With moving further from image to video data, video super-resolution aims to reconstruct a temporally consistent HR video from an LR input video. Huang *et al*. learn a bidirectional recurrent network [9] to directly predict the HR video. Several recent approaches [13, 4, 38, 32] rely on optical flow to compensate the motion in the input video. Another group of methods implicitly compensate motion with the dynamic filter network [11], deformable alignment [39], and 3D convolution [21].

2.2. Temporal Upsampling

Temporal upsampling, or video frame interpolation, aims to synthesize intermediate frames for increasing the temporal resolution of an input video while maintaining the temporal smoothness simultaneously. With the advancement of learning-based optical flow estimation methods [7, 10, 19, 30], recent approaches learn to estimate optical flow tailored for video frame interpolation [25, 12, 44, 43]. Niklaus *et al*. [27] adopt bi-directional flows to warp both images and contextual features for synthesizing the intermediate frame. While flow-based methods are able to handle large motion, the predicted frames often contain severe visual artifacts when the estimated flows are not accurate. On the other hand, the kernel-based method [28, 29] learns local adaptive kernels to blend the neighboring pixels for prediction. However, the memory footprint and computational load of the kernel-based approaches are too heavy for high-resolution input videos. Recently, Bao *et al*. [2] propose an adaptive warping layer to integrate the optical flow with local adaptive kernels. By using optical flow to warp input frames and then synthesizing pixels with local adaptive kernels, the model can handle large motion effectively and use smaller kernel sizes to reduce the memory usage. This approach is later extended to incorporate the depth prediction to explicitly detect occlusion [1] when synthesizing the intermediate frames.

2.3. Spatiotemporal Upsampling

Unlike spatial or temporal upsampling, spatiotemporal upsampling is a more challenging task but attracts less attention in the field. Early approaches [35, 26] use multiple low-resolution and low frame-rate videos of the same scene to reconstruct a high-resolution and high-frame-rate video. Shahar *et al*. [33] exploit the recurrences of space-time patches to propose an example-based method for spatiotemporal upsampling from a single input video. The CDCA method [34] learns convolutional auto-encoders to map the LR video to HR video. However, the mapping requires a pre-defined tricubic interpolation, which may not be able to reconstruct the missing high-frequency details in both spatial and temporal domains. Recently, Kim *et al*. [16] propose the FISR model, which uses multi-scale and temporal regularization to upscale the spatial and temporal resolutions of videos from 2K 30fps to 4K 60fps. Instead of introducing a brand-new architecture for realizing the spatiotemporal upsampling, we well utilize the power of existing spatial upsampling approaches and the temporal ones for building our spatiotemporal super-resolution framework, which shows favorable performance against FISR.

3. Proposed Method

Our goal here is to simultaneously upsample the spatial and temporal resolutions of a low-resolution low frame-rate video. To this end, we first analyze two baseline architectures by concatenating the spatial upsampling sub-network with the temporal upsampling sub-network, and vice versa (i.e., different orders of these two sub-networks for cascade). We discover the complementary property of the two baseline approaches, where one performs well on handling large motion and the other reconstructs finer details. Then, we propose a unified dual-stream fusion framework to adaptively merge their results for a better prediction. As shown in Fig. 1, the proposed framework consists of the following components: 1) a spatiotemporal upsampling module, 2) a fusion module, and 3) a refinement module. In the following, we introduce the function of each component as well as the loss functions for training our model.

3.1. Spatiotemporal Upsampling Module

Given two LR video frames $L^{(t-1)}$ and $L^{(t+1)}$ at timestamp $t-1$ and $t+1$, the spatiotemporal upsampling module generates three consecutive HR frames, $\hat{H}^{(t-1)}$, $\hat{H}^{(t)}$, and $\hat{H}^{(t+1)}$. We start with two basic upsampling components: a spatial upsampling sub-network $\mathbb{M}_S$, and a temporal upsampling sub-network $\mathbb{M}_T$. The $\mathbb{M}_S$ subnetwork takes a single LR frame L as input and generates a HR frame:

$$\hat{H} = \mathbb{M}_S(L). \tag{1}$$

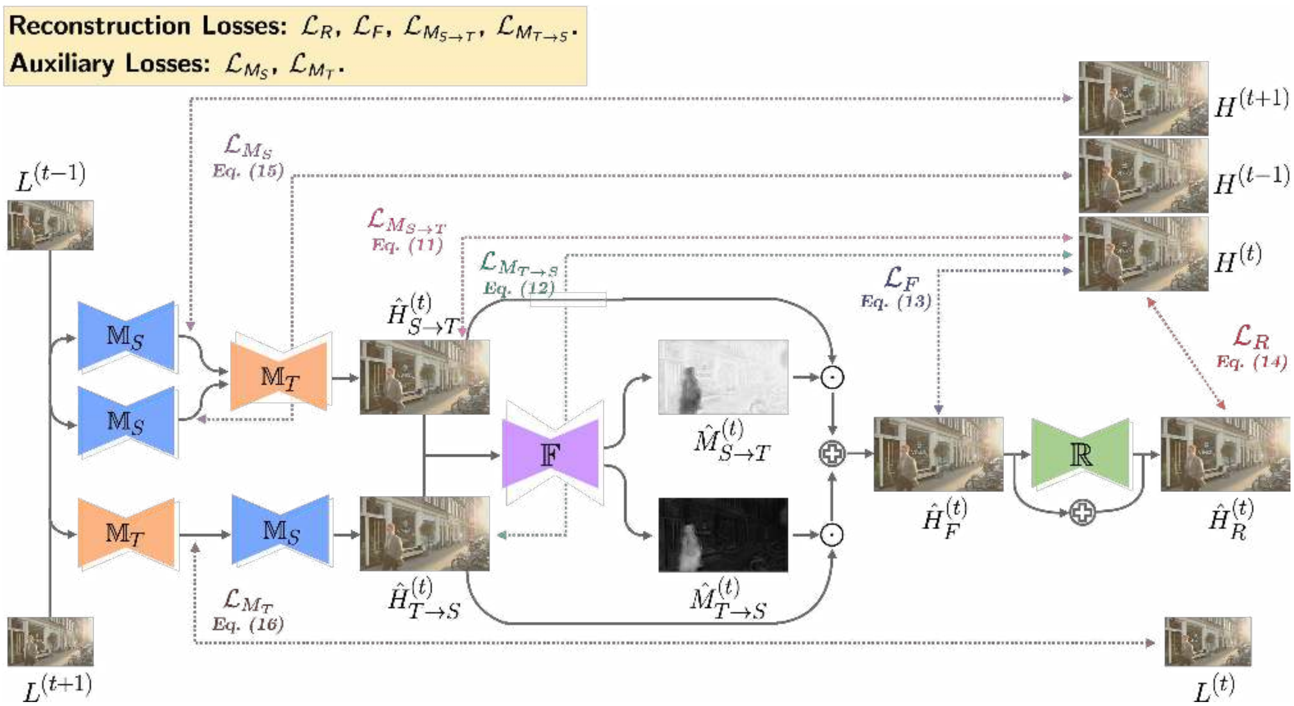

Figure 1: **Overview of the proposed dual-stream fusion framework.** Our spatiotemporal upsampling framework consists of three modules: (1) a spatiotemporal upsampling module, which generates two HR intermediate frames, $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$, from the LR input frames, $L^{(t-1)}$ and $L^{(t+1)}$, (2) a fusion module where the fusion network $\mathbb{F}$ predicts two blending masks to adaptively merge $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$ into $\hat{H}^{(t)}_F$, and (3) a refinement module $\mathbb{R}$ that refines $\hat{H}^{(t)}_F$ with a residual learning scheme and generates the final prediction $\hat{H}^{(t)}_R$. In particular, the sptaiotemporal upsampling module is composed of two basic upsampling streams, where the orders of cascading spatial upsampling sub-network $\mathbb{M}_S$ and temporal upsampling sub-network $\mathbb{M}_T$ are opposite across streams.

On the other hand, the $\mathbb{M}_T$ subnetwork generates an intermediate frame from the two input frames, $I^{(t-1)}$ and $I^{(t+1)}$:

$$\hat{I}^{(t)} = \mathbb{M}_T(I^{(t-1)}, I^{(t+1)}). \tag{2}$$

where $I^{(t-1)}$ and $I^{(t+1)}$ are the input temporal adjacent frames of arbitrary resolution, and $\hat{I}^{(t)}$ is the synthesized intermediate frame. The output HR frames $\hat{H}^{(t-1)}$ and $\hat{H}^{(t+1)}$ can be directly generated from the spatial upsampling sub-network, where $\hat{H}^{(t-1)} = \mathbb{M}_S(L^{(t-1)})$ and $\hat{H}^{(t+1)} = \mathbb{M}_S(L^{(t+1)})$. To generate the intermediate HR frame $\hat{H}^{(t)}$, we explore the following two strategies.

Spatial upsampling followed by temporal upsampling $\mathbb{M}_{\mathbf{S\to T}}$. We first generate the HR frames $\hat{H}^{(t-1)}$ and $\hat{H}^{(t+1)}$ with the spatial upsampling sub-network $\mathbb{M}_S$ and then synthesize the intermediate HR frame with the temporal upsampling sub-network $\mathbb{M}_T$:

$$\begin{aligned}\hat{H}^{(t)}_{S\to T} &= \mathbb{M}_{S\to T}(L^{(t-1)}, L^{(t+1)}), &(3)\\ &= \mathbb{M}_T(\mathbb{M}_S(L^{(t-1)}), \mathbb{M}_S(L^{(t+1)})), &(4)\\ &= \mathbb{M}_T(\hat{H}^{(t-1)}, \hat{H}^{(t+1)}). &(5)\end{aligned}$$

Temporal upsampling followed by spatial upsampling $\mathbb{M}_{\mathbf{T\to S}}$. We first synthesize the intermediate LR frame $\hat{L}^{(t)}$ with the $\mathbb{M}_T$ sub-network and then generate the intermediate HR frame with the $\mathbb{M}_S$ sub-network:

$$\begin{aligned}\hat{H}^{(t)}_{T\to S} &= \mathbb{M}_{T\to S}(L^{(t-1)}, L^{(t+1)}), &(6)\\ &= \mathbb{M}_S(\mathbb{M}_T(L^{(t-1)}, L^{(t+1)})), &(7)\\ &= \mathbb{M}_S(\hat{L}^{(t)}). &(8)\end{aligned}$$

The two spatiotemporal upsampling steams $\mathbb{M}_{S\to T}$ and $\mathbb{M}_{T\to S}$ use the same spatial and temporal upsampling sub-networks but apply them in a different order. In our experiments, we discover that the two streams show complementary results for spatiotemporal upsampling, where the stream $\mathbb{M}_{S\to T}$ generates finer details on areas with smaller motion, while the stream $\mathbb{M}_{T\to S}$ provides better reconstruction on areas with larger motion. More analyses and discussions are provided in Section 4.2.

3.2. Fusion Module

Due to the complementary property of the two spatiotemporal upsampling strategies, we propose a unified framework to take advantages from both streams. To this end, we train a fusion network $\mathbb{F}$ to blend the prediction results from $\mathbb{M}_{S\to T}$ and $\mathbb{M}_{T\to S}$. The fusion network learns to estimate two blending masks, $\hat{M}_{T\to S}$ and $\hat{M}_{S\to T}$, and fuse $\hat{H}^{(t)}_{T\to S}$ and $\hat{H}^{(t)}_{S\to T}$ by:

$$\begin{aligned}\hat{H}^{(t)}_F &= \mathbb{F}(\hat{H}^{(t)}_{S\to T}, \hat{H}^{(t)}_{T\to S}),\\ &= \hat{M}_{S\to T} \odot \hat{H}^{(t)}_{S\to T} + \hat{M}_{T\to S} \odot \hat{H}^{(t)}_{T\to S},\end{aligned} \quad (9)$$

where $\mathbb{M}_{S\to T} \in [0,1]$, $\mathbb{M}_{T\to S} \in [0,1]$, and $\odot$ denotes the element-wise multiplication. Note that here we can constrain the two masks to be complementary with each other, where $\hat{M}^{S\to T} = 1 - \hat{M}^{T\to S}$. In this way, the prediction $\hat{H}^{(t)}_F$ is a simple linear interpolation of $\hat{H}^{(t)}_{T\to S}$ and $\hat{H}^{(t)}_{S\to T}$. On the other hand, without such constraint (i.e., $\hat{M}^{S\to T}$ and $\hat{M}^{T\to S}$ are separate masks), each pixel is able to have one extra degree of freedom, and the prediction $\hat{H}^{(t)}_F$ becomes a linear combination of $\hat{H}^{(t)}_{T\to S}$ and $\hat{H}^{(t)}_{S\to T}$. We discuss the performance of these two design choices in our fusion network in Section 4.2.

3.3. Refinement Module

As the prediction $\hat{H}^{(t)}_F$ is a linear combination of two estimated frames (*i.e.*, $\hat{H}^{(t)}_{T\to S}$ and $\hat{H}^{(t)}_{S\to T}$), the output may inevitably look blurry and overly smoothed. In order to overcome this issue, we learn a small refinement network $\mathbb{R}$ to further enhance the details in the predicted frame. As shown in Fig. 1, the final output frame is generated via a residual learning scheme:

$$\hat{H}^{(t)}_R = \mathbb{R}(\hat{H}^{(t)}_F) + \hat{H}^{(t)}_F. \quad (10)$$

3.4. Objective Functions

We optimize the following losses to train the proposed model.

Reconstruction losses. We adopt the L_1 loss between the ground-truth frame $H^{(t)}$ and the intermediate predictions $\hat{H}^{(t)}_{S\to T}$, $\hat{H}^{(t)}_{T\to S}$, merged frame $\hat{H}^{(t)}_F$, and final prediction $\hat{H}^{(t)}_R$:

$$\mathcal{L}_{M_{S\to T}} = \left\| \hat{H}^{(t)}_{S\to T} - H^{(t)} \right\|_1, \quad (11)$$

$$\mathcal{L}_{M_{T\to S}} = \left\| \hat{H}^{(t)}_{T\to S} - H^{(t)} \right\|_1, \quad (12)$$

$$\mathcal{L}_F = \left\| \hat{H}^{(t)}_F - H^{(t)} \right\|_1, \quad (13)$$

$$\mathcal{L}_R = \left\| \hat{H}^{(t)}_R - H^{(t)} \right\|_1, \quad (14)$$

where $\mathcal{L}_{M_{S\to T}}$ is applied to the output of the stream $\mathbb{M}_{S\to T}$, $\mathcal{L}_{M_{T\to S}}$ is applied to the output of the stream $\mathbb{M}_{T\to S}$, $\mathcal{L}_F$ is applied to the output of the fusion module $\mathbb{F}$, and $\mathcal{L}_R$ is applied to the output of the refinement module $\mathbb{R}$.

Auxiliary losses. To stabilize the network training, we also enforce the following losses to the intermediate images that are generated during the two spatiotemporal upsampling streams:

$$\mathcal{L}_{M_S} = \left\| \hat{H}^{(t-1)} - H^{(t-1)} \right\|_1 + \left\| \hat{H}^{(t+1)} - H^{(t+1)} \right\|_1, \quad (15)$$

$$\mathcal{L}_{M_T} = \left\| \hat{L}^{(t)} - L^{(t)} \right\|_1, \quad (16)$$

where $\hat{H}^{(t-1)}$ and $\hat{H}^{(t+1)}$ are the upsampled frames from the spatial upsampling sub-network in the stream $\mathbb{M}_{S\to T}$, and $\hat{L}^{(t)}$ is the intermediate LR frame from the temporal upsampling sub-network in the stream $\mathbb{M}_{T\to S}$.

Overall loss. The overall objective to optimize our proposed spatiotemporal upsampling framework is a summation of the aforementioned losses:

$$\mathcal{L}_{total} = \mathcal{L}_{M_{S\to T}} + \mathcal{L}_{M_{T\to S}} + \mathcal{L}_F + \mathcal{L}_R + \mathcal{L}_{M_S} + \mathcal{L}_{M_T}. \quad (17)$$

We apply equal weights for all the loss functions to avoid any extra hyper-parameter tuning.

3.5. Implementation Details

Network architecture. We adopt state-of-the-art image super-resolution and video frame interpolation models as our basic spatial and temporal upsampling sub-networks, respectively (described in Section 4). Our fusion network $\mathbb{F}$ uses a U-Net architecture [31], which contains five symmetric downsampling and upsampling convolution layers with skip connections. The refinement network $\mathbb{R}$ has three residual blocks without any downsampling and upsampling layers. The details for all the network architecture are provided in the supplementary materials.

Training procedure. We adopt the following procedure for training:

1. Pre-train the basic upsampling sub-networks $\mathbb{M}_S$ and $\mathbb{M}_T$ independently.

2. Freeze the basic upsampling sub-network $\mathbb{M}_S$ and $\mathbb{M}_T$, and train the fusion network $\mathbb{F}$ and refinement network $\mathbb{R}$ by optimizing the reconstruction losses $\mathcal{L}_F$ and $\mathcal{L}_R$.

3. Jointly fine-tune all the (sub-)networks in an end-to-end manner by optimizing all the reconstruction losses and auxiliary losses.

Such a training procedure makes the entire model converge stably and achieve better results. The batch size is set to 24. We use the RAdam [24] optimizer with initial learning rate of $5e-5$ in all three training stages.

4. Experimental Results

We first introduce the datasets and evaluation metrics used in our experiments. We then provide quantitative and qualitative comparisons, as well as the ablation study between the proposed model and its variants.

4.1. Datasets and Evaluation Metrics

Datasets. Three commonly used video datasets are considered for both training and evaluation.

- **Vimeo-90K**: The Vimeo-90K [44] dataset contains 51312 triplets for training and 3782 triplets for evaluation, where each triplet contains three continuous frames of 448×256 pixels.
- **UCF101**: The UCF101 dataset [37] contains videos with a wide variety of human actions and camera motion. We randomly select 200 triplets from the full training set for our training, and use the 379-triplet test set, which is commonly adopted for evaluating the frame interpolation methods [25, 2, 1]. Each video frame is resized to 256×256 pixels.
- **FISR dataset**: The FISR dataset [16] contains 10 test videos with diverse objects and camera motions. Each video has five temporally-subsampled frames of 1920×1080 pixels as the input (*i.e.*, $\{L^{(t)}|t = 1, 3, 5, 7, 9\}$) and the corresponding seven consecutive frames of 3840×2160 pixels as the ground-truth for spatiotemporal upsampling (*i.e.*, $\{H^{(t)}|t = 2, 3, \cdots, 8\}$). This dataset is more challenging due to the high spatial resolution and large motion displacement.

In all our experiments, we consider upsampling the spatial resolution for $2\times$ and increasing the temporal frame rate for $2\times$. For each triplet in Vimeo-90K and UCF-101, we downsample the spatial resolution of the first and the third frames by $2\times$ as the input. Then the the model performance is evaluated on the second frame of each triplet.

Metrics. PSNR and SSIM [42] are adopted for quantitative evaluation, which are widely used in low-level vision tasks.

4.2. Quantitative and Qualitative Evaluations

Complementary property of two spatiotemporal upsampling streams. We first demonstrate the complementary property of the two baseline spatiotemporal upsampling streams: $\mathbb{M}_{S\to T}$ and $\mathbb{M}_{T\to S}$. Here we use the ESPCN [36] and SuperSloMo [12] as the spatial and temporal upsampling sub-networks, respectively. In Fig. 2, we compare the predicted intermediate frames, $\hat{H}^{(t)}_{S\to T}$ (generated from $\mathbb{M}_{S\to T}$) and $\hat{H}^{(t)}_{T\to S}$ (generated from $\mathbb{M}_{T\to S}$), and show their error maps with respect to the ground-truth frame. We observe that $\hat{H}^{(t)}_{S\to T}$ has finer details in areas with smaller motion, while $\hat{H}^{(t)}_{T\to S}$ provides better reconstruction in areas with larger motion. Such an observation guides us to develop the proposed framework for utilizing the benefits from both of the streams.

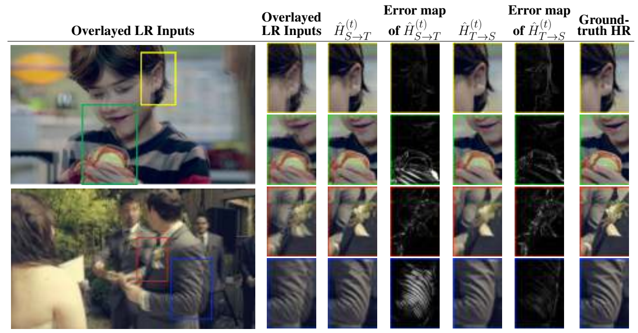

Figure 2: **Complementary property of the spatiotemporal upsampling methods.** We visualize the predicted intermediate frames $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$, the ground-truth frame $H^{(t)}$, and the error maps of the two predictions w.r.t the ground-truth. The stream $\mathbb{M}_{S\to T}$ (i.e., spatial upsampling followed by temporal upsampling) generates finer details but shows larger errors when input frames have complex motion. On the other hand, the stream $\mathbb{M}_{T\to S}$ (i.e., temporal upsampling followed by spatial upsampling) performs well with large motion but cannot reconstruct fine details.

Analysis on the fusion module. As mentioned in Section 3.2, our fusion module learns to predict two separate masks or a single mask (*i.e.*, having the complementary constraint between two masks) for blending. Table 1 compares the performance of these two design choices. First, we observe that the predictions $\hat{H}^{(t)}_{F}$ from the fusion module are more accurate than both $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$ as the fusion module adaptively blends the pixels from which they reconstruct well. Second, the two-mask fusion performs much better than the one-mask fusion with only introducing 0.001% more parameters in the fusion network (*i.e.*, the only modification is the number of channels in the last layer of the fusion network $\mathbb{F}$). The two-mask fusion network allows each pixel to have one extra degree-of-freedom for blending, effectively increasing the solution space to find a better reconstruction. Therefore, we choose the two-mask fusion network in our framework. Fig. 3 shows the visual comparisons between the one-mask and two-mask designs, while Fig. 4 visualizes the blending masks $\hat{M}_{S\to T}$ and $\hat{M}_{T\to S}$.

Analysis on the training procedure. In Table 2, we compare the model performance in each stage of our training procedure. While the stream $\mathbb{M}_{S\to T}$ (2nd row) typically

Table 1: **Quantitative comparisons on design choices for fusion network.** Our fusion network learns to blend the intermediate predictions, $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$, leading to better reconstruction with respect to $H^{(t)}$ than both of the streams. The two-mask fusion module further improves the accuracy by predicting the independent masks for the outputs of both streams.

Dataset	$\hat{H}^{(t)}_{S\to T}$		$\hat{H}^{(t)}_{T\to S}$		One-mask $\hat{H}^{(t)}_{F}$		Two-mask $\hat{H}^{(t)}_{F}$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Vimeo-90K	31.17	0.9187	31.41	0.9179	32.03	0.9288	32.23	0.9313
UCF-101	30.87	0.9247	30.71	0.9251	31.23	0.9290	31.38	0.9308

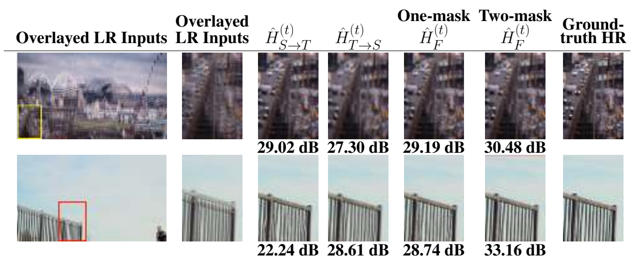

Figure 3: **Visual comparisons of the fusion network.** The fusion network $\mathbb{F}$ estimates blending masks to adaptively fuse the predictions $\hat{H}^{(t)}_{S\to T}$ and $\hat{H}^{(t)}_{T\to S}$. We show that the two-mask design reconstructs more accurate details than the one-mask variation.

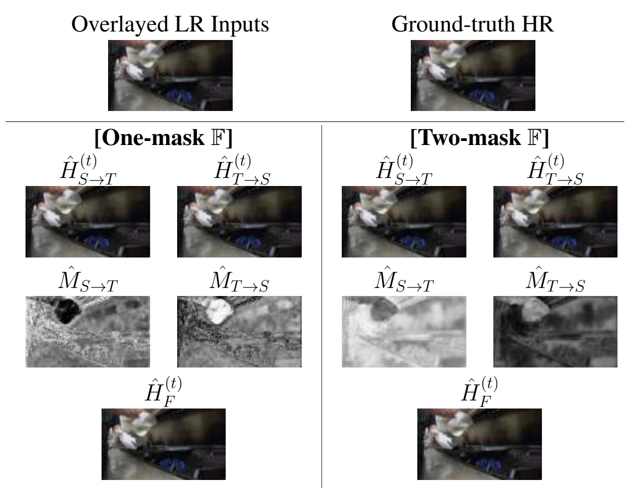

Figure 4: **Visualization of the blending masks.** The one-mask fusion module predicts a single mask (*i.e.*, under the constraint $\hat{M}_{T\to S} + \hat{M}_{S\to T} = 1$), while the two-mask fusion module only requires $\hat{M}_{T\to S} \in [0,1]$ and $\hat{M}_{S\to T} \in [0,1]$, allowing each pixel to have one extra degree-of-freedom for blending.

performs better than the stream $\mathbb{M}_{T\to S}$ (1st row), the fusion module utilizes the prediction from both streams and leads to better reconstruction with respect to the ground-truth (3rd row). The refinement module further improves the accuracy (4th row). Note that in the 3rd and 4th rows, the two streams $\mathbb{M}_{S\to T}$ and $\mathbb{M}_{T\to S}$ are frozen with both fusion network $\mathbb{F}$

Figure 5: **Visual comparisons of results from different training stages of the proposed framework.** Please see Table 2 for the specific setting of each model variant. PSNR values are provided below the corresponding output frames.

and refinement network $\mathbb{R}$ being trained. The 3rd row is actually the *intermediate result* obtained from the output of fusion network $\mathbb{F}$. Finally, we jointly fine-tune the whole pipeline to significantly boost the reconstruction accuracy (5th rows). Fig. 5 compares the reconstructed frames at each of the training stage. Our full pipeline with joint fine-tuning obtains the sharper results with finer details.

Comparisons of different upsampling sub-networks. We analyze the performance of the proposed framework by replacing the fundamental spatial and temporal upsampling sub-networks with different backbones. For the spatial upsampling sub-network $\mathbb{M}_S$, we use the SAN [5], which is a state-of-the-art single image super-resolution method, and ESPCN [36], which is an efficient image super-resolution model using the pixel shuffling. For the temporal upsampling sub-network $\mathbb{M}_T$, we compare the state-of-the-art video frame interpolation methods, DAIN [1] and Super-SloMo [12]. For fair comparisons, we fix the spatial and temporal upsampling sub-networks (i.e., use their off-the-shelf pre-trained weights) and only update our fusion and refinement networks. Table 4 shows the quantitative comparisons of different combinations on the Vimeo-90K and UCF-101 test sets. In each row, we observe that our fusion

Table 2: **Quantitative comparisons on each training stage.** Our framework starts with pre-training the baseline streams $\mathbb{M}_{T\rightarrow S}$ and $\mathbb{M}_{S\rightarrow T}$. Then, we freeze $\mathbb{M}_{T\rightarrow S}$ and $\mathbb{M}_{S\rightarrow T}$ to train the fusion network $\mathbb{F}$ and refinement network $\mathbb{R}$. Finally, we jointly fine-tune all the sub-networks end-to-end.

Setting		Vimeo-90K		UCF101	
		PSNR	SSIM	PSNR	SSIM
I.	$\mathbb{M}_{T\rightarrow S}$ (pre-trained)	31.41	0.9179	30.71	0.9251
II.	$\mathbb{M}_{S\rightarrow T}$ (pre-trained)	31.17	0.9187	30.87	0.9247
III.	$\mathbb{M}_{T\rightarrow S}$ (fixed) + $\mathbb{M}_{S\rightarrow T}$ (fixed) + $\mathbb{F}$	32.23	0.9313	31.38	0.9308
IV.	$\mathbb{M}_{T\rightarrow S}$ (fixed) + $\mathbb{M}_{S\rightarrow T}$ (fixed) + $\mathbb{F}$ + $\mathbb{R}$	32.35	0.9326	31.45	0.9313
V.	$\mathbb{M}_{T\rightarrow S} + \mathbb{M}_{S\rightarrow T} + \mathbb{F} + \mathbb{R}$ (jointly fine-tuned)	**32.85**	**0.9401**	**31.54**	**0.9317**

and refinement networks consistently improve the performance, demonstrating the capability of our framework for being integrated with existing spatial and temporal upsampling methods. Several examples for the qualitative comparisons among different combinations of upsampling sub-networks are provided in Fig. 6.

Table 3: **Quantitative comparisons with the state-of-the-art spatiotemporal upsampling method, FISR [16] and STARnet [8].** The experiments are conducted on the test sets of the FISR and Vimeo-90K datasets, and the performance is evaluated on the spatiotemporal upsampling output frames in RGB color space.

Spatiotemporal Upsampling	FISR dataset		Vimeo -90K	
	PSNR	SSIM	PSNR	SSIM
FISR	32.04	0.9241	25.09	0.7612
STARNet	31.84	0.9273	**33.07**	0.9418
Ours	**33.27**	**0.9360**	32.97	**0.9423**

Comparison to FISR and STARnet. We compare the proposed method with the recently proposed state-of-the-art spatiotemporal upsampling methods, FISR [16] and STARnet [8]. The pre-trained model provided by [16] is used in our experiments. For our method, we utilize SAN [5] and DAIN [1] as upsampling sub-networks $\mathbb{M}_S$ and $\mathbb{M}_T$ respectively. And for STARnet, we retrain it on Vimeo-90K following our settings. We use every two consecutive LR frames (*i.e.*, $L^{(t-1)}$ and $L^{(t+1)}$) as the input to reconstruct the spatiotemporal upsampling frame (*i.e.*, $\hat{H}^{(t)}$). Table 3 shows the quantitative evaluation for spatiotemporal upsampling on the FISR and Vimeo-90K test sets. On the FISR test set, our method performs favorably against FISR and STARnet. On the Vimeo-90K test set, the proposed method has superior performance with respect to FISR and comparable performance with respect to STARnet. It is worth noting that our method is able to generalize well to the FISR test set even the FISR training set is not included in our training, while the FISR model does not produce satisfactory results on Vimeo-90K. In the supplementary materials, we provide more qualitative comparisons between our method and FISR, in which our method can better handle the large motion displacement and generate fewer artifacts on challenging examples.

4.3. Limitations and Discussions

The proposed method leverages the complementary property of two spatiotemporal upsampling streams. There are two main limitations in our current framework. First, the reconstruction is solely guided by the pixel-wise reconstruction losses of the intermediate and final output frames. The temporal consistency between output frames $\hat{H}^{(t-1)}$, $\hat{H}^{(t)}$, and $\hat{H}^{(t+1)}$ is not explicitly enforced. A future direction may consider enforcing the temporal warping losses in the image and feature spaces, or exploring temporal recurrent components such as long short-term memory (LSTM). Second, although our framework utilizes existing image super-resolution and video frame interpolation models, the model size and computational load could inevitably increase with the size of the backbone upsampling modules. One future work is to develop an one-stage pipeline to directly perform spatiotemporal upsampling to reduce the model complexity.

5. Conclusions

We propose a novel end-to-end spatiotemporal upsampling framework which increases both the video frame-rate and the spatial resolution of video frames simultaneously for better visual experience. Based on two cascades of spatial and temporal upsampling sub-networks with different execution orders, we take advantage of the complementary property between the cascades by proposing a fusion module to effectively combine their outputs, and further utilize a refinement module to enhance the fine details. We conduct extensive experiments to demonstrate the efficacy of our proposed framework against several baselines, in terms of both visual quality and quantitative results. In addition, the thorough ablation studies are performed to verify our design choices. Moreover, in comparison to the other methods that build the spatiotemporal upsampling model from scratch, our framework is beneficial as it can be easily boosted, once

Table 4: **Quantitative comparisons among different combinations of upsampling sub-networks.** Our proposed framework can be integrated with any off-the-shelf CNN-based spatial/temporal upsampling sub-networks. We evaluate the combinations in basis of two single-image super-resolution methods, ESPCN [36] and SAN [5], and two video-frame interpolation approaches, SuperSloMo [12] and DAIN [1]. Our fusion and refinement modules consistently improve the performance on both Vimeo-90K and UCF-101 datasets.

Vimeo-90K	$\hat{H}^{(t)}_{S\to T}$		$\hat{H}^{(t)}_{T\to S}$		$\hat{H}^{(t)}_{F}$		$\hat{H}^{(t)}_{R}$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ESPCN + SuperSloMo	31.17	0.9187	31.41	0.9179	32.23	0.9313	32.35	0.9326
ESPCN + DAIN	32.32	0.9347	31.67	0.9248	32.72	0.9396	32.83	0.9407
SAN + SuperSloMo	31.35	0.9215	31.73	0.9225	32.41	0.9339	32.51	0.9350
SAN + DAIN	32.70	0.9394	31.93	0.9279	32.89	0.9414	32.97	0.9423
UCF-101	$\hat{H}^{(t)}_{S\to T}$		$\hat{H}^{(t)}_{T\to S}$		$\hat{H}^{(t)}_{F}$		$\hat{H}^{(t)}_{R}$	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ESPCN + SuperSloMo	30.87	0.9247	30.71	0.9251	31.38	0.9308	31.45	0.9313
ESPCN + DAIN	31.12	0.9303	31.27	0.9284	31.54	0.9328	31.60	0.9331
SAN + SuperSloMo	31.12	0.9253	30.90	0.9261	31.37	0.9298	31.43	0.9306
SAN + DAIN	31.33	0.9310	31.48	0.9286	31.59	0.9317	31.64	0.9323

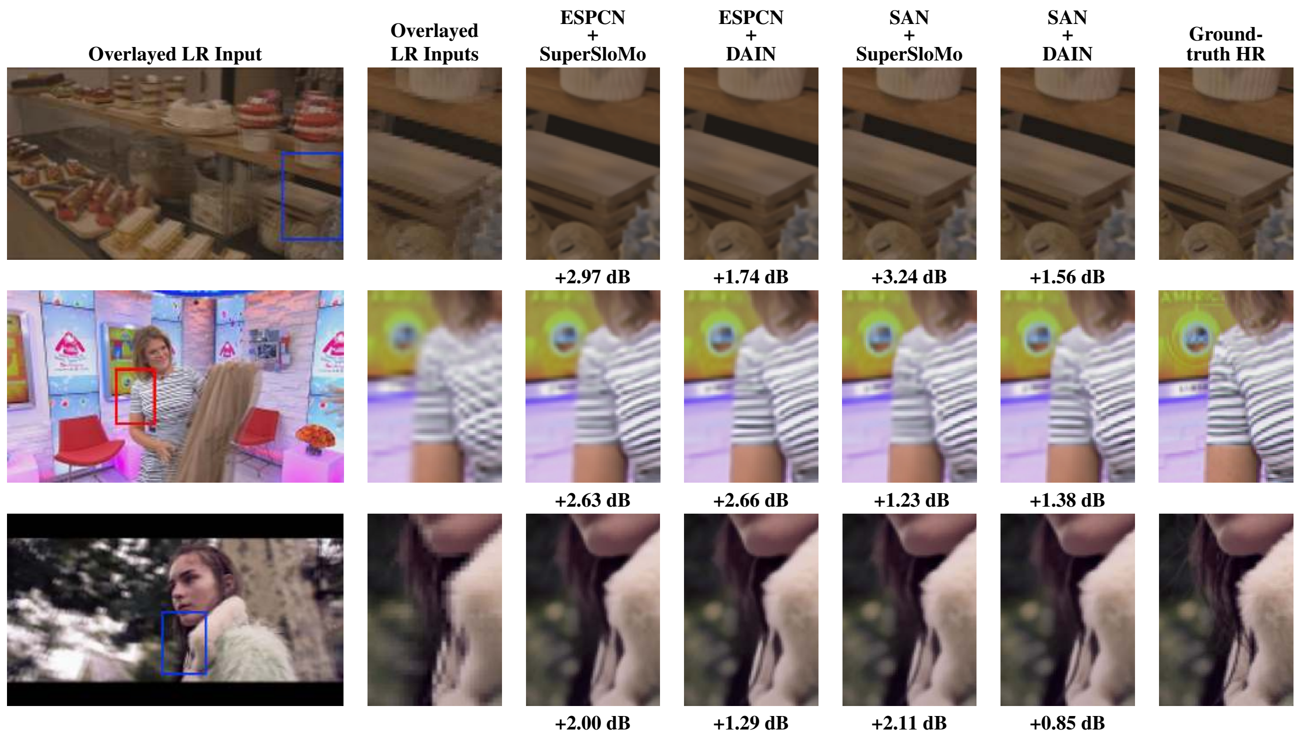

Figure 6: **Visual comparison among different combinations of upsampling sub-networks.** We show that the proposed framework can be integrated with state-of-the-art image super-resolution and video frame interpolation methods to achieve high-quality spatiotemporal upsampling results. Numbers below the reconstructed frames indicate the PSNR gains between $\hat{H}^{(t)}_{R}$ and the maximum one among $\{\hat{H}^{(t)}_{S\to T}, \hat{H}^{(t)}_{T\to S}\}$ (*i.e.*, the improvement made by our fusion and refinement modules).

either the spatial upsampling or temporal upsampling module is improved, without requiring additional efforts in designing new operations or network architectures.

Acknowledgement This project is supported by MediaTek (under MediaTek-NCTU Research Center) and Ministry of Science and Technology of Taiwan (under grants MOST-109-2634-F-009-015, MOST-109-2634-F-009-020, and MOST-109-2636-E-009-018). We are also grateful to the National Center for High-performance Computing of Taiwan for computer time and facilities.

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

[3] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a GAN to learn how to do image degradation first. In *European Conference on Computer Vision (ECCV)*, 2018.

[4] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[9] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[10] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[12] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[13] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging (TCI)*, 2016.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Fisr: Deep joint frame interpolation and super-resolution with a multi-scale temporal loss. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[17] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[19] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[20] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Sheng Li, Fengxiang He, Bo Du, Lefei Zhang, Yonghao Xu, and Dacheng Tao. Fast spatio-temporal residual network for video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[22] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[23] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[24] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *ArXiv:1908.03265*, 2019.

[25] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[26] Uma Mudenagudi, Subhashis Banerjee, and Prem Kumar Kalra. Space-time super-resolution using graph-cut optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2010.

[27] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[29] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[30] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

[32] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] Oded Shahar, Alon Faktor, and Michal Irani. Space-time super-resolution from a single video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[34] Manoj Sharma, Santanu Chaudhury, and Brejesh Lall. Space-time super-resolution using deep learning based framework. In *International Conference on Pattern Recognition and Machine Intelligence (PReMI)*, 2017.

[35] Eli Shechtman, Yaron Caspi, and Michal Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2005.

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv:1212.0402*, 2012.

[38] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[39] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally deformable alignment network for video super-resolution. *ArXiv:1812.02898*, 2018.

[40] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[41] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 2004.

[43] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[44] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 2019.

[45] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[46] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[47] Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[48] Tianyu Zhao, Wenqi Ren, Changqing Zhang, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. *ArXiv:1812.04240*, 2018.