

# Blurry Video Frame Interpolation

Wang Shen Wenbo Bao Guangtao Zhai\* Li Chen Xiongkuo Min Zhiyong Gao  
 Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

## Abstract

Existing works reduce motion blur and up-convert frame rate through two separate ways, including frame deblurring and frame interpolation. However, few studies have approached the joint video enhancement problem, namely synthesizing high-frame-rate clear results from low-frame-rate blurry inputs. In this paper, we propose a blurry video frame interpolation method to reduce motion blur and up-convert frame rate simultaneously. Specifically, we develop a pyramid module to cyclically synthesize clear intermediate frames. The pyramid module features adjustable spatial receptive field and temporal scope, thus contributing to controllable computational complexity and restoration ability. Besides, we propose an inter-pyramid recurrent module to connect sequential models to exploit the temporal relationship. The pyramid module integrates a recurrent module, thus can iteratively synthesize temporally smooth results without significantly increasing the model size. Extensive experimental results demonstrate that our method performs favorably against state-of-the-art methods. The source code and pre-trained model are available at <https://github.com/laomao0/BIN>.

## 1. Introduction

Shutter speed and exposure time of camera sensors are two fundamental factors that affect the quality of captured videos [33]. Slow shutter speed and long exposure time may lead to two kinds of degradations: motion blur and low frame rate. Eliminating these degradations is critical for enhancing the quality of captured videos. However, few studies have approached the joint problem, namely synthesizing high-frame-rate clear results from low-frame-rate blurry inputs. Existing methods may help address this problem by image deblurring and frame interpolation, but are often sub-optimal due to the lack of a joint formulation.

Frame interpolation aims to recover unseen intermediate frames from the captured ones [1, 9, 2, 3]. It can up-convert frame rate and improve visual smoothness. Most state-of-the-art frame interpolation methods [1, 9, 2] first estimate objects' motion, and then perform frame warping



Figure 1. Examples of synthesizing an intermediate frame from blurry inputs. We show the results of (a) overlapped blurry inputs, (b) cascaded interpolation and deblurring model, (c) cascaded deblurring and interpolation model, (d) our model.

to synthesize pixels using reference frames. However, if the original reference frames are degraded by motion blur, the motion estimation may not be accurate. Consequently, it is challenging to restore clear intermediate frames via existing frame interpolation approaches.

Considering the above problems introduced by motion blur, some existing methods generally employ a pre-deblurring procedure [32, 35, 30]. A straightforward approach is to perform frame deblurring, followed by the frame interpolation, which we refer to as the *cascade* scheme. However, this approach is sub-optimal in terms of interpolation quality. First, the interpolation performance is highly dependent on the quality of the deblurred images. The pixel errors introduced in the deblurring stage will be propagated to the interpolation stage, thus degrading the overall performance. Second, most of the frame interpolation methods use two consecutive frames as a reference, namely those methods have a *temporal scope* of two. However, given imperfect deblurred frames in the cascade scheme, the interpolation model with a short temporal scope can hardly maintain the long-term motion consistency among adjacent frames. An alternative strategy is to perform frame interpolation and then frame deblurring. However, the overall quality deteriorates because the interpolated frames suffer from blurry textures of the inputs, as shown in Figure 1.

In this paper, we formulate the joint video enhancement problem with a unified degradation model. Then we pro-

\*Corresponding author

pose a Blurry video frame INterpolation (BIN) method, including a pyramid module and an inter-pyramid recurrent module. The structure of our pyramid module resembles a pyramid that consists of multiple backbone networks. The pyramid module is flexible. As the scale increases, the model creates a larger spatial receptive field and a broader temporal scope. The flexible structure can also make a trade-off between computational complexity and restoration quality. Besides, we adopt cycle loss [17, 27, 38, 6, 34, 26] to enforce the spatial consistency between the input frames and the re-generated frames of the pyramid module.

Based on the pyramid structure, we propose an inter-pyramid recurrent module which effectively exploits the time information. Specifically, the recurrent module adopts ConvLSTM units to propagate the frame information across time. The propagated frame information helps the model restore fine details and synthesize temporally consistent images. Besides conventional restoration evaluation criteria, we also propose an optical-flow based metric to evaluate the motion smoothness of synthesized video sequences. We use both existing databases as well as a new composed dataset crawled from YouTube for performance evaluation. Extensive experiments on the Adobe240 dataset [30] and our YouTube240 dataset demonstrate that the proposed BIN performs favorably against state-of-the-art methods.

Our main contributions are summarized as follows:

- We formulate the joint frame deblurring and interpolation problem by exploring the camera’s intrinsic properties related to motion blur and frame rate.
- We propose a blurry video frame interpolation method to jointly reduce blur and up-convert frame rate, and we propose an inter-pyramid recurrent module to enforce temporal consistency across generated frames.
- We demonstrate that the proposed method can fully exploit space-time information and performs favorably against state-of-the-art methods.

## 2. Related Work

In this section, we introduce the related literature for frame interpolation, video deblurring, and the joint restoration problem.

**Video Frame Interpolation.** Existing methods for frame interpolation generally utilize optical flow to process motion information [18, 1, 2, 9, 15, 37, 23] or use kernel-based models [19, 24, 25]. As a pioneer of learning-based methods, Long *et al.* [19] train a generic convolutional neural network to synthesize the intermediate frame directly. The AdaConv [24] and SepConv [25] estimate spatially-adaptive interpolation kernels to synthesize pixels from a large neighborhood. Meyer *et al.* [20] use the phase shift of single-pixel to represent motion and construct intermediate frames using a modified per-pixel phase without using optical flow. Bao *et al.* [2] integrate the flow-based and kernel-

based approaches. Their adaptive warping layer synthesizes a new pixel using a local convolutional kernel where the position of the kernel window is determined by optical flow.

Estimating accurate optical flow is very difficult when the interpolation model encounters blurry inputs. We use a variation of the residual dense network [41] as the *backbone network*. It can generate the intermediate frame without using optical flow. Moreover, we use multiple backbone networks to construct a pyramid module, which can simultaneously reduce blur and up-convert frame rate.

**Video Deblurring.** Existing learning-based deblurring methods reduce motion blur using multiple frames [30, 11, 12, 11, 21, 7] or single image [32, 30, 14]. Wang *et al.* [35] first extract feature information from multiple inputs, then use feature alignment and fusion module to restore high-quality deblurred frames. To further exploit the temporal information, existing algorithms use the recurrent mechanism [8, 40, 28, 16, 39]. Kim *et al.* [8] introduce a spatio-temporal recurrent architecture with a dynamic temporal blending mechanism that enables adaptive information propagation. Zhou *et al.* [42] use a spatio-temporal filter adaptive network to integrate feature alignment and deblurring. Their model recurrently uses information of the previous frame and current inputs. Nah *et al.* [22] adapt the hidden states transferred from past frames to the current frame to exploit information between video frames.

We integrate the backbone network with the proposed inter-pyramid recurrent module to operate iteratively. The proposed recurrent module adopts ConvLSTM units [36] to propagate the frame information between adjacent backbone networks. Due to the recurrence property, the proposed module can iteratively synthesize temporally smooth results without significantly increasing model size.

**Joint Video Deblurring and Interpolation.** Few studies have approached the joint video enhancement problem. Jin *et al.* [10] introduce the closest related work. Their model can be categorized into the jointly optimized cascade scheme. It first extracts several clear keyframes, and then synthesizes intermediate frames using those keyframes. Their model adopts an approximate recurrent approach by unfolding and distributing the extraction of the frames over multiple processing stages.

Our method differs from Jin *et al.* [10]’s algorithm in two aspects. First, our model is jointly optimized, and we do not explicitly distinguish the frame deblurring stage or the frame interpolation stage. We use the proposed backbone network to associate frame deblurring and interpolation uniformly. Second, instead of constructing an approximate recurrent mechanism, we explicitly use the proposed inter-pyramid recurrent module that adopts ConvLSTM units to propagate the frame information across time.

### 3. Joint Frame Deblurring and Interpolation

In this section, we introduce the degradation model for motion blur and low frame rate, and we formulate the joint frame deblurring and interpolation problem.

#### 3.1. Degradation Model

Generally, a camera captures videos by periodically turning on and off its shutter [33]. While the shutter is on, also known as exposure, the sensors integrate the luminous intensity reflected by objects to acquire the brightness of objects' pixels. Therefore, the exposure time accounts for the pixel brightness, and the shutter on-off frequency determines the video frame rate. Formally, we assume that there exists a latent image  $\mathbf{L}(\tau)$  at each instant time  $\tau$ , as shown in Figure 2. We integrate the latent images from time  $t_1$  over an interval of time (the exposure interval  $e$ ) to obtain one captured frame. We formulate the acquisition of a single frame as:

$$\mathbf{B}_{t_1} = \frac{1}{e} \int_{t_1}^{t_1+e} \mathbf{L}(\tau) d\tau. \quad (1)$$

Then at the next shutter time  $t_2$ , the camera generates another frame denoted by  $\mathbf{B}_{t_2}$ . The frame rate of the captured video is defined by:

$$f = \frac{1}{t_2 - t_1}. \quad (2)$$

Particularly, fast objects movement or camera shake during the exposure time would deteriorate the pixel brightness. This deterioration is often in the form of visual blur.

#### 3.2. Problem Formulation

Given low-frame-rate blurred inputs, we aim to generate high-frame-rate clear outputs. Our goal is to enhance the input video to provide a clear and smooth visual experience. We formulate the joint blur reduction and frame rate up-conversion problem as maximizing a *posteriori* of the output frames conditioned on the blurred inputs:

$$\mathcal{F}^* \doteq \max_{\mathcal{F}} p(\hat{\mathbf{I}}_{1:1:2N-1} | \mathbf{B}_{0:2:2N}), \quad (3)$$

where  $\mathbf{B}_{0:2:2N}$  denotes the low-frame-rate blurry inputs starting from index 0 to  $2N$  with a time step of 2,  $\hat{\mathbf{I}}_{1:1:2N-1}$  represents the restored and frame rate up-converted results, and  $\mathcal{F}^*$  refers to the optimal joint space-time enhancement model. We propose to use trainable neural networks to approximate the optimal model  $\mathcal{F}^*$ . We reformulate the problem in Equation (3) as a minimization of the loss function  $\mathcal{L}$  over dataset  $\mathcal{S}$ :

$$\begin{aligned} & \underset{\mathcal{F}(\cdot; \Theta)}{\text{minimize}} \quad \sum_{s \in \mathcal{S}} \mathcal{L}(\hat{\mathbf{I}}_{1:1:2N-1} | \mathbf{I}_{1:1:2N-1}) \\ & \text{subject to} \quad \hat{\mathbf{I}}_{1:1:2N-1} = \mathcal{F}(\mathbf{B}_{0:2:2N}), \end{aligned} \quad (4)$$

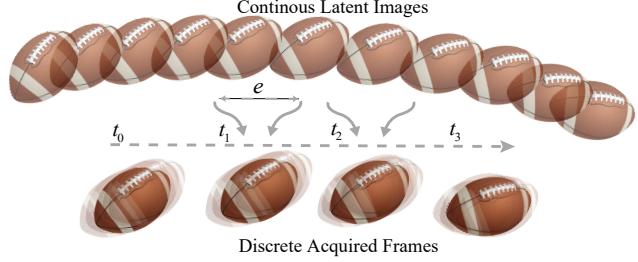


Figure 2. **Example of frame capturing.** Camera sensors capture the discrete frames at time step  $t_0, t_1, t_2, t_3$ , each of which requires continuous latent images within an exposure time interval of  $e$ .

where  $\mathbf{I}_{1:1:2N-1}$  denotes the ground-truth frames in the video sample  $s \in \mathcal{S}$ , and  $\mathcal{F}(\cdot; \Theta)$  refers to the proposed BIN with network parameters  $\Theta$ .

### 4. Blurry Video Frame Interpolation

The proposed model consists of two key components: the pyramid module and the inter-pyramid recurrent module. We use the pyramid module to reduce blur and up-convert frame rate simultaneously. The inter-pyramid recurrent module can further enforce temporal consistency between neighboring frames. We show the overall network architecture in Figure 3. Below we describe the design of each sub-network and the implementation details.

#### 4.1. Pyramid Module

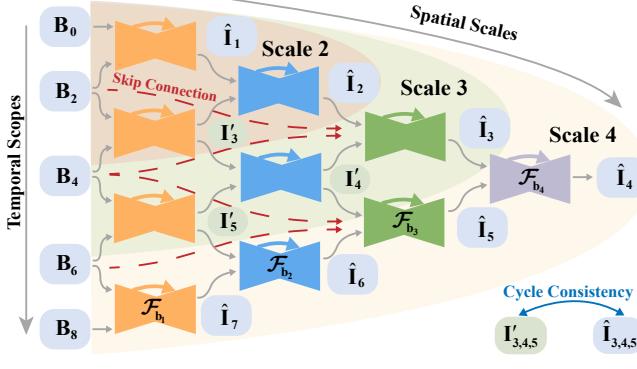
The proposed pyramid module integrates frame deblurring and frame interpolation by the following operation:

$$\hat{\mathbf{I}}_{1:1:2N-1} = \mathcal{F}(\mathbf{B}_{0:2:2N}), \quad (5)$$

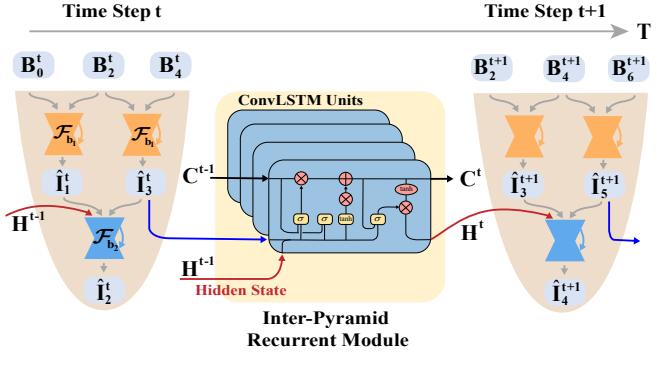
where  $\mathcal{F}$  refers to the pyramid module. It takes  $N+1$  frames  $\mathbf{B}_{0:2:2N}$  as input, and outputs the deblurred and the interpolated frames  $\hat{\mathbf{I}}_{1:1:2N-1}$ . We construct multiple *backbone networks* to build the pyramid module, as shown in Figure 3(a). The backbone network  $\mathcal{F}_b$  interpolates an intermediate frame using two consecutive inputs:

$$\hat{\mathbf{I}}_1 = \mathcal{F}_b(\mathbf{B}_0, \mathbf{B}_2). \quad (6)$$

The pyramid module has an adjustable spatial receptive field and temporal scope by alternating the scales of the model architecture. We show networks with three different scales in Figure 3(a), denoted by Scale 2, Scale 3 and Scale 4. The increase of scales makes the entire network deeper, thus creating a larger spatial receptive field. At the same time, the increase of scales also extends the number of inputs, namely the temporal scope, which facilitates the utilization of contextual temporal information. For example, the module of scale 2 has a temporal scope of three, while the module of scale 4 can exploit information from five frames, and it has a deeper receptive field compared to the module of scale 2.



(a) Pyramid Module



(b) Computation Flow of BIN<sub>2</sub>

Figure 3. **Architectures of the proposed blurry video frame interpolation model.** The proposed BIN consists of two key components, the pyramid module and the inter-pyramid recurrent module. The pyramid module in (a) consists of multiple backbone networks. The backbone networks in the same color have shared weights. It takes two consecutive frames as input, then synthesizes one intermediate frame. The pyramid module can reduce motion blur and interpolate intermediate frames simultaneously. Based on the pyramid module in (a), we integrate it with inter-pyramid recurrent module to realize the recurrent mechanism in (b). The proposed inter-pyramid recurrent module uses ConvLSTM units to propagate the frame information across different pyramid modules.

Besides the output frames  $\hat{\mathbf{I}}_{1:1:2N-1}$ , the pyramid module also generates multiple temporary frames. As shown in Figure 3(a), the pyramid module with scale 4 has three temporary frames  $\{\mathbf{I}'_3, \mathbf{I}'_4, \mathbf{I}'_5\}$ . We use a cycle consistency loss to ensure the spatial consistency between temporary frames with the cycle-paired frames (e.g.,  $\{\mathbf{I}'_3, \hat{\mathbf{I}}_3\}$ ).

#### 4.2. Inter-Pyramid Recurrent Module

Temporal motion smoothness is a critical factor in affecting human visual experiences. Based on the pyramid structure, we propose an inter-pyramid recurrent module to construct multi-scale Blurry frame INterpolation models, denoted by BIN<sub>l</sub>, where  $l$  is the scale of the pyramid structure. The recurrent module can further enforce temporal motion consistency between neighboring frames. The inter-pyramid recurrent module consists of multiple ConvLSTM units. Each ConvLSTM unit uses hidden states to propagate previous frame information to the current pyramid module.

For brevity, we illustrate the computation flow of BIN<sub>2</sub>, which takes one ConvLSTM unit and one pyramid module with scale 2. As shown in Figure 3(b), at time  $t \in [1, T]$ , given three inputs  $\mathbf{B}_{0:2:4}^t$ , we first generate two intermediate frames  $\hat{\mathbf{I}}_1^t$  and  $\hat{\mathbf{I}}_3^t$  by feed-forwarding network  $\mathcal{F}_{b_1}$  twice:

$$\hat{\mathbf{I}}_1^t = \mathcal{F}_{b_1}(\mathbf{B}_0^t, \mathbf{B}_2^t), \quad (7)$$

$$\hat{\mathbf{I}}_3^t = \mathcal{F}_{b_1}(\mathbf{B}_2^t, \mathbf{B}_4^t). \quad (8)$$

Then, we use the synthesized intermediate frames  $\hat{\mathbf{I}}_1^t, \hat{\mathbf{I}}_3^t$  as well as the hidden state  $\mathbf{H}^{t-1}$  to synthesize the deblurred frame  $\hat{\mathbf{I}}_2^t$ . We extend the backbone network  $\mathcal{F}_{b_2}$  to take the previous hidden state as input, which can be formulated by:

$$\hat{\mathbf{I}}_2^t = \mathcal{F}_{b_2}(\mathbf{H}^{t-1}, \hat{\mathbf{I}}_1^t, \hat{\mathbf{I}}_3^t). \quad (9)$$

Besides synthesizing the target frames, the ConvLSTM module also requires to maintain its cell state for temporal recurrence. We formulate the updating equation of the inter-pyramid recurrent module by:

$$\mathbf{H}^t, \mathbf{C}^t = \mathcal{F}_c(\hat{\mathbf{I}}_3^t, \mathbf{H}^{t-1}, \mathbf{C}^{t-1}), \quad (10)$$

where  $\mathcal{F}_c$  refers to the ConvLSTM unit,  $\mathbf{C}^{t-1}$  and  $\mathbf{C}^t$  are previous cell state and current cell state,  $\mathbf{H}^t$  refers to the current hidden state, and  $\hat{\mathbf{I}}_3^t$  denotes the current input. At time  $t$  and  $t+1$ , we obtain  $\{\hat{\mathbf{I}}_1^t, \hat{\mathbf{I}}_2^t\}$  and  $\{\hat{\mathbf{I}}_3^t, \hat{\mathbf{I}}_4^t\}$ , respectively. By extending the iteration to time  $T$ , we can synthesize all the deblurred and interpolated frames  $\hat{\mathbf{I}}_{1:1:2N}$ .

Following the computation flow of BIN<sub>2</sub>, we can extend networks with larger scales (e.g., BIN<sub>3</sub>, BIN<sub>4</sub>). The network with large scales can utilize a wide receptive field and a broad temporal scope to exploit time information, which can synthesize temporally smooth results.

#### 4.3. Implementation Details

**Temporal Skip Connection.** We use multiple identity skip connections to pass the pre-stage frame information into later backbone networks, as shown in Figure 3(a). We use identity skip connections to regulate the flow of frame signals for better gradient backward propagation. Take BIN<sub>3</sub> as an example, the identity skip connections concatenate the inputs  $\{\mathbf{B}_2, \mathbf{B}_4\}$  and the synthesized frames  $\{\hat{\mathbf{I}}_2, \hat{\mathbf{I}}_4\}$  to help the network  $\mathcal{F}_{b_3}$  synthesize the frame  $\hat{\mathbf{I}}_3$ .

**Backbone Network.** We use a variation of the residual dense network [41] as the backbone network. As shown in Figure 4, the backbone module consists of one Down-Shuffle layer and one UpShuffle layer [29], six convolutional layers, and six residual dense blocks [41]. The residual dense block consists of four  $3 \times 3$  convolutional layers,

one  $1 \times 1$  convolutional layer, and four ReLU activation layers. All of the hierarchical features from the residual dense blocks are concatenated for successive network modules.

**Loss Function.** Our loss function consists of two terms including the pixel reconstruction and cycle-consistency loss:

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_c. \quad (11)$$

*Pixel reconstruction loss*  $\mathcal{L}_p$  measures the overall pixel difference between the ground-truth frames  $\mathbf{G}_n^t$  and the reconstructed frames  $\hat{\mathbf{I}}_n^t$ :

$$\mathcal{L}_p = \frac{1}{T} \sum_{t=1}^T \sum_{n=1}^{2M-1} \rho \left( \hat{\mathbf{I}}_n^t - \mathbf{G}_n^t \right), \quad (12)$$

where  $\rho(x) = \sqrt{x^2 + \epsilon^2}$  is the Charbonnier penalty function [5].  $T$  represents the iterations executed on the recurrent module. We use *cycle consistency loss*  $\mathcal{L}_c$  to ensure the spatial consistency between temporary inputs  $\mathbf{I}'_n^t$  and the re-generated frames  $\hat{\mathbf{I}}_n^t$  in the pyramid architecture:

$$\mathcal{L}_c = \frac{1}{T} \sum_{t=1}^T \sum_{n \in \Omega} \rho \left( \mathbf{I}'_n^t - \hat{\mathbf{I}}_n^t \right), \quad (13)$$

where  $\Omega$  is the index of all cycle-paired frames.

**Training Dataset.** We use the Adobe240 dataset [30] to train the proposed network. It consists of 120 videos at 240 fps with the resolution of  $1280 \times 720$ . We use 112 of the videos to construct the training set. The following discrete degradation model is used to generate the training data:

$$\mathbf{B}_{2i} = \frac{1}{2\tau + 1} \sum_{j=iK-\tau}^{j=iK+\tau} \mathbf{L}_j, \quad i = 0, 1, \dots, N, \quad (14)$$

where  $\mathbf{L}_i$  is the  $i$ -th high-frame-rate latent image,  $\mathbf{B}_{2i}$  is the  $i$ -th acquired low-frame-rate blurred frame, the parameter  $K$  determines the frame rate of acquired frames,  $2\tau + 1$  corresponds to the equivalent long exposure time, that restricts the degree of blur [4]. We down-sample the high-frame-rate sequences to generate ground-truth frames. The frame rate of the ground-truth sequence is two times that of the blurry sequence. We use Equation (14) with parameters  $K = 8$  and  $\tau = 5$  to generate the training data. The resolution of training images is  $640 \times 352$ . Considering the computational complexity, we choose the temporal length of  $T = 2$ . We augment the training data by horizontal and vertical flipping, randomly cropping as well as reversing the temporal order of the training samples.

**Training Strategy.** We utilize the AdaMax [13] optimizer with parameters  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . We use a batch size of 2, and the initial learning rate is  $1e^{-3}$ . We train the model for 40 epochs, then reduce the learning rate by a factor of 0.2 and fine-tune the entire model for another 5 epochs. We train the network on an RTX-2080 Ti GPU card. It takes about two days to converge.

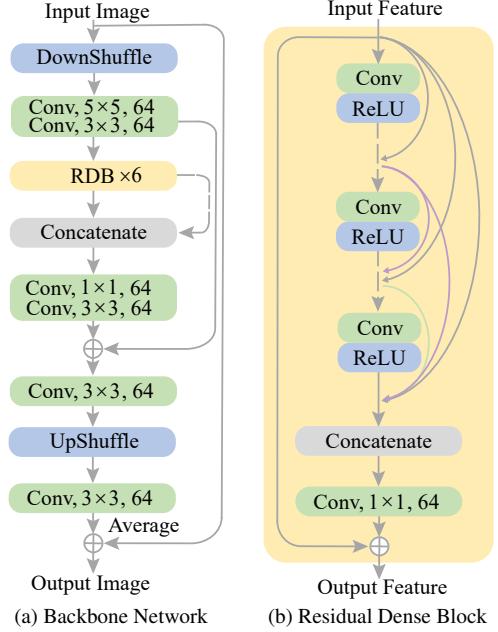


Figure 4. **Architecture of the backbone network.** We use a DownShuffle layer in the backbone network to distribute the motion information into multiple channels. We use residual dense blocks to learn hierarchical features.

## 5. Experimental Results

In this section, we first introduce the evaluation datasets and then conduct ablation studies to analyze the contribution of each proposed component. Finally, we compare the proposed model with state-of-the-art algorithms.

### 5.1. Evaluation Datasets and Metrics

We evaluate the proposed model on two video datasets and measure the motion smoothness of the synthesized video sequences for a comprehensive understanding.

**Adobe240.** We use 8 videos of the Adobe240 dataset [30] for evaluation. Each video comes at 240 fps with the resolution of  $1280 \times 720$ .

**YouTube240.** We download 59 slow-motion videos from the YouTube website to construct our YouTube240 evaluation dataset. The videos are of the same resolution and frame rate with Adobe240. For both Adobe240 and YouTube240 datasets, we use Equation (14) with parameter  $K = 8$  and  $\tau = 5$  to generate the evaluation data. All of the frames are resized to  $640 \times 352$ .

**Motion Smoothness.** Our motion smoothness metric is based on optical flow estimation [10, 31]. We first compute the differential optical flow  $\mathbf{D}$  using three inputs  $\mathbf{I}_{0:1:2}$  and three reference frames  $\mathbf{R}_{0:1:2}$  with the following equation:

$$\mathbf{D} = (\mathbf{F}_{\mathbf{I}_1 \rightarrow \mathbf{I}_2} - \mathbf{F}_{\mathbf{I}_0 \rightarrow \mathbf{I}_1}) - (\mathbf{F}_{\mathbf{R}_1 \rightarrow \mathbf{R}_2} - \mathbf{F}_{\mathbf{R}_0 \rightarrow \mathbf{R}_1}), \quad (15)$$

where  $\mathbf{F}_{\mathbf{x} \rightarrow \mathbf{y}}$  is the estimated optical flow from frame  $\mathbf{x}$  to frame  $\mathbf{y}$ . We use the state-of-the-art PWC-Net [31]



Figure 5. **Effect of network scales.** The model with larger scales can generate clear and sharper content.

algorithm to estimate optical flow. PWC-Net integrates the classic pyramid processing, flow warping and cost volume filtering techniques into a convolutional neural network framework. Our motion smoothness metric is defined by:

$$\mathcal{M}(s) = \log \sum_{d \in \mathbf{D}} \mathbf{1}_{[s, s+1]}(\|d\|_2) - \log |\mathbf{D}| \quad (16)$$

where  $d$  denotes the 2-dimensional vector of matrix  $\mathbf{D}$ ,  $|\mathbf{x}|$  represents the size of the matrix  $\mathbf{x}$ , and the indicator function  $\mathbf{1}_{\mathbf{A}}(x)$  equals to 1 if  $x$  belongs to set  $\mathbf{A}$ . The  $\mathcal{M}(s)$  measures the motion smoothness of three consecutive input frames concerning the pixel error length  $s$ , and lower  $\mathcal{M}(s)$  indicates better performance.

## 5.2. Model Analysis

To analyze the contributions of the proposed pyramid module, inter-pyramid recurrent module, ConvLSTM unit, and cycle consistency loss, we perform the following extensive experiments:

**Architecture Scalability.** We first investigate the scalability of the pyramid module by evaluating networks with three different scales ( $\text{BIN}_2$ ,  $\text{BIN}_3$ ,  $\text{BIN}_4$ ). We show the quantitative results in Table 1, and provide the visual comparisons in Figure 5. We find that the module using larger scales generates more clear details in Figure 5. We observe that with the parameters of  $\text{BIN}$  increasing from 2.29, 3.49 to 4.68 million, the networks steadily obtain better PSNR results from 31.87dB, 32.39dB to 32.59dB on the Adobe240

Table 1. **Analysis on network scales and recurrent module.** The numbers in **red** and **blue** represent the best and second-best results.

Method	Runtime (seconds)	Parameters (million)	Adobe240 [30]		YouTube240	
			PSNR	SSIM	PSNR	SSIM
$\text{BIN}_2$ - <i>w/o rec</i>	0.01	2.27	31.37	0.9129	34.10	0.9374
$\text{BIN}_3$ - <i>w/o rec</i>	0.06	3.44	31.67	0.9181	34.54	0.9392
$\text{BIN}_4$ - <i>w/o rec</i>	0.12	4.62	32.06	0.9190	34.72	0.9411
$\text{BIN}_2$	0.02	2.29	31.87	0.9183	34.41	0.9400
$\text{BIN}_3$	0.10	3.49	<b>32.39</b>	<b>0.9212</b>	<b>34.77</b>	<b>0.9419</b>
$\text{BIN}_4$	0.28	4.68	<b>32.59</b>	<b>0.9258</b>	<b>35.10</b>	<b>0.9443</b>

Table 2. **Analysis on ConvLSTM unit.** We evaluate three variations, including the model using LSTM ( $\text{BIN}_2$  -*LSTM*), the model using ConvLSTM ( $\text{BIN}_2$  -*ConvLSTM*), and the model without using recurrent model ( $\text{BIN}_2$  -*None*).

Method	Adobe240 [30]		YouTube240	
	PSNR	SSIM	PSNR	SSIM
$\text{BIN}_2$ - <i>None</i>	30.39	0.8974	33.32	0.9263
$\text{BIN}_2$ - <i>LSTM</i>	<b>31.38</b>	<b>0.9120</b>	<b>34.07</b>	<b>0.9283</b>
$\text{BIN}_2$ - <i>ConvLSTM</i>	<b>31.87</b>	<b>0.9183</b>	<b>34.41</b>	<b>0.9400</b>

dataset. However, the runtime costs also increase from 0.02, 0.10, to 0.28 seconds. The comparisons show that the pyramid module is scalable, where the scales balance the computational complexity (execution time and model parameters) and restoration quality.

**Inter-Pyramid Recurrent Module.** We then study the contributions of the proposed recurrent module by evaluating the model using recurrent modules and the model without using recurrent modules (i.e.,  $\text{BIN}_l$  versus  $\text{BIN}_l$  -*w/o rec*,  $l = 2, 3, 4$ ). In Table 1, we find that  $\text{BIN}_4$  obtains a better SSIM of 0.9258 than the SSIM of 0.9212 achieved by  $\text{BIN}_4$  -*w/o rec* in the Adobe240 set. The model using the recurrent module improves the restoration performance, it achieves about 0.5dB gain in the Adobe240 set and 0.3dB gain in the YouTube240 set.

**ConvLSTM Module.** To analyze the contribution of ConvLSTM unit, we evaluate the model using LSTM ( $\text{BIN}_2$  -*LSTM*), ConvLSTM ( $\text{BIN}_2$  -*ConvLSTM*), and that without using any recurrent unit ( $\text{BIN}_2$  -*None*). The  $\text{BIN}_2$  -*None* directly concatenates previous frames to propagate information recurrently. The results in Table 2 show that the ConvLSTM unit performs better than the LSTM unit as well as the model without using recurrent unit. The ConvLSTM unit provides about 0.49dB PSNR gain in the Adobe240 set and 0.34dB gain in the YouTube240 set.

**Cycle Consistency Loss.** Finally, we compare the model with cycle loss ( $\text{BIN}_4$  -*w/ cycle loss*) versus the model without cycle loss ( $\text{BIN}_4$  -*w/o cycle loss*). On the Adobe240 dataset, the PSNR of the model w/ and w/o cycle loss is 32.59dB and 32.42dB, respectively. Namely, the cycle loss

Table 3. Quantitative comparisons on the Adobe240 [30] and YouTube240 EVALUATION sets.

Method	Runtime (seconds)	Parameters (million)	Deblurring				Interpolation				Comprehensiveness			
			Adobe240 [30]		YouTube240		Adobe240 [30]		YouTube240		Adobe240 [30]		YouTube240	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Blurry Inputs	—	—	28.68	0.8584	31.96	0.9119	—	—	—	—	—	—	—	—
Super SloMo	—	39.6	—	—	—	—	27.52	0.8593	30.84	0.9107	—	—	—	—
MEMC-Net	—	70.3	—	—	—	—	30.83	0.9128	34.91	0.9596	—	—	—	—
DAIN	—	24.0	—	—	—	—	31.03	0.9172	35.06	0.9615	—	—	—	—
EDVR + Super SloMo	0.42	63.2					27.79	0.8671	31.15	0.9136	30.28	0.9003	32.91	0.9292
EDVR + MEMC-Net	0.27	93.9	32.76	0.9335	34.66	0.9448	30.22	0.9058	33.49	0.9367	31.49	0.9197	34.08	0.9408
EDVR + DAIN	1.13	47.6					30.28	0.9070	33.53	0.9378	31.52	0.9203	34.10	0.9413
SRN + Super SloMo	0.27	47.7					27.22	0.8454	30.42	0.8970	28.32	0.8604	31.21	0.9044
SRN + MEMC-Net	0.22	78.4	29.42	0.8753	32.00	0.9118	28.25	0.8625	31.60	0.9107	28.84	0.8689	31.80	0.9113
SRN + DAIN	0.79	32.1					27.83	0.8562	31.15	0.9059	28.63	0.8658	31.58	0.9089
Jin [10]	0.25	10.8	29.40	0.8734	32.06	0.9119	29.24	0.8754	32.24	0.9140	29.32	0.8744	32.15	0.9130
BIN <sub>4</sub> (Ours)	0.28	4.68	32.67	0.9236	35.10	0.9417	32.51	0.9280	35.10	0.9468	32.59	0.9258	35.10	0.9443

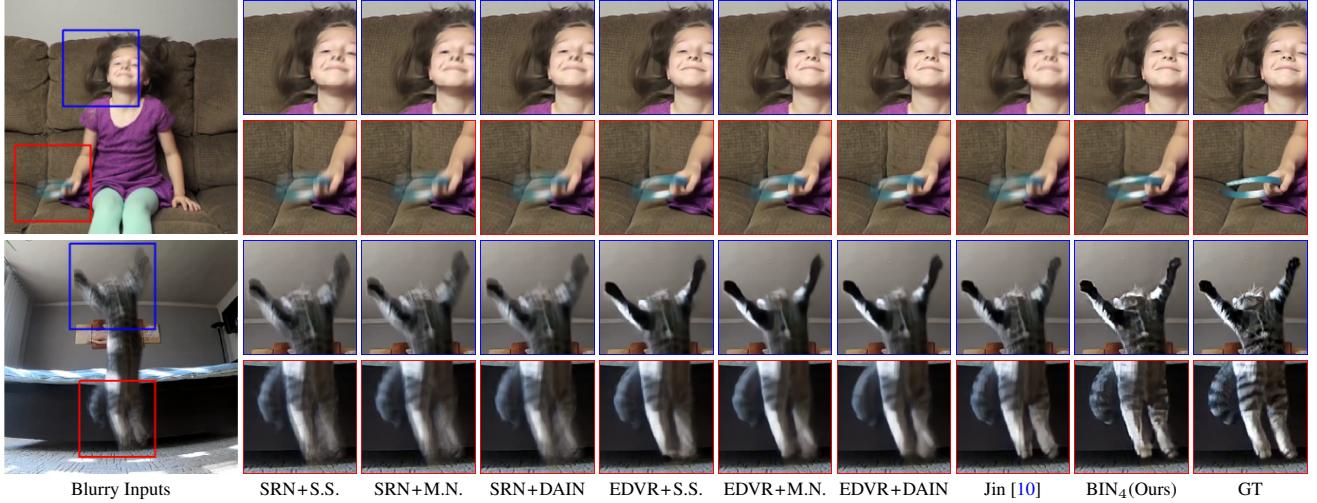


Figure 6. **Visual comparisons on the YouTube240 EVALUATION set.** The pictures in the first two rows and the last two rows show the deblurred frames and the interpolated frames, respectively. Our method generates clearer and sharper content. S.S. is short for Super SloMo [9] and M.N. is short for MEMC-Net [2].

provides 0.17dB gain. The results demonstrate that the cycle loss ensures consistency of frames and it helps the model to generate fine details of moving objects.

### 5.3. Compare with the State-of-the-arts

We evaluate the proposed method against the algorithm proposed by Jin *et al.* [10]. Their model synthesizes nine intermediate frames using two blurred inputs. We extract the center interpolated frame to compare with our results. Besides, we construct several cascade methods by connecting deblurring and interpolation models, including EDVR [35], SRN [32] for deblurring, and Super SloMo [9], MEMC [2], DAIN [1] for interpolation. We compare our model with the state-of-the-art algorithms in the following aspects:

**Interpolation Evaluation.** As shown in Table 3 and Figure 6, our model performs favorably against all the compared methods. Moreover, we find that our model performs better than the frame interpolation method using *sharp* frames (e.g., DAIN). For example, the PSNR of our model is 32.51dB, while the PSNR of DAIN is 31.03dB on the Adobe240 dataset. The main reason is that one blurred frame contains information of multiple sharp frames, and our method synthesizes the intermediate frame using several blurred frames, while the interpolation method only uses two sharp frames. Therefore, our model can exploit more space-time information from multiple blurred frames, resulting in more satisfactory intermediate frames.

**Deblurring Evaluation.** We then compare the deblurring aspects with the state-of-the-art methods. As shown in Ta-

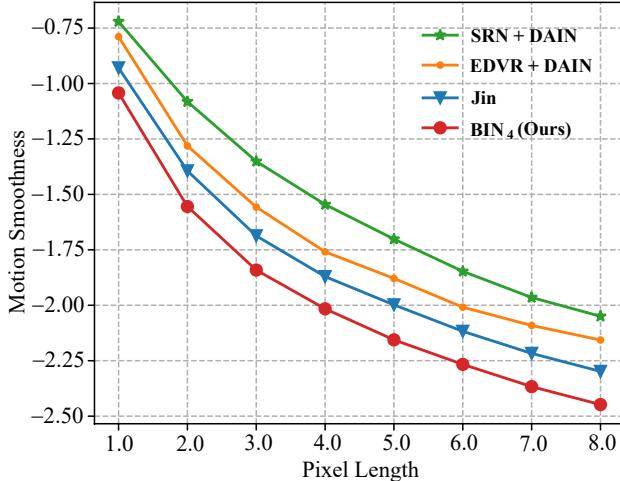


Figure 7. **Motion smoothness comparisons on the Adobe240 EVALUATION set.** The proposed model reaches the best performance in terms of motion smoothness. The smoothness metric is displayed in logarithm to show the difference.

ble 3, our model performs slightly inferior to the state-of-the-art EDVR algorithm. Our model achieves 0.09dB less than EDVR in terms of PSNR, but our model size (4.68 million) is much smaller than that of the EDVR (23.6 million), and our model requires less execution time.

**Comprehensive Evaluation.** We compare the comprehensive performance of deblurring and interpolation. A high-performance pre-deblurring model in the cascade method helps the subsequent interpolation network to restore better results. As shown in Table 3, the SRN model performs slightly inferior to the EDVR. Thus the EDVR + DAIN has a better performance than SRN + DAIN. However, the best-performing cascade method (EDVR + DAIN) is still sub-optimal in terms of the overall performance. The overall PSNR of EDVR + DAIN is 31.52dB, while our model obtains PSNR of 32.59dB on the Adobe240 dataset.

Compared to Jin *et al.* [10]’s method, our approach obtains up to 3.27dB gain on the Adobe240 dataset. Their training dataset has less fast-moving screens and camera shakes than the Adobe240 dataset. Therefore, the Adobe240 dataset has a severer blur than Jin’s training dataset. We note that Jin *et al.* do not publish their training code at the time of submission. We cannot optimize their model on the Adobe240 dataset for fair comparisons. Nevertheless, compared with their method, our network benefits from scalable structure and recurrent information propagation, thus obtains significant performance gains.

**Motion Smoothness Evaluation.** We compare the motion smoothness performance based on the metric introduced in Section 5.1. A lower metric indicates a better performance. As shown in Figure 7, Jin [10]’s model performs favorably against all the cascade methods (we show SRN +

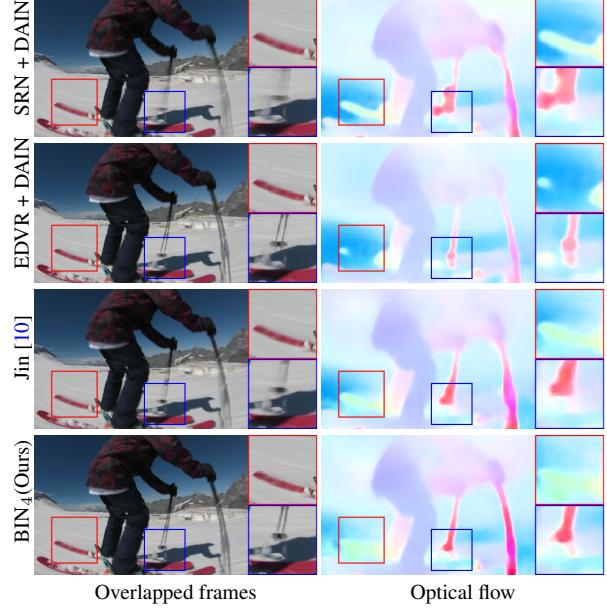


Figure 8. **Visual comparisons on the YouTube240 dataset.** We use the PWC-net [31] to estimate optical flow of two adjacent output frames. The optical flow of our model has smoother shapes.

DAIN and EDVR + DAIN for brevity), and our algorithm has a better smoothness metric than Jin’s model. In Figure 8, the optical flow of our model has smoother shapes compared with the cascade methods. Our network is a unified model with a broad temporal scope, which helps generate smooth frames. Besides, compared to the approximate recurrent mechanism of Jin [10], our proposed inter-pyramid recurrent module adopts ConvLSTM cells to propagate the frame information across time. It can further enforce temporal consistency between the deblurred and interpolated frames. Thus, our method is superior to all the cascade methods and Jin’s model.

## 6. Conclusion

In this work, we propose a blurry video frame interpolation method to address the joint video enhancement problem. Our model consists of pyramid modules and inter-pyramid recurrent modules. The pyramid module is scalable, where the scales balance the computational complexity and restoration quality. We use the cycle consistency loss to ensure the consistency of inter-frames in the pyramid module. Furthermore, the inter-pyramid recurrent module utilizes the spatial-temporal information to generate temporally smoother results. Extensive quantitative and qualitative evaluations demonstrate that the proposed method performs favorably against the existing approaches.

**Acknowledgment.** To be add.

## References

- [1] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. Depth-aware video frame interpolation. In *CVPR*, 2019. [1](#), [2](#), [7](#)
- [2] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang. MEMC-Net: Motion Estimation and Motion Compensation Driven Neural Network for Video Interpolation and Enhancement. *TPAMI*, 2019. [1](#), [2](#), [7](#)
- [3] W. Bao, X. Zhang, L. Chen, L. Ding, and Z. Gao. High-order model and dynamic filtering for frame rate up-conversion. *TIP*, 2018. [1](#)
- [4] T. Brooks and J. T. Barron. Learning to synthesize motion blur. In *CVPR*, 2019. [5](#)
- [5] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICCV*, 1994. [5](#)
- [6] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman. Temporal cycle-consistency learning. In *CVPR*, 2019. [2](#)
- [7] T. Hyun Kim and K. Mu Lee. Generalized video deblurring for dynamic scenes. In *CVPR*, 2015. [2](#)
- [8] T. Hyun Kim, K. Mu Lee, B. Scholkopf, and M. Hirsch. Online video deblurring via dynamic temporal blending network. In *CVPR*, 2017. [2](#)
- [9] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018. [1](#), [2](#), [7](#)
- [10] M. Jin, Z. Hu, and P. Favaro. Learning to extract flawless slow motion from blurry videos. In *CVPR*, 2019. [2](#), [5](#), [7](#), [8](#)
- [11] M. Jin, G. Meishvili, and P. Favaro. Learning to extract a video sequence from a single motion-blurred image. In *CVPR*, 2018. [2](#)
- [12] T. H. Kim, S. Nah, and K. M. Lee. Dynamic video deblurring using a locally adaptive blur model. *TPAMI*, 2017. [2](#)
- [13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. [5](#)
- [14] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. [2](#)
- [15] W. H. Lee, K. Choi, and J. B. Ra. Frame rate up conversion based on variational image fusion. *TIP*, 2013. [2](#)
- [16] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu. Feedback network for image super-resolution. In *CVPR*, 2019. [2](#)
- [17] Y.-L. Liu, Y.-T. Liao, Y.-Y. Lin, and Y.-Y. Chuang. Deep video frame interpolation using cyclic frame generation. In *AAAI*, 2019. [2](#)
- [18] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala. Video frame synthesis using deep voxel flow. In *CVPR*, 2017. [2](#)
- [19] G. Long, L. Kneip, J. M. Alvarez, H. Li, X. Zhang, and Q. Yu. Learning image matching by simply watching video. In *ECCV*, 2016. [2](#)
- [20] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, and C. Schroers. Phasenet for video frame interpolation. In *CVPR*, 2018. [2](#)
- [21] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. [2](#)
- [22] S. Nah, S. Son, and K. M. Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *CVPR*, 2019. [2](#)
- [23] S. Niklaus and F. Liu. Context-aware synthesis for video frame interpolation. In *CVPR*, 2018. [2](#)
- [24] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive convolution. In *CVPR*, 2017. [2](#)
- [25] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive separable convolution. In *ICCV*, 2017. [2](#)
- [26] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *CVPR*, 2019. [2](#)
- [27] F. A. Reda, D. Sun, A. Dundar, M. Shoeybi, G. Liu, K. J. Shih, A. Tao, J. Kautz, and B. Catanzaro. Unsupervised video interpolation using cycle consistency. In *ICCV*, 2019. [2](#)
- [28] M. S. Sajjadi, R. Vemulapalli, and M. Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. [2](#)
- [29] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. [4](#)
- [30] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *CVPR*, 2017. [1](#), [2](#), [5](#), [6](#), [7](#)
- [31] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. [5](#), [8](#)
- [32] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. [1](#), [2](#), [7](#)
- [33] J. Telleen, A. Sullivan, J. Yee, O. Wang, P. Gunawardane, I. Collins, and J. Davis. Synthetic shutter speed imaging. In *CGF*, 2007. [1](#), [3](#)
- [34] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv*, 2018. [2](#)
- [35] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPR*, 2019. [1](#), [2](#), [7](#)
- [36] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015. [2](#)
- [37] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019. [2](#)
- [38] Y. Yuan, S. Liu, J. Zhang, Y. Zhang, C. Dong, and L. Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *CVPR*, 2018. [2](#)
- [39] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese. Feedback networks. In *CVPR*, 2017. [2](#)
- [40] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li. Adversarial spatio-temporal learning for video deblurring. *TIP*, 2018. [2](#)

- [41] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. [2](#), [4](#)
- [42] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren. Spatio-temporal filter adaptive network for video deblurring. In *ICCV*, 2019. [2](#)