ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation

Akash Gupta
University of California, Riverside
agupt013@ucr.edu

Abhishek Aich
University of California, Riverside
aaich001@ucr.edu

Amit K. Roy-Chowdhury
University of California, Riverside
amitrc@ece.ucr.edu

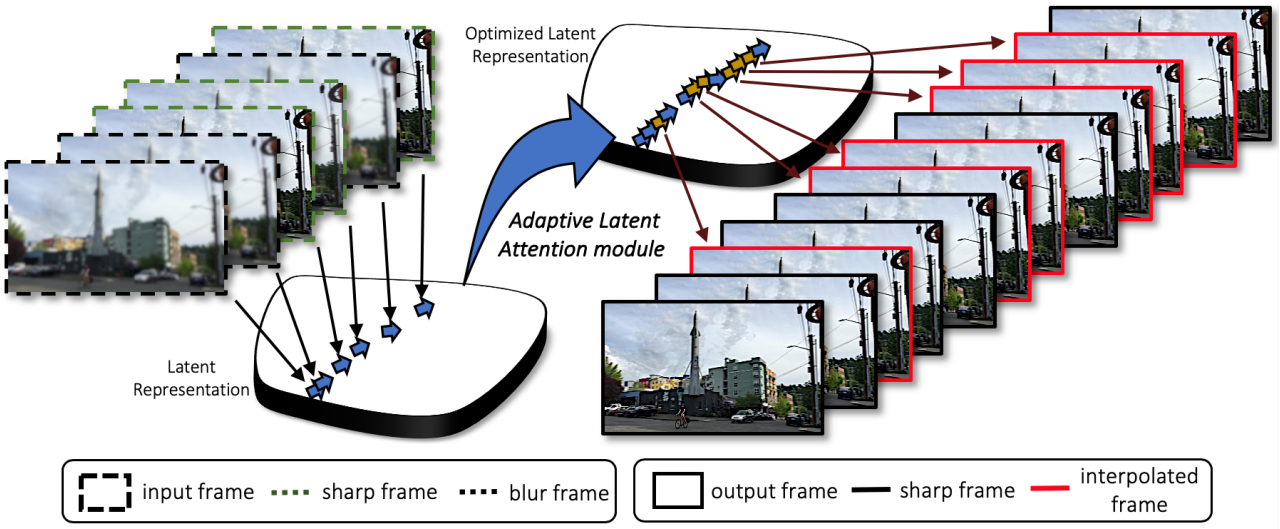


Figure 1: *Conceptual Overview of* ALANET. Given a poor-quality video consisting both blurry and sharp frames, the frames are projected on a latent space. These latent representations are modulated and interpolated using the proposed Adaptive Latent Attention module to generate optimized latent representations for deblurring and interpolation. These optimized representations are then used to generate a high frame-rate sharp video.

ABSTRACT

Existing works address the problem of generating high frame-rate sharp videos by separately learning the frame deblurring and frame interpolation modules. Most of these approaches have a strong prior assumption that all the input frames are blurry whereas in a real-world setting, the quality of frames varies. Moreover, such approaches are trained to perform either of the two tasks - deblurring or interpolation - in isolation, while many practical situations call for both. Different from these works, we address a more realistic problem of high frame-rate sharp video synthesis with no prior assumption that input is always blurry. We introduce a novel architecture, Adaptive Latent Attention Network (ALANET), which synthesizes sharp high frame-rate videos with no prior knowledge of input frames being blurry or not, thereby performing the task of both deblurring and interpolation. We hypothesize that information from the latent representation of the consecutive frames can be utilized to generate optimized representations for both frame deblurring and frame interpolation. Specifically, we employ combination of self-attention and cross-attention module between consecutive frames in the latent space to generate optimized representation for each frame. The optimized representation learnt using these attention modules help the model to generate and interpolate sharp frames. Extensive experiments on standard datasets demonstrate that our method performs favorably against various state-of-the-art approaches, even though we tackle a much more difficult problem. The project page is available at `https://agupt013.github.io/ALANET.html`

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3413686

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

KEYWORDS

Video Synthesis, Interpolation, Deblurring, Cross-Attention, Self-Attention, Generative Model

ACM Reference Format:
Akash Gupta, Abhishek Aich, and Amit K. Roy-Chowdhury. 2020. ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413686

1 INTRODUCTION

Motion blur and low frame-rate are often commonplace in videos captured by mobile devices, whether hand-held or on a moving platform. The reasons vary, including low shutter frequency, long exposure times, and the movement of the device itself [10, 28]. These factors limit the quality of videos captured. As vast majority of video media is captured using mobile cameras these days, it calls for improved quality of the video captured by these devices. Enhancing video quality requires restoring the degradation caused by motion blur along with increase in the frame-rate for temporal smoothness.

Most existing approaches have addressed the problem of high frame-rate sharp video generation by frame deblurring and frame interpolation, separately. In [10], separate models are used to deblur input frames and to interpolate between frames. The phenomenon of motion blur and frame-rate at which video is captured are related. Thus, a joint formulation is needed when addressing the task of high frame-rate sharp video generation from a low frame-rate blurry video. Recently, [25] studied the problem of joint video deblurring and interpolation. Here, authors proposed to use pyramid deep models to deblur and interpolate along with a pyramid of convolutional Long-Short Term Memory (LSTM) to capture temporal smoothness. However, these methods assume that all the input frames are blurry, which is often unrealistic because the quality of a video usually varies non-uniformly over time.

In this paper, we introduce a novel architecture **A**daptive **L**atent **A**ttention **NET**work (**ALANET**) which aims to jointly deblur and interpolate frames from a poor quality video input without an assumption that all input frames are blurry. Specifically, we construct a Adaptive Latent Attention module that leverages the latent space with attention mechanisms to generate high frame-rate sharp video. **ALANET** has a U-Net variant [24] as it's backbone, combined with the proposed attention module. Similar to U-Net, we utilize contracting path (encoder) of the network for latent space representation and expanding path (generator) for video generation. However unlike U-Net, we do not pass the bottleneck features extracted from the encoder directly to the generator. We introduce our proposed adaptive attention module to modulate and interpolate the latent features for deblurring and interpolating frames from the input video. Figure 1 illustrates the concept of proposed adaptive attention module. Given a set of input blurry and sharp frames, their projection in latent space can be modulated and interpolated using Adaptive Latent Attention module, to generate optimized representations for sharp frames. These modulated and interpolated latent representations are then used by the generator to synthesize the high frame-rate sharp video.

1.1 Approach Overview

An overview of our approach is illustrated in Figure 2. Given a low frame-rate poor quality input, our objective is to generate a high frame-rate sharp video. Our proposed architecture, **ALANET**, consists of three modules: the frame encoding network $\mathcal{E}$, the Adaptive Latent Attention network $\mathcal{M}$, and the high frame-rate sharp video generator $\mathcal{G}$. We modulate and interpolate the frame features by applying **self-attention** and **cross-attention** on channels of the latent features of consecutive frames using our proposed adaptive attention module. Self-attention on the feature space helps the model to focus on important features of the same frame whereas cross-attention helps the model to retrieve information from neighbouring frames that can be useful for either deblurring or interpolation tasks. In turn, the Adaptive Latent Attention module will give less importance to the neighbouring frame feature if the input is a sharp frame, and utilize this information from the neighbours if input frame is blurry. Hence, our proposed approach is able to deblur and generate high quality interpolated frames using self-attention and cross-attention on frame representations. To the best of our knowledge, *our approach is the first work to exploit the ability of learning optimized latent representation for generation of high frame-rate sharp video using self-attention and cross-attention.*

1.2 Contributions

The key contributions of our proposed framework are summarized as follows.

- We introduce a novel framework **ALANET**, Adaptive Latent Attention Network, designed to jointly deblur and interpolate for high frame-rate visually sharp video generation.
- This is the first work to generate high frame-rate sharp video from low frame-rate poor quality video by applying attention in the latent space without any assumption on the uniformity of blurriness in different frames of the video.
- Our framework demonstrates consistently effective results on two datasets, the benchmark Adobe240 and crawled YouTube240 with better or at par performance with state-of-the-art in both deblurring and interpolation tasks.

2 RELATED WORK

Our work relates to research in video deblurring, video interpolation, attention model, and joint video deblurring and interpolation. In this section, we discuss some representative methods closely related to our work (see Table 1).

Video Deblurring. Inversion of motion blur is an ill-posed problem [21, 23]. Recent works have used deep learning based methods to solve this restoration problem either using a single frame [26, 27] or multiple frames [7, 10, 13, 18, 26]. [5] attempts to deblur a video by exploring similarity between the frames of the video and exploiting sharp patches of neighbouring frames. DeBlurNet [26] proposes to use consecutive frames stacked as input to generate a single clean central frame. ESVR [30] tries to align the features of multiple frames using a temporal and spatial fusion module for feature fusion from different layer to deblur a video. [12] proposes an integrated model to jointly predict the defocus blur, optical flow and latent frames. [8] proposed a spatio-temporal recurrent neural network that enforces temporal consistency between neighbouring frames. [35] proposes a spatio-temporal recurrent architecture with dynamic temporal blending mechanism. In contrast, we do not estimate any extra information like optical flow (which can be noisy and computationally heavy) in our approach and rely on proposed attention model to generate high frame-rate sharp videos.

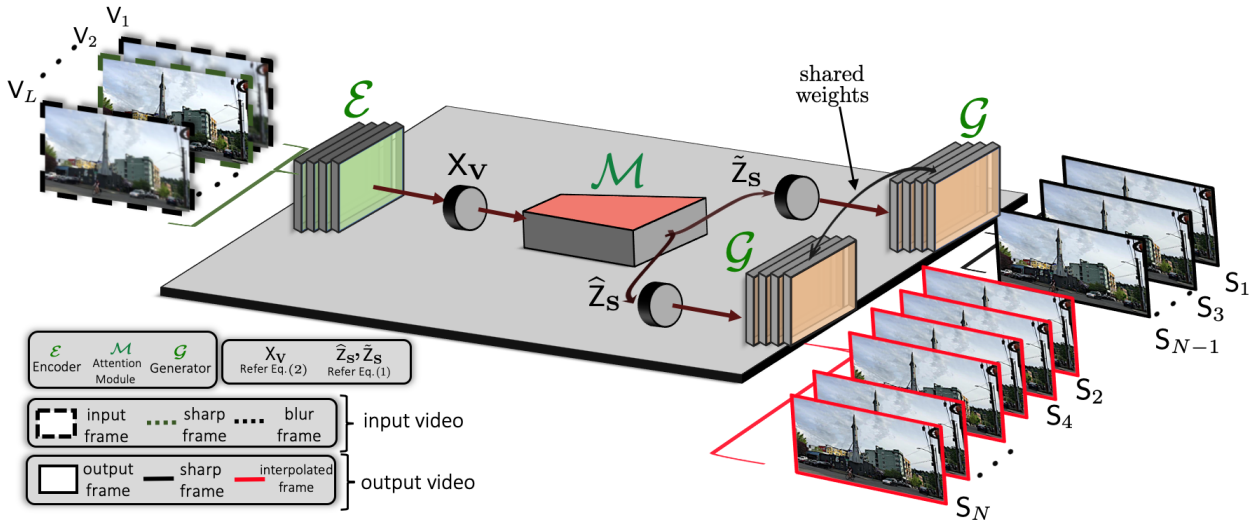


Figure 2: ***Architectural Overview of*** **ALANET. Given a low frame-rate poor quality video $\mathbf{V} = [V_1, V_2, \cdots, V_L]$, we extract latent representations $X_V = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L]$ using encoder network $\mathcal{E}$. Adaptive Latent Attention module $\mathcal{M}$ utilizes combination of self-attention and cross-attention on X_V to generate optimized representations for deblurring ($\widetilde{Z}_S$) and interpolation ($\widehat{Z}_S$). These optimized representations are used by the generative network $\mathcal{G}$ to synthesize deblurred frames ($S_1, S_3, \cdots, S_{N-1}$) from $\widetilde{Z}_S$ and interpolated frames ($S_2, S_4, \cdots, S_N$) from $\widehat{Z}_S$, thereby generating a high frame-rate video $\mathbf{S} = [S_1, S_2, \cdots, S_N]$.**

Video Interpolation. Many of the existing approaches [3, 4, 9, 15, 17, 36] for frame interpolation use optical flow estimation between input frames. Consequently, the quality of estimated optical flow governs the quality of frame interpolation. Recent learning based methods have demonstrated effectiveness in frame interpolation tasks. A direct application of convolutional neural networks (CNNs) for intermediate frame synthesis is presented in [16]. Some methods [19, 20] apply CNNs to estimate space-varying and separable convolutional kernels for synthesis using neighbourhood pixels. [1] proposes to generate videos by learning optimized representation by a non-adversarial approach and then interpolating between the optimized latent representation of two frames to synthesize central frame. However, they average the latent representations of two frames for frame interpolation which often generates a blurry image. Unlike these methods, our approach utilizes adaptive attention in the latent space for interpolation.

Attention Model. Attention mechanism has garnered a lot of interest due to their learnable guidance ability. With pioneering work in language translation [29], variations of attention mechanism have shown promising results in object recognition [2], image generation [32] and image super-resolution [33]. Residual channel attention mechanism for super-resolution is introduced in [33]. Authors in [31] used different length sequences to deblur the center frame and attention is applied on different outputs to generate a single central frame. Recently, variations of attention models are proposed for video deblurring [31] and video interpolation [6]. In [6], attention is applied channel-wise on concatenated down-shuffled frames for video interpolation. In contrast to our work, where we apply attention in latent space, the existing methods employ attention for video deblurring and interpolation tasks in pixel space.

Joint Video Deblurring and Interpolation. Joint video deblurring and interpolation still remains a challenging problem. [10] proposed DeBlurNet, to deblur, and InterpNet, for interpolating input frames in a jointly optimized cascade scheme to generate sharp slow motion videos using blurry input. Blurry Video Frame Interpolation proposed in [25] uses pyramid structure to deblur and interpolate along with a pyramid convolutional LSTM to capture temporal information. However, both these methods strongly assume that all the input frames are blurry. We relax this assumption to address a more difficult problem where we do not know which input frames are blurry and where to interpolate. Hence, the proposed **ALANET** framework is *self-sufficient to make decisions on which frames to deblur using information from neighbouring frames.*

Table 1: ***Categorization of prior works in video deblurring and interpolation.*** **Different from the state-of-the-art approaches, ALANET demonstrates adaptive attention in latent space to perform joint deblurring and interpolation.**

Methods	Settings			
	Interpolate?	Deblur?	Joint Deblur & Interpolate?	Latent Attention?
DAIN [3]	✓	✗	✗	✗
Jin [10]	✓	✓	✗	✗
BIN [25]	✓	✓	✓	✗
ALANET (Ours)	✓	✓	✓	✓

3 PROBLEM FORMULATION

Given a low frame-rate poor quality video $\mathbf{V} = \left[\mathsf{V}_1, \mathsf{V}_2, \cdots, \mathsf{V}_L \right]$, with L frames, we aim to generate a high frame-rate sharp video $\mathbf{S} = \left[\mathsf{S}_1, \mathsf{S}_2, \cdots, \mathsf{S}_N \right]$ with N frames, where $N > L$. Our objective is to deblur and increase the frame-rate of the given input video $\mathbf{V}$. Corresponding to each input frame $\mathsf{V}_i \, \forall \, i = 1, 2, \cdots, L$, let there be a feature representation $\mathbf{x}_i$ in latent space $\mathsf{X} \in \mathbb{R}^{H_1 \times W_1 \times C_1 \times L}$ such that $\mathsf{X}_\mathbf{V} = \left[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L \right]$ where $H_1 \times W_1 \times C_1$ is the dimension of the latent representation.

We propose to generate a high frame-rate video by adaptive attention modeling (see Section 4.2) of the feature representations of input video frames in the latent space. Our hypothesis is that in latent space, information from neighbouring frames can help learn optimized representations for deblurring and interpolation. Thus, the proposed Adaptive Latent Attentive model transforms input blurry frame representation ($\mathsf{X}_\mathbf{V}$) to the optimized representations ($\mathsf{Z}_\mathbf{S} \in \mathbb{R}^{H_1 \times W_1 \times C_1 \times N}$) for deblurring and interpolation in the latent space given by

$$\mathsf{Z}_\mathbf{S} = \left[\mathbf{z}_1, \widehat{\mathbf{z}}_2, \mathbf{z}_3, \widehat{\mathbf{z}}_4, \cdots, \mathbf{z}_N \right] = \widetilde{\mathsf{Z}}_\mathbf{S} \bigcup \widehat{\mathsf{Z}}_\mathbf{S} \tag{1}$$

where $\mathbf{z}_{2i}$ is the representation for a deblurred frame S_{2i}, and $\widehat{\mathbf{z}}_{2i+1}$ is the representation for an interpolated frame between S_{2i} and S_{2i+2}, i.e., S_{2i+1}. We denote all latent representations for deblurred frames by $\widetilde{\mathsf{Z}}_\mathbf{S}$ and for interpolated frames by $\widehat{\mathsf{Z}}_\mathbf{S}$. These optimized representations $\mathsf{Z}_\mathbf{S} = \widetilde{\mathsf{Z}}_\mathbf{S} \bigcup \widehat{\mathsf{Z}}_\mathbf{S}$ are used to deblur and interpolate sharp frames to generate a high frame-rate video.

4 ALANET: ADAPTIVE LATENT ATTENTION NETWORK

In this section, we describe the proposed framework, **ALANET**, in detail. Our framework consists of three components: the encoder $\mathcal{E}$, the Adaptive Latent Attention module $\mathcal{M}$ and, the generator $\mathcal{G}$. We use the encoder module to extract latent representation for each input frame. The Adaptive Latent Attention module generates optimized representations for frames to reduce blur and to interpolate frames, simultaneously. Finally, the optimized representations are used by the generator to synthesize a high frame-rate sharp video. Our overall framework is shown in Figure 2.

4.1 Latent Representation of Frames

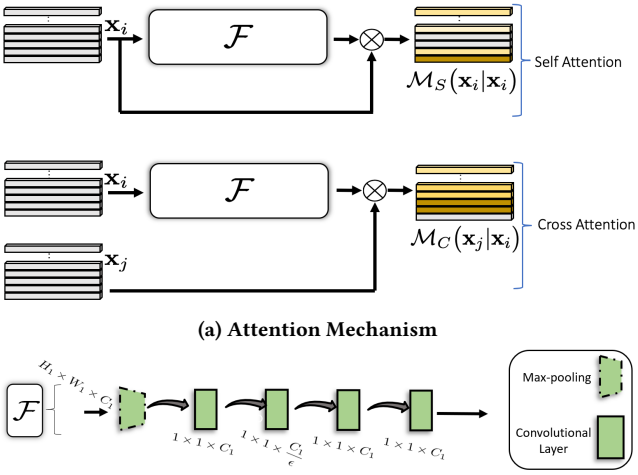
The encoder $\mathcal{E}$ is a trainable convolutional neural network which projects the input video into a latent representation for each frame.

$$\begin{aligned} \mathcal{E}(\mathbf{V}) &= \mathcal{E}\Big(\left[\mathsf{V}_1, \mathsf{V}_2, \cdots, \mathsf{V}_L \right] \Big) \\ &= \left[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L \right] = \mathsf{X}_\mathbf{V} \end{aligned} \tag{2}$$

Here, $\mathbf{x}_i \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ is the latent representation corresponding to V_i. The representations generated by the encoder $\mathcal{E}$ are used by the Adaptive Latent Attention module $\mathcal{M}$ to generate optimized representations for deblurring and interpolation.

4.2 Adaptive Latent Attention

The latent representation of a frame generated by the encoder may not be optimized as all the channels of the input representation are not equally important for generation task. Also, since frames of a video are temporally correlated, their latent representation can be leveraged to extract information from neighbouring frames to generate an optimized representation for deblurring and interpolation.



(a) Attention Mechanism

(b) Channel Attention Computation Network

Figure 3: *Proposed Attention Module.* (a) Self-Attention (*top*) on latent representation $\mathbf{x}_i$ and Cross-Attention (*bottom*) for representation $\mathbf{x}_j$ conditioned on $\mathbf{x}_i$. Symbol $\otimes$ denotes element-wise multiplication of each attention weight with respective channel of the representation. (b) The channel weight computation function $\mathcal{F}$. It generates channel descriptor by channel-wise global average pooling to learn attention weights for each channel.

To extract important information from the latent representation of the given frame and utilize the information from the neighbouring frames, we propose an Adaptive Latent Attention module $\mathcal{M}$. The proposed module $\mathcal{M}$ applies attention on the input latent representations to generate the optimized representations for deblurring and interpolation. This module takes two latent representations ($\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{H_1 \times W_1 \times C_1}$) as input, where $H_1 \times W_1$ is dimension of each feature in C_1 channels of the latent representation. A combination of **self-attention** $\mathcal{M}_S$ and **cross-attention** $\mathcal{M}_C$ is then used to generate latent representations to jointly deblur and interpolate between two consecutive frames in an adaptive manner.

The basic building block of the attention mechanism is the channel attention function $\mathcal{F}$. It computes attention weights of each channel in the latent representation. As in [33], the channel-wise global spatial information is extracted using global average pooling to condense input features to a channel descriptor. Then, a gating mechanism is applied to learn non-linear interactions and correlation between multi-channel features such that $\mathcal{F} : \mathbb{R}^{H_1 \times W_1 \times C_1} \rightarrow \mathbb{R}^{1 \times 1 \times C_1}$, where $H_1 \times W_1 \times C_1$ is the dimension of the latent representation. Figure 3 shows the self-attention $\mathcal{M}_S$ and cross-attention $\mathcal{M}_C$ modules along with the basic building block $\mathcal{F}$ for computation of the channel attention.

Self-Attention ($\mathcal{M}_S$) correlates different channels of the latent representation of a frame in order to generate an informative representation. This is achieved by computing attention weights for each

of the channels of the input representation followed by element-wise multiplication of the channels with their attention weights. This self-attention on $\mathbf{x}_i$ can then be expressed as in (3).

Cross-Attention ($\mathcal{M}_C$) provides attention weights for each channel of the latent representation $\mathbf{x}_j$ conditioned on another latent representation $\mathbf{x}_i$. Cross-attention leverages information from other frames to generate a conditional representation. The conditional representation provides insight on what information is useful from other frames. This cross-attention on $\mathbf{x}_j$ given the input $\mathbf{x}_i$ can then be computed as in (4).

$$\mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) = \mathbf{x}_i \otimes \mathcal{F}(\mathbf{x}_i) \tag{3}$$

$$\mathcal{M}_C\Big(\mathbf{x}_j|\mathbf{x}_i\Big) = \mathbf{x}_j \otimes \mathcal{F}(\mathbf{x}_i) \tag{4}$$

Note that, $\otimes$ in (3) and (4) represents element-wise multiplication, $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ are the encoded feature representations of frames and $\mathcal{M}_\mathbf{S}(\mathbf{x}_i|\mathbf{x}_i),\ \mathcal{M}_\mathbf{C}(\mathbf{x}_j|\mathbf{x}_i) \in \mathbb{R}^{H_1 \times W_1 \times C_1}$.

Deblurred and Interpolated Representations. A combination of self-attention and cross-attention modules is employed to obtain optimized latent representations for deblurring and interpolation. Given a window $\mathbf{W}$, the optimized latent representations $\mathrm{Z}_\mathbf{V} = \big[\ \mathbf{z}_1,\ \widehat{\mathbf{z}}_2,\ \mathbf{z}_3,\ \widehat{\mathbf{z}}_4, \cdots,\ \mathbf{z}_N\ \big]$ for a high frame-rate video $\mathbf{S}$ is computed as follows:

$$\mathbf{z}_{2i} = \mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) + \sum_{j \in \mathbb{Q}} \mathcal{M}_C\Big(\mathbf{x}_j|\mathbf{x}_i\Big) \tag{5}$$

$$\begin{aligned}\widehat{\mathbf{z}}_{2i+1} = {} & \mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) + \mathcal{M}_C\Big(\mathbf{x}_i|\mathbf{x}_{i+1}\Big) \\ & + \mathcal{M}_S\Big(\mathbf{x}_{i+1}|\mathbf{x}_{i+1}\Big) + \mathcal{M}_C\Big(\mathbf{x}_{i+1}|\mathbf{x}_i\Big)\end{aligned} \tag{6}$$

where $\mathbb{Q}$ denotes integer values in $[\ i - 0.5\mathbf{W},\ i\ ) \bigcup (\ i,\ i + 0.5\mathbf{W}\ ]$, $\mathbf{z}_{2i}$ is the optimized representation for deblurred frame S_{2i} and $\widehat{\mathbf{z}}_{2i+1}$ is the optimized representation for the interpolated frame between S_{2i} and S_{2i+2}.

As defined by (5), an optimized representation $\mathbf{z}_{2i}$ for sharp output S_{2i} is computed using self-attention on i^{th} input representation $\mathbf{x}_i$ and cross-attention of all the remaining input latent representation $\mathbf{x}_j$ in a neighbourhood of $\mathbf{W}$ frames. Cross-attention is computed in a temporal window of $\mathbf{W}$ frames as the significant information for deblurring and interpolation is available in neighbouring frames compared to temporally distant frames. Similarly, a latent representation $\widehat{\mathbf{z}}_{2i+1}$ for interpolated frame S_{2i+1} between S_{2i} and S_{2i+2} is given by (6), where we consider self-attention on each latent representations $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, and cross-attention for each representation conditioned on the other.

4.3 High Frame-Rate Video Generation

To generate a high frame-rate video from blurry inputs, we employ a generative neural network $\mathcal{G}$ that transforms the optimized representations to a sequence of frames. The optimized representations generated by the adaptive attention module $\mathcal{M}$ are used by generator $\mathcal{G}$ to synthesize deblurred frames as well as interpolate between frames represented by $\mathbf{S} = \mathcal{G}\big(\big[\ \mathbf{z}_1,\ \widehat{\mathbf{z}}_2,\ \mathbf{z}_3,\ \widehat{\mathbf{z}}_4, \cdots,\ \mathbf{z}_N\ \big]\big)$ where $\mathbf{z}_{2i}$ and $\widehat{\mathbf{z}}_{2i+1}$ are optimized representation used to deblur and interpolate frames S_{2i} and S_{2i+1}, respectively.

4.4 Network Architecture

In this section, we describe the network architecture used for different modules in the proposed **ALANET** framework.

Encoder-Generator Network. A variation of U-Net [9] is employed to design the backbone network for the proposed framework. The contracting path is used as the encoder network $\mathcal{E}$ and the expansive path is used as the generator network $\mathcal{G}$. The encoder-decoder network also retains the skip-connections as in the original U-Net architecture [24]. However unlike the U-Net architecture, our proposed Adaptive Latent Attention module $\mathcal{M}$ is introduced after the bottleneck to optimize the latent representations before they are fed to the generator $\mathcal{G}$.

Adaptive Latent Attention Network. In order to make the generator model, $\mathcal{G}$, focus more on informative features, we exploit the inter-dependencies within frame feature (self-attention) and across frame features (cross-attention). The basic building block of self-attention and cross-attention is the attention weight computation module, $\mathcal{F}$. We adopt the channel attention module as in [33] for $\mathcal{F}$. This channel attention module first extracts the channel-wise global spatial information into a channel descriptor using global average pooling. Then, a gating mechanism is applied to learn non-linear interactions and non-mutually-exclusive relationship between multi-channel features [33]. Unlike self-attention for super-resolution in [33], we also employ cross-attention between consecutive features to learn interactions between these features for deblurring and interpolation.

5 EXPERIMENTS

In this section, we first introduce the benchmark datasets, and evaluation metrics. Next, the model used for generation of blurry training data is described. Finally, extensive experiments are shown to demonstrate the effectiveness of our proposed approach in generating high frame-rate sharp videos.

5.1 Datasets and Metrics

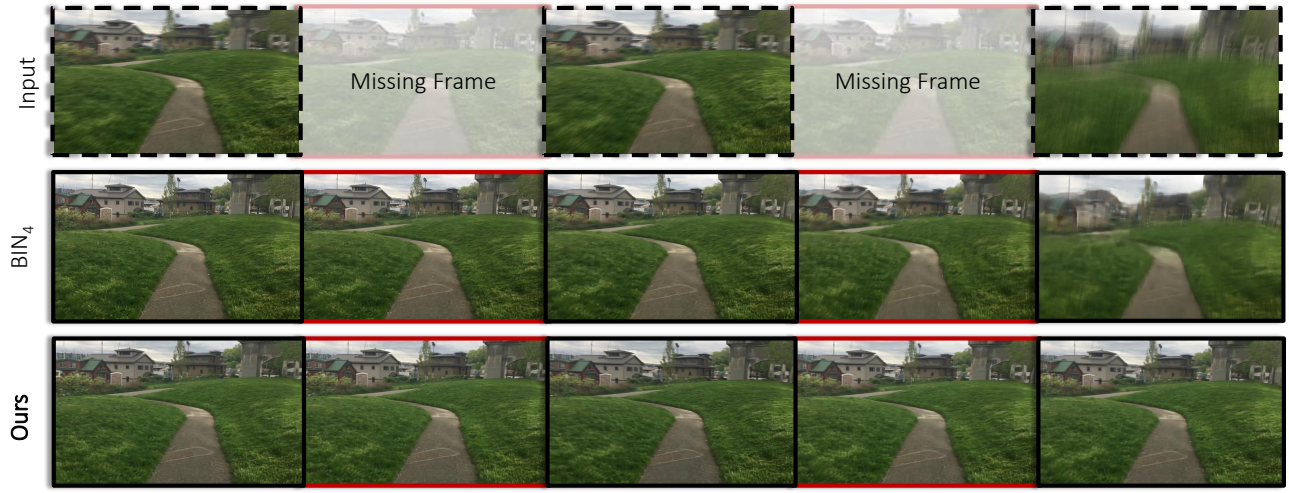
We evaluate the performance of our approach using publicly available Adobe240 [26] dataset which has been used in many prior works and a dataset crawled from YouTube as in [25].

Adobe240 Dataset. This dataset contains 118 videos captured at 240 frames per second (fps) with the resolution of 1280×720 . We choose 110 videos for training and remaining 8 for evaluation following the split provided in [9] for fair comparison.

YouTube240 Dataset. We download 60 random video videos captured at 240fps from the YouTube website to construct an evaluation dataset similar to that used in [25]. The resolution of the downloaded video is 1280×720. For this dataset, we train the model in Adobe240 but test on YouTube240 without any fine-tuning.

Dataset Preparation. For Adobe240 [26] and crawled YouTube240 dataset, low frame-rate poor quality videos of 30fps are generated using process described in section 5.2. All the frames are resized to 640×352 for training and evaluation purposes.

(a) ***Representative result from*** **Adobe240** ***dataset.*** **Observe zoomed-in patch of the car. The motion of car introduces motion blur. ALANET is able to significantly reduce the motion blur in all the frames and also generate superior quality interpolated frames.**



(b) ***Representative result from*** **Adobe240** ***dataset.*** **The last frame in blurry input (top row) is of poor quality. ALANET is able to deblur and interpolate clear frame (last two frame in the bottom row) as compared to the state-of-the-art (last two frame in the middle row).**

Figure 4: ***Qualitative result comparison with the state-of-the-art.*** **Top row consists of the input blurry frames and the missing frames faded. We show two high frame-rate videos generated by our proposed method (bottom row) and compare it with the state-of-the-art BIN$_4$ (middle row). ALANET is able to generate superior quality high frame-rate video.**

5.2 Implementation Details

Our framework is implemented in PyTorch [22]. All the experiments are trained for 200 epochs with a batch size of 2. We use ADAM [14] optimizer with initial learning rate of 0.0001 and weight decay 5×10^{-4}. The learning rate is reduced by a factor of 10 after 100 and 150 epochs. The proposed framework takes a 30fps blurry video as an input and generates a 60fps sharp video.

Blurry Video Formation. Camera shutter frequency affects degradation due to motion blur in each frame of a captured video. A low shutter frequency may not be able to capture temporal smoothness and hence generate blurry frames. To simulate the motion blur, we approximate the blurry frame as a discrete averaging of sharp frames within an overlapping window as defined in [9, 10, 26]. Let $2\tau + 1$ be the number of sharp frames between two blurry frames and β be the rate at which frames are captured. Then, a blurry frame V_i is approximated as:

$$V_i = \frac{1}{2\tau + 1} \sum_{k=i\beta-\tau}^{i\beta+\tau} S_k \tag{7}$$

where, S_k's are the sharp frames in the given video. Since we do not assume that all the input frames are blurry, we average 11 consecutive frames randomly using (7) on a sharp video to generate a poor quality video with low frame-rate.

Table 2: *Quantitative results comparison on* Adobe240 *and* YouTube240. We obtained better average PSNR and SSIM index on Adobe240 dataset. Our proposed approach performs at-par on YouTube240 dataset when evaluated using the model trained on Adobe240. Best scores have been highlighted in bold. † indicates results reported from [25].

Method	Deblurring				Interpolation				Joint Deblurring and Interpolation			
	Adobe240		YouTube240		Adobe240		YouTube240		Adobe240		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† [9]	-	-	-	-	27.52	0.8593	30.84	0.9107	-	-	-	-
MEMC-Net† [4]	-	-	-	-	30.83	0.9128	34.91	0.9596	-	-	-	-
DAIN† [3]	-	-	-	-	31.03	0.9172	35.09	0.9615	-	-	-	-
Jin† [10]	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_4† [25]	32.67	0.9236	35.10	0.9417	32.51	0.9280	35.10	0.9468	32.59	0.9258	35.10	0.9443
ALANET (Ours)	**33.71**	**0.9429**	35.94	0.9496	**32.98**	**0.9362**	35.85	0.9513	**33.34**	**0.9355**	35.89	0.9504

Training and Testing Protocol. During training, random blurry frames are generated on-the-fly by averaging 11 frames as defined in (7). The 5^{th} and 9^{th} sharp frames are considered as the ground-truth for deblurring and interpolation, respectively. The framework is jointly optimized for deblurring and interpolation using Adaptive Latent Attention Network. During testing, a low frame-rate (30fps) poor quality video is used as an input to the trained model and a high frame-rate (60fps) sharp video is generated.

Objective Function. Our objective function consists of a ℓ_1 pixel reconstruction loss[1] and the perceptual loss [11] defined as follows.

$$\mathcal{L} = \mathcal{L}_r + \lambda \mathcal{L}_p \tag{8}$$

Here, $\mathcal{L}_r = \sum_i |G_i - S_i|_1$ denotes ℓ_1 reconstruction loss with G_i being the ground-truth frame corresponding to the generated frame S_i. $\mathcal{L}_p$ denotes the perceptual loss computed using a pre-trained VGG16 network [11], and λ is a hyper-parameter. We use $\lambda = 0.2$ for all our experiments.

5.3 Qualitative Results

Figure 4 shows some examples of high frame-rate videos generated using the proposed method and state-of-the-art BIN_4 [25] given a low frame-rate video (top row). From Figure 4a, it can be seen that our approach is able to tackle the motion blur introduced due to the object motion (car in the bottom left corner for this particular example) along with the blur produced by averaging of consecutive sharp frames. As our approach is extracting information by applying attention on latent representation of input frame, our method is able to deblur and interpolate visually more appealing videos. In Figure 4b, the last two frames of middle and bottom row show that the proposed method is able to deblur and interpolate visually good quality frames whereas BIN_4 generates a blurry interpolated frame. As the BIN_4 utilizes the deblurred frame to interpolate, the error from deblurred frame may propagate during interpolation and hence produce a blurry interpolated frame as shown in Fig 4b (middle row, last frame). Our approach overcomes this by generating optimized representation using attention mechanisms, which extracts relevant information from neighbouring frames in the latent space for both deblurring and interpolation.

[1]For pixel reconstruction loss, we choose ℓ_1-loss instead of Mean-Squared Error (MSE) ℓ_2 loss as latter has inherent property of generating blurry output as shown in the literature [34].

5.4 Quantitative Results

Our proposed method performs joint deblurring and interpolation. There are several methods that only solve the tasks of either deblurring or interpolation. We compare our proposed approach with these state-of-the-art methods that either perform deblurring or interpolation [3, 4, 9] given an input blurry video. We also compare **ALANET** with two recent approaches where deblurring and interpolation is performed jointly [10, 25]. Quantitative result comparison with these baselines are shown in Table 2.

Results on Adobe240 Dataset. For deblurring task on Adobe240 dataset, we report a relative improvement of 1.04dB in the average PSNR value and 2.09% improvement in SSIM metric when compared to [25]. Our method achieves 32.98dB average PSNR in interpolation task as opposed 32.51dB reported by state-of-the-art method BIN_4 [25]. Overall, for the joint task of deblurring and interpolation the proposed method achieves relative improvement of 2.3% in average PSNR and 1.04% in SSIM index against BIN_4. It can be observed that BIN_4 and **ALANET** both jointly formulate the deblurring and interpolation tasks which helps to outperform [10]. We again highlight that our method does not know which frames are blurry or where to interpolate, unlike BIN_4 [25].

Results on YouTube240 Dataset. We evaluate the performance of our model trained on Adobe240 dataset for deblurring and interpolation on YouTube240 dataset. For this experiment we crawled 60 videos from YouTube to create this dataset following authors in [25]. However, we do not have the same set of videos as in [25] as the list of videos is not publicly available. From Table 2, it can be

Figure 5: *Ablation study on different attention modules.* Frame generated using different attention mechanisms (top) and the residue image (bottom) computed by taking its difference with the ground-truth frame. Scale for the error range [0. 255] is given on the bottom left. Our proposed ALANET which combines self-attention and cross-attention produces superior results compared to using only one of the attention mechanisms. Results best viewed when zoomed-in.

observed that network trained on Adobe240 performs at-par when evaluated on YouTube240 dataset with average PSNR of 35.89dB and SSIM index of 0.9504 for joint deblurring and interpolation.

5.5 Ablation Study

In this section, we investigate the contribution of self-attention and cross-attention in the proposed approach. First, we study the impact of self-attention on video deblurring and interpolation. We remove the cross-attention $\mathcal{M}_C$ terms from (5) and (6) and train the network using only self-attention in the latent space. Secondly, we study the impact of cross-attention in absence of self-attention by removing $\mathcal{M}_S$ terms from (5) and (6) for training the network.

Figure 5 presents the qualitative results of the ablation study. It can be observed that the network trained using only self-attention produces inferior results as compared to that of using only cross-attention. The network trained with only self-attention module assumes that all the information to deblur and interpolate resides in a single frame and discards the temporal information available in consecutive frames. This loss in information results in poor quality frame when using only self-attention. On the other hand, using only cross-attention produces better results than using only self-attention module as it exploits the available temporal information by applying cross-attention on latent representation of the consecutive frames.

The quantitative results of impact of different attention mechanisms are shown in Table 3. Network trained on only cross-attention achieves improvement of 0.38dB PSNR as compared to using only self-attention for deblurring. However, for interpolation there is improvement of 1.90dB when using only cross-attention as, unlike self-attention, it exploits the temporal information available from neighbouring frames. From Table 3, we can observe that **ALANET** performs best as it extracts quality information from the latent representation by exploiting combination of self-attention and cross-attention for deblurring and interpolation.

Table 3: *Ablation study on attention mechanism.* We evaluate contribution of self-attention and cross-attention for high frame-rate video generation on Adobe240 dataset.

Attention	Deblurring		Interpolation	
	PSNR	SSIM	PSNR	SSIM
only Self-Attention	31.98	0.9373	30.87	0.9233
only Cross-Attention	32.36	0.9385	32.77	0.9340
ALANET	33.71	0.9429	32.98	0.9362

6 CONCLUSION

We present an Adaptive Latent Attention Network (**ALANET**) for generating high frame-rate sharp videos with no knowledge that either an input frame is blurry or not. The proposed approach employs self-attention and cross-attention mechanism in the latent representations of input video frames for deblurring and interpolation. Specifically, the self-attention module extracts information local to the input frame and the cross-attention module exploits the temporal relationship from latent representations of neighbouring frame. Using combination of self-attention and cross-attention our approach is able to generate high frame-rate sharp video. Experiments on standard datasets show the efficacy of our proposed attention module in task of joint deblurring and interpolation over state-of-the-art methods.

ACKNOWLEDGMENTS

This work was partially supported by NSF grants 33397 and 33425, ONR grant N00014-18-1-2252, and a gift from CISCO. We thank Padmaja Jonnalagedda, JVL Venkatesh and JV Megha for valuable discussions and feedback on the paper.

REFERENCES

[1] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M Salman Asif, and Amit K Roy-Chowdhury. 2020. Non-Adversarial Video Synthesis with Learned Priors. *arXiv preprint arXiv:2003.09565* (2020).

[2] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755* (2014).

[3] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. Depth-aware video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3703–3712.

[4] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. 2019. MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[5] Sunghyun Cho, Jue Wang, and Seungyong Lee. 2012. Video deblurring for hand-held cameras using patch-based synthesis. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–9.

[6] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. 2020. Channel attention is all you need for video frame interpolation. AAAI.

[7] Tae Hyun Kim and Kyoung Mu Lee. 2015. Generalized video deblurring for dynamic scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5426–5434.

[8] Tae Hyun Kim, Kyoung Mu Lee, Bernhard Scholkopf, and Michael Hirsch. 2017. Online video deblurring via dynamic temporal blending network. In *Proceedings of the IEEE International Conference on Computer Vision*. 4038–4047.

[9] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9000–9008.

[10] Meiguang Jin, Givi Meishvili, and Paolo Favaro. 2018. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6334–6342.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.

[12] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. 2016. Dynamic scene deblurring using a locally adaptive linear blur model. *arXiv preprint arXiv:1603.04265* (2016).

[13] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Dynamic video deblurring using a locally adaptive blur model. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2374–2387.

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

[15] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. 2017. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*. 4463–4471.

[16] Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. 2016. Learning image matching by simply watching video. In *European Conference on Computer Vision*. Springer, 434–450.

[17] Dhruv Mahajan, Fu-Chung Huang, Wojciech Matusik, Ravi Ramamoorthi, and Peter Belhumeur. 2009. Moving gradients: a path-based method for plausible image interpolation. *ACM Transactions on Graphics (TOG)* 28, 3 (2009), 1–11.

[18] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3883–3891.

[19] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 670–679.

[20] Simon Niklaus, Long Mai, and Feng Liu. 2017. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 261–270.

[21] Thekke Madam Nimisha, Akash Kumar Singh, and Ambasamudram N Rajagopalan. 2017. Blur-invariant deep learning for blind-deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*. 4752–4760.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *NIPS AutoDiff Workshop*.

[23] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. 2006. Coded exposure photography: motion deblurring using fluttered shutter. In *ACM SIGGRAPH 2006 Papers*. 795–804.

[24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[25] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. 2020. Blurry Video Frame Interpolation. arXiv:cs.CV/2002.12259

[26] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. 2017. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1279–1288.

[27] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. 2018. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8174–8182.

[28] Jacob Telleen, Anne Sullivan, Jerry Yee, Oliver Wang, Prabath Gunawardane, Ian Collins, and James Davis. 2007. Synthetic shutter speed imaging. In *Computer Graphics Forum*, Vol. 26. Wiley Online Library, 591–598.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[30] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[31] Junru Wu, Xiang Yu, Ding Liu, Manmohan Chandraker, and Zhangyang Wang. 2020. DAVID: Dual-Attentional Video Deblurring. In *The IEEE Winter Conference on Applications of Computer Vision*. 2376–2385.

[32] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* (2018).

[33] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 286–301.

[34] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2015. Loss functions for neural networks for image processing. *arXiv preprint arXiv:1511.08861* (2015).

[35] Shangchen Zhou, Jiawei Zhang, Jinshan Pan, Haozhe Xie, Wangmeng Zuo, and Jimmy Ren. 2019. Spatio-temporal filter adaptive network for video deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*. 2482–2491.

[36] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2004. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)* 23, 3 (2004), 600–608.