UNIVERSITY OF CALIFORNIA
RIVERSIDE

Video Enhancement with Internal Learning and Blind Priors

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering

by

Akash Ashok Gupta

December 2021

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Nael Abu-Ghazaleh
Dr. Evangelos Papalexakis

The Dissertation of Akash Ashok Gupta is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The work presented in this dissertation would not have been possible without the inspiration and support of a number of wonderful individuals. I express my sincere gratitude to all of them for supporting me and being part of this amazing journey and making this Ph.D. dissertation possible. First and foremost I am extremely grateful to my supervisors, Prof. Amit K. Roy-Chowdhury for his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. Through his constant guidance, I learned various aspects of doing good research including looking at problems holistically, identifying the critical assumptions in existing works to formulate and solve practical research problems, and how to approach a problem by critical thinking. I feel extremely fortunate and privileged to work under his supervision.

I would also like to express my heartfelt gratitude to my dissertation committee members, Dr. Nael Abh-Ghazaleh, and Dr. Evangelos Papalexakis for giving me valuable feedback and constructive comments in improving the quality of this dissertation. I would like thank my co-authors and research colleagues - Prof. B. S. Manjunath, Prof. Shiv Chandrasekharan, Prof. Bir Bhanu, Dr. Jingen Liu, Dr. Rameswar Panda, Dr. Niluthpol Chowdhury Mithun, Dr. Sujoy Paul, Dr. Lakshmanan Natarajan and Abhishek Aich for their stimulating discussions and constructive feedback. I would like to thank Dr. Ashwin Kothari, who instilled a sense of research in me.

I would like to thank the NSF (1664172, 1724341 and 1911197), ONR (N00014-18-1-2252), Cisco and Vimaan Robotics for their grants to Dr. Roy-Chowdhury, which partially supported my research. I thank Victor Hill for setting up the computing infrastructure used in most of the works presented in this dissertation. A hearty thank you to Kim Underhill for being an amazing advisor through these years. I am also grateful for advisors in the ECE Graduate Division, UCR Graduate Division and International Center for their help and support. I would also like to thank Dr. Jordan Edwards for his constant encouragement and support.

I am deeply indebted to my late grandmother Smt. Radhadevi and my late grandfather Adv. Mohanlalji Gupta for their heartfelt blessings, love, concern, care and affection towards me. My grandparents have been a source of inspiration and motivation to me. I owe my deepest gratitude towards my father Adv. Ashok Gupta and my mother Sarla Gupta for their unconditional trust, support and encouragement. I am also grateful to my loving younger brother Dr. Amol Gupta for love and encouragement. I would also like to thank my father-in-law Dr. J.V.L. Venkatesh, my mother-in-law Dr. Megha Jonnalagedda and my sister-in-law Priya Jonnalagedda for their love and constant support. I'm also grateful to have a wonderful support system of friends and family.

Finally, I thank with love to my wife. She has been my best friend and a great companion. Her unconditional love, support, encouragement and endless patience helped me get through the biggest milestone and difficult times in the most positive way.

Acknowledgment of previously published materials: The text of this dissertation, in part or in full, is a reprint of the material as appeared in four previously published papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all four publications, directed and supervised the research which forms the basis for this dissertation.

The papers are as follows:

1. Akash Gupta, Sudhir Singh, Amit K. Roy-Chowdhury, “Patch Attention Network for Joint Rolling Shutter Correction and Super-Resolution”, Preprint, 2021

2. Akash Gupta, Padmaja Jonnalagedda, Bir Bhanu, Amit K. Roy-Chowdhury, “AdaVSR: Adaptive Video Super-Resolution with Meta-Learning”, ACM International Conference on Multimedia (ACM-MM), 2021.

3. Akash Gupta, Abhishek Aich, Amit K. Roy-Chowdhury, “ALANET: Adaptive Latent Attention Network for Joint Video Deblurring and Interpolation”, ACM International Conference on Multimedia (ACM-MM), 2020.

4. Akash Gupta, Abhishek Aich, Amit K. Roy-Chowdhury, “Deep Quantized Representation for Enhanced Reconstruction”, IEEE International Symposium on Biomedical Imaging Workshops, 2020.

To my family for all the support.

ABSTRACT OF THE DISSERTATION

Video Enhancement with Internal Learning and Blind Priors

by

Akash Ashok Gupta

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, December 2021
Dr. Amit K. Roy-Chowdhury, Chairperson

With the increasing popularity of mobile cameras in various computer vision and multimedia applications, the demand for high-quality visual content is also increasing. However, videos captured using current consumer-grade cameras often suffer from a variety of quality issues such as motion blur, low frame-rate, low resolution, and rolling shutter artifacts. The reasons vary, including low shutter frequency, long exposure times, type of imaging sensors, and the movement of the device itself. These factors limit the quality of videos captured. As a vast majority of videos is captured using mobile cameras these days, it calls for improved quality of the video captured by these devices. In this thesis, we focus on enhancing the quality of videos by leveraging the spatio-temporal internal structure of the given video along with the external information available from the external dataset. Most of the existing works make a prior assumption on the degradation model that affects the quality of videos. Examples of such assumptions in degradation models include knowledge that all input frames are blurry, known degradation kernel for spatio-temporal down-sampling, and absence of rolling shutter artifacts. Nevertheless, in many real-world applications, these assumptions don't hold true as the input priors are usually unknown. We term these unknown priors as blind priors for the task of video enhancement. In this regard, we first present our work on joint video deblurring and interpolation with no prior assumption that input frames are always blurry. We utilize internal information available from neighbouring frames to deblur and interpolate between frames. Then, we describe our approach on blind spatio-temporal video super-resolution with the unknown down-sampling kernel, by leveraging an external dataset and internal structure of a given video. Next, we present our work on joint rolling shutter correction and super-resolution to recover the high-resolution global shutter video using patch-recurrence property in videos. Finally, we show an application of enhancement techniques in biomedical imaging, where we utilize quantization in feature space for unsupervised image denoising. We demonstrate that the proposed approaches effectively utilize the internal learning for the task of video enhancement and show impressive performance in different real-world blind prior settings.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With increasing popularity of high-performance higher resolution displays such as 4K Ultra HD (UHD), the consumer expectation for high quality visual content is also increasing [40, 51, 50]. Recently, with prevalence of mobile cameras, more and more videos are being captured using mobile devices. However, motion blur, low frame-rate, low resolution and rolling shutter artifacts are often commonplace in videos captured using these devices [106, 44]. Enhancing video quality at device level requires restoring the degradation caused by motion blur, increasing the frame-rate for temporal smoothness and rolling shutter rectification. Consequently, video super-resolution is necessary for the compatibility of videos captured using mobile devices with high-resolution displays for high perceptual quality. 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.

The success of deep learning methods and the availability large-scale datasets [97, 43, 119, 61] have greatly facilitated the research in video restoration techniques. Although the deep learning approaches have shown remarkable performance in individual video enhancement tasks, most of the existing works ignore the correlation between different factors affecting quality of any video. Moreover, these approaches often make unrealistic assumptions regarding the video quality degradation model such as the knowledge that all input frames are blurry [98, 105, 44, 53, 38, 98, 73, 128], fixed and known degradation kernel for spatio-temporal down-samplings [47, 45, 59, 119, 104, 10, 112], and absence of rolling shutter artifacts in CMOS cameras [47, 45, 119, 32]. The joint formulation of different degradation models with blind priors on image formation is hardly explored.

In this thesis, we address three novel video enhancement problems in realistic settings. In the first chapter, we study the task of high frame-rate sharp video generation. Existing works address the problem of generating high frame-rate sharp videos by separately learning the frame deblurring [98, 105, 44, 53, 38, 98, 73, 128] and frame interpolation [66, 130, 4, 5, 43, 64, 74, 75, 1, 65] 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 dataset and web-crawled dataset demonstrate that our method performs favorably against various state-of-the-art approaches, even though we tackle a much more difficult problem.

Most of the existing works in supervised spatio-temporal video super-resolution (STVSR) heavily rely on a large-scale external dataset consisting of paired low-resolution low-frame rate (LR-LFR) and high-resolution high-frame rate (HR-HFR) videos. Despite their remarkable performance, these methods make a prior assumption that the low-resolution video is obtained by down-scaling the high-resolution video using a known degradation kernel, which does not hold in practical settings [47, 45, 59, 119, 104, 10, 112]. Another problem with these methods is that they cannot exploit instance-specific internal information of a video at testing time. Recently, deep internal learning approaches have gained attention due to their ability to utilize the instance-specific statistics of a video. However, these methods have a large inference time as they require thousands of gradient updates to learn the intrinsic structure of the data. In the second chapter, to address these challenges in real-world video super-resolution task, we propose a novel Adaptive Video Super-Resolution (Ada-VSR) framework which leverages external as well as internal information through meta-transfer learning and internal learning, respectively. Specifically, meta-learning is employed to obtain adaptive parameters, using a large-scale external dataset, such that the obtained parameters can adapt quickly to the novel condition (degradation model) of the given test video during internal learning task, thereby exploiting external and internal information for video super-resolution task. The model trained using our approach can quickly adapt to a specific video condition with only a few gradient updates, which reduces the inference time significantly. Extensive experiments on standard datasets demonstrate that our method performs favorably against various state-of-the-art approaches in terms of perceptual quality and computational resources.

Our next work on video enhancement addresses the problem of rolling shutter correction and super-resolution. With the prevalence of CMOS cameras in many computer vision applications, there is increase in appearance of rolling shutter (RS) artifacts in captured videos. However, existing video super-resolution algorithms assume that the motion in the input video is global and no rolling shutter effect is present [47, 45, 119, 32]. The problem of video super-resolution for video captured using RS cameras is challenging as the model needs to learn the row-wise local pixel displacements and the global structure of the objects for RS correction and super-resolution, respectively. We propose Patch Attention Network (PatchNet) to address the problem of joint rolling shutter correction and super-resolution (RS-SR). Our conjecture is that the combination of information from the neighbouring patches in feature space can span more detailed feature space for the task of super-resolution. In particular, the Patch Attention Network leverages bi-directional motion information in feature space to extract relevant information from neighbouring patches using attention mechanism, and deformable fields using deformable convolution layers to extract local pixel-level information. We perform extensive experiments on real as well as synthetic datasets and demonstrate that our model is favourable against various benchmark baselines for the task of rolling shutter correction and super-resolution.

We also explore application of enhancement technique in biomedical imaging. In Chapter 5, we extend the process of quantization in the feature space and show that quantization, that is usually utilized to denoise any image in pixel space, can also be applied in the internal feature space for the task of unsupervised denoising.

In this thesis, we demonstrate that significant information regarding video enhancement is available within each video. We show that internal structure from neighbouring frames and patches can be utilized for video enhancement tasks. Additional, we show that external information available from external datasets can be leveraged to effectively tackle video enhancement problem in blind prior setup. Various experiments with combination of internal learning, blind prior and external dataset show promising results for various video enhancement tasks.

Organization of the Thesis. The rest of the thesis is organized as follows. In Chapter 2, we address the problem joint deblurring and interpolation using self-attention and cross-attention mechanisms in the latent representations. We present a novel meta-learning framework for blind spatio-temporal super-resolution where degradation kernel is not known in Chapter 3. We leverage meta-training using external dataset to learn a model that can easily adapt to unseen degradation models. Furthermore, we exploit the internal structure of the test video to adapt the model, trained using external learning, specific to the given video. In Chapter 4, we exploit the patch-recurrence property in frames to recover high-resolution global shutter frames from low-resolution rolling shutter video. In Chapter 5, we presented application of vector quantization in biomedical microscopy imaging for the task of unsupervised denoising. We conclude the thesis in Chapter 6 by providing some future directions related to the problem of video enhancement and video compression.

Chapter 2

Joint Video Deblurring and Interpolation

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.

2.1 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 [106, 44]. 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 videos captured by these devices. Enhancing video quality requires restoring the degradation caused by motion blur along with increase in the frame-rate at which video is captured 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 [44], 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

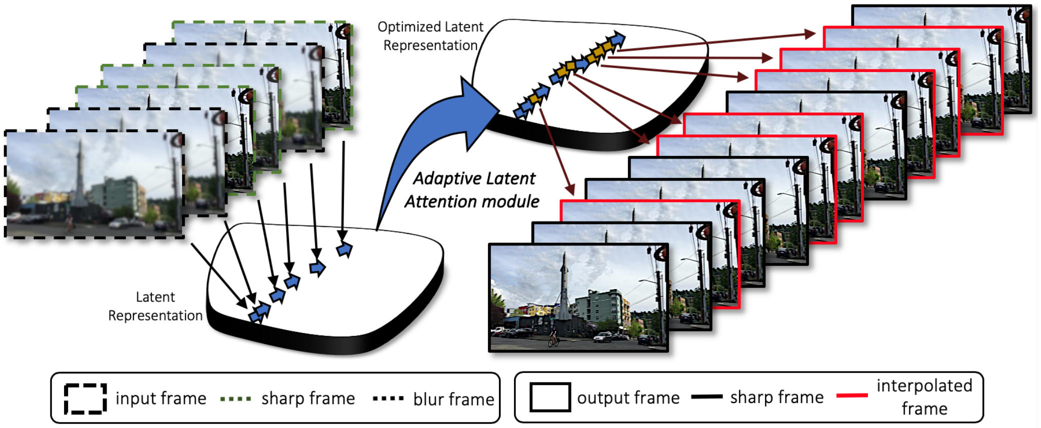


Figure 2.1: Conceptual Overview of ALANET. Given a poor-quality video consisting both blurry and sharp frames, the frames are projected on a latent space. These latent representations are modulated and interpolated using the proposed Adaptive Latent Attention module to generate optimized latent representations for deblurring and interpolation. These optimized representations are then used to generate a high frame-rate sharp video.

a low frame-rate blurry video. Recently, [92] 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 input frames are blurry, which is often unrealistic because the quality of a video usually varies non-uniformly over time.

In this work, 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 [87] 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 2.1 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.

Approach Overview. An overview of our approach is illustrated in Figure 2.2. 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.*

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.

2.2 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 2.1).

Video Deblurring. Inversion of motion blur is an ill-posed problem [82, 77]. Recent works have used deep learning based methods to solve this restoration problem either using a single frame [98, 105] or multiple frames [44, 53, 38, 98, 73]. [15] attempts to deblur a video by exploring similarity between the frames of the video and exploiting sharp patches of neighbouring frames. DeBlurNet [98] proposes to use consecutive frames stacked as input to generate a single clean central frame. ESVR [112] 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. [52] proposes an integrated model to jointly predict the defocus blur, optical flow and latent frames. [39] proposed a spatio-temporal recurrent neural network that enforces temporal consistency between neighbouring frames. [128] 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.

Video Interpolation. Many of the existing approaches [66, 130, 4, 5, 43, 64] 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 [65]. Some methods [74, 75] 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 [110], variations of attention mechanism have shown promising results in object recognition [3], image generation [120] and image super-resolution [124]. Residual channel attention mechanism for super-resolution is introduced in [124]. Authors in [115] 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 [115] and video interpolation [16]. In [16], 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. [44] 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 [92] 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*

Table 2.1: 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 [4]	✔	✗	✗	✗
Jin [44]	✔	✔	✗	✗
BIN [92]	✔	✔	✔	✗
ALANET (Ours)	✔	✔	✔	✔

deblur using information from neighbouring frames.

2.3 Problem Formulation

Given a low frame-rate poor quality video $\mathbf{V} = [\ \mathsf{V}_1,\ \mathsf{V}_2, \cdots,\ \mathsf{V}_L]$, with L frames, we aim to generate a high frame-rate sharp video $\mathbf{S} = [\mathsf{S}_1,\ \mathsf{S}_2, \cdots,\ \mathsf{S}_N]$ 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 $\mathsf{X}_{\mathbf{V}} = [\ \mathbf{x}_1,\ \mathbf{x}_2, \cdots,\ \mathbf{x}_L\ ]$ 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 2.4.2) 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

$$\mathsf{Z}_{\mathbf{S}} = [\mathbf{z}_1,\ \widehat{\mathbf{z}}_2,\ \mathbf{z}_3,\ \widehat{\mathbf{z}}_4, \cdots,\ \mathbf{z}_N] = \widetilde{\mathsf{Z}}_{\mathbf{S}} \bigcup \widehat{\mathsf{Z}}_{\mathbf{S}} \tag{2.1}$$

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.

2.4 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 2.2.

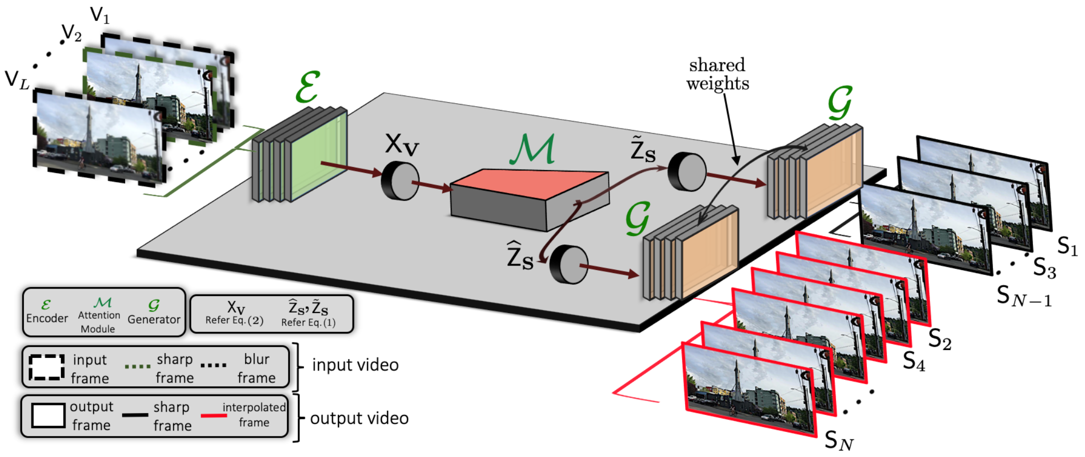


Figure 2.2: Architectural Overview of ALANET. Given a low frame-rate poor quality video $\mathbf{V} = [\ \mathsf{V}_1,\ \mathsf{V}_2, \cdots,\ \mathsf{V}_L]$, we extract latent representations $\mathsf{X}_{\mathbf{V}} = [\ \mathbf{x}_1,\ \mathbf{x}_2, \cdots,\ \mathbf{x}_L\ ]$ using encoder network $\mathcal{E}$. Adaptive Latent Attention module $\mathcal{M}$ utilizes combination of self-attention and cross-attention on $\mathsf{X}_{\mathbf{V}}$ to generate optimized representations for deblurring ($\widetilde{\mathsf{Z}}_{\mathbf{S}}$) and interpolation ($\widehat{\mathsf{Z}}_{\mathbf{S}}$). These optimized representations are used by the generative network $\mathcal{G}$ to synthesize deblurred frames $(\mathsf{S}_1, \mathsf{S}_3, \cdots, \mathsf{S}_{N-1})$ from $\widetilde{\mathsf{Z}}_{\mathbf{S}}$ and interpolated frames $(\mathsf{S}_2, \mathsf{S}_4, \cdots, \mathsf{S}_N)$ from $\widehat{\mathsf{Z}}_{\mathbf{S}}$, thereby generating a high frame-rate video $\mathbf{S} = [\mathsf{S}_1,\ \mathsf{S}_2, \cdots,\ \mathsf{S}_N]$.

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.

$$\begin{aligned} \mathcal{E}(\mathbf{V}) &= \mathcal{E}\Big([\ \mathsf{V}_1,\ \mathsf{V}_2, \cdots,\ \mathsf{V}_L]\Big) \\ &= [\ \mathbf{x}_1,\ \mathbf{x}_2, \cdots,\ \mathbf{x}_L] = \mathsf{X}_{\mathbf{V}} \end{aligned} \tag{2.2}$$

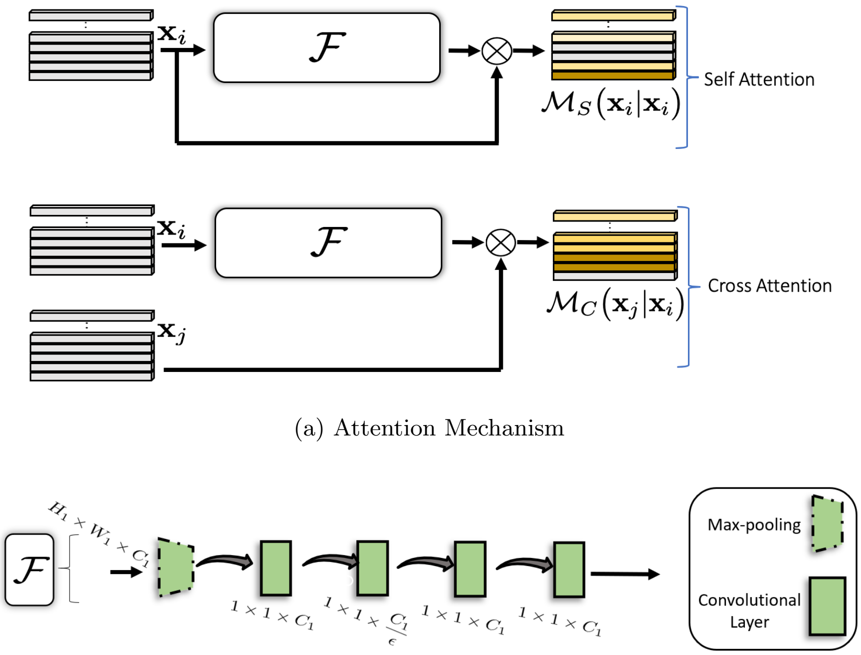
Here, $\mathbf{x}_i \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ is the latent representation corresponding to V_i. The representations generated by the encoder $\mathcal{E}$ are used by the Adaptive Latent Attention module $\mathcal{M}$ to generate optimized representations for deblurring and interpolation.

2.4.2 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.

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 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 [124], 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 2.3 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.



(a) Attention Mechanism

(b) Channel Attention Computation Network

Figure 2.3: 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.

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 (2.3).

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 (2.4).

$$\mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) = \mathbf{x}_i \otimes \mathcal{F}(\mathbf{x}_i) \tag{2.3}$$

$$\mathcal{M}_C\Big(\mathbf{x}_j|\mathbf{x}_i\Big) = \mathbf{x}_j \otimes \mathcal{F}(\mathbf{x}_i) \tag{2.4}$$

Note that, symbol $\otimes$ in (2.3) and (2.4) represents element-wise multiplication, input variables $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ are the encoded feature representations of frames and output of $\mathcal{M}_\mathbf{S}(\mathbf{x}_i|\mathbf{x}_i),\ \mathcal{M}_\mathbf{C}(\mathbf{x}_j|\mathbf{x}_i) \in \mathbb{R}^{H_1 \times W_1 \times C_1}$.

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 $\mathsf{Z}_\mathbf{V} = [\ \mathbf{z}_1,\ \widehat{\mathbf{z}}_2,\ \mathbf{z}_3,\ \widehat{\mathbf{z}}_4, \cdots,\ \mathbf{z}_N]$ for a high frame-rate video $\mathbf{S}$ is computed as follows:

$$\mathbf{z}_{2i} = \mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) + \sum_{j \in \mathbb{Q}} \mathcal{M}_C\Big(\mathbf{x}_j|\mathbf{x}_i\Big) \tag{2.5}$$

$$\begin{aligned}\widehat{\mathbf{z}}_{2i+1} = {} & \mathcal{M}_S\Big(\mathbf{x}_i|\mathbf{x}_i\Big) + \mathcal{M}_C\Big(\mathbf{x}_i|\mathbf{x}_{i+1}\Big) \\ & + \mathcal{M}_S\Big(\mathbf{x}_{i+1}|\mathbf{x}_{i+1}\Big) + \mathcal{M}_C\Big(\mathbf{x}_{i+1}|\mathbf{x}_i\Big)\end{aligned} \tag{2.6}$$

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 (2.5), 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 (2.6), 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.

2.4.3 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 $\mathbf{S} = \mathcal{G}([\ \mathbf{z}_1,\ \widehat{\mathbf{z}}_2,\ \mathbf{z}_3,\ \widehat{\mathbf{z}}_4, \cdots,\ \mathbf{z}_N])$ 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.

2.4.4 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 [43] 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 [87]. 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 [124] 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 [124]. Unlike self-attention for super-resolution in [124], we also employ cross-attention between consecutive features to learn interactions between these features for deblurring and interpolation.

2.5 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.

2.5.1 Datasets and Metrics

We evaluate the performance of our proposed approach using publicly available Adobe240 [98] dataset which has been used in many prior works and a dataset crawled from YouTube as in [92].

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 [43] 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 [92]. 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 [98] and crawled YouTube240 dataset, low frame-rate poor quality videos of 30fps are generated using process described in section 2.5.2. All the frames are resized to 640 × 352 for training and evaluation purposes.

2.5.2 Implementation Details

Our framework is implemented in PyTorch [80]. All the experiments are trained for 200 epochs with a batch size of 2. We use ADAM [54] 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 [44, 43, 98]. 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:

$$\mathsf{V}_i = \frac{1}{2\tau + 1} \sum_{k=i\beta-\tau}^{i\beta+\tau} \mathsf{S}_k \tag{2.7}$$

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 (2.7) on a sharp video to generate a poor quality video with low frame-rate.

Training and Testing Protocol. During training, random blurry frames are generated on-the-fly by averaging 11 frames as defined in (2.7). 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 [46] defined as follows.

$$\mathcal{L} = \mathcal{L}_r + \lambda \mathcal{L}_p \tag{2.8}$$

Here, $\mathcal{L}_r = \sum_i |\mathsf{G}_i - \mathsf{S}_i|_1$ 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 [46], and λ is a hyper-parameter. We use $\lambda = 0.2$ for all our experiments.

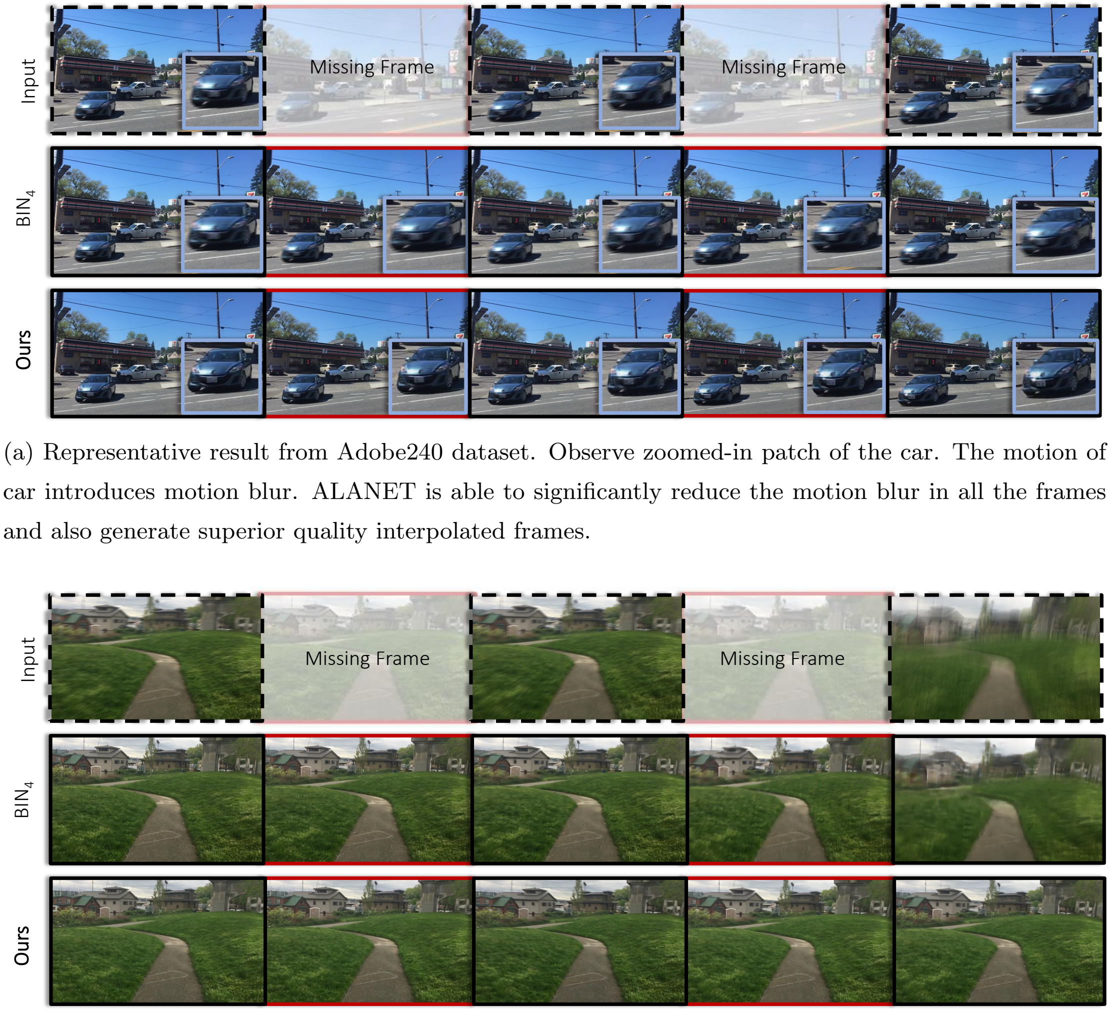
[1] For pixel reconstruction loss, we choose ℓ_1-loss instead of Mean-Squared Error (MSE) ℓ_2 loss as latter has inherent property of generating blurry output as shown in the literature [125].

2.5.3 Qualitative Results

Figure 2.4 shows some examples of high frame-rate videos generated using the proposed method and state-of-the-art BIN_4 [92] given a low frame-rate video (top row). From Figure 2.4a, 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 2.4b, 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 2.4b (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.

2.5.4 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 [43, 5, 4] given an input blurry video. We also compare **ALANET** with



(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 2.4: 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.

Table 2.2: 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 [92].

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† [43]	-	-	-	-	27.52	0.8593	30.84	0.9107	-	-	-	-
MEMC-Net† [5]	-	-	-	-	30.83	0.9128	34.91	0.9596	-	-	-	-
DAIN† [4]	-	-	-	-	31.03	0.9172	35.09	0.9615	-	-	-	-
Jin† [44]	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$_4$† [92]	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

two recent approaches where deblurring and interpolation is performed jointly [44, 92]. Quantitative result comparison with these baselines are shown in Table 2.2.

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 [92]. Our method achieves 32.98dB average PSNR in interpolation task as opposed 32.51dB reported by state-of-the-art method BIN_4 [92]. 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 [44]. We again highlight that our method does not know which frames are blurry or where to interpolate, unlike BIN_4 [92].

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 [92]. However, we do not have the same set of videos as in [92] as the list of videos is not publicly available. From Table 2.2, it can be 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.

2.5.5 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

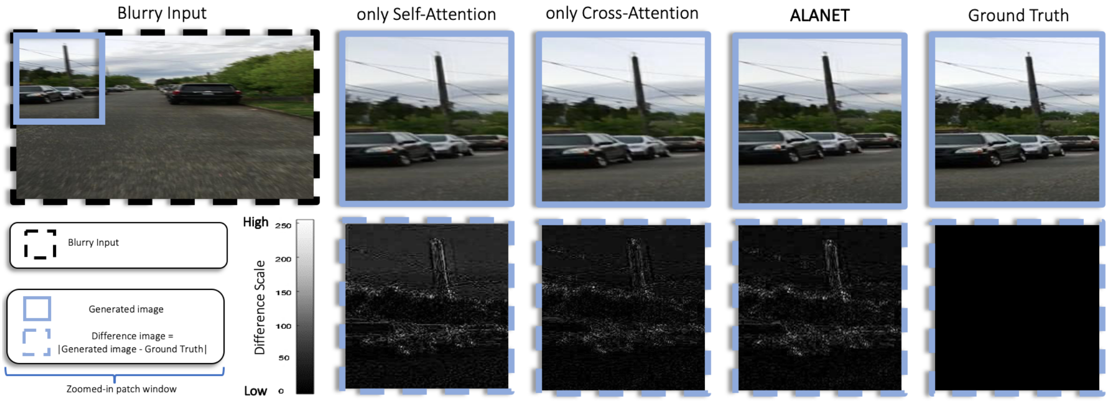


Figure 2.5: Ablation study on different attention modules. Frame generated using different attention mechanisms (top) and the residue image (bottom) computed by taking its difference with the ground-truth frame. Scale for the error range [0. 255] is given on the bottom left. Our proposed ALANET which combines self-attention and cross-attention produces superior results compared to using only one of the attention mechanisms. Results best viewed when zoomed-in.

and interpolation. We remove the cross-attention $\mathcal{M}_C$ terms from (2.5) and (2.6) 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 (2.5) and (2.6) for training the network.

Figure 2.5 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.

Table 2.3: 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

The quantitative results of impact of different attention mechanisms are shown in Table 2.3. 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 2.3, 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.

2.6 Additional Qualitative Results

We present a few videos generated by the proposed **ALANET** on Adobe240 and crawled YouTube240 dataset. Figure 2.6- 2.9 presents a few of the test videos in Adobe240 dataset [98]. An example video from crawled YouTube240 dataset is show in Figure. 2.10. It can be observed from Figure 2.6- 2.10 that the video generated by the proposed approach are clear and interpolated frame are of high perceptual quality. Not only blur caused by averaging of frame but **ALANET** is able to tackle motion blur caused by motion of the object as can be seen from Figure 2.7. These examples show efficacy of the Adaptive Latent Attention mechanism in synthesis of high-frame rate sharp video.

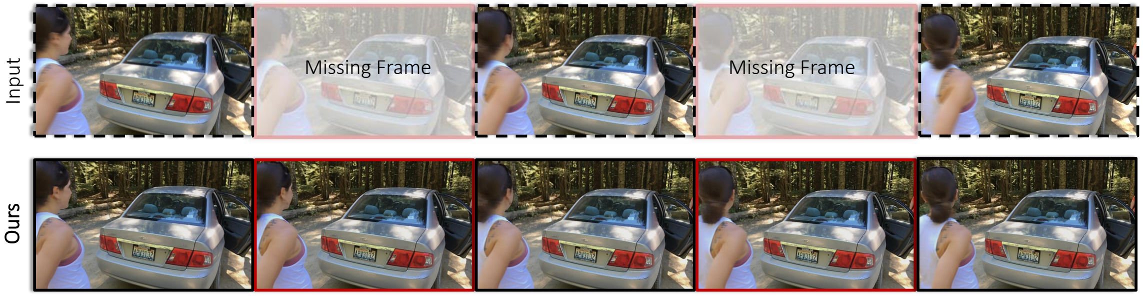


Figure 2.6: Representative video from Adobe240 dataset. The girl at the left hand side of the input frames is blurry. ALANET is able to deblur the blur portion of the frames and interpolate high quality frames

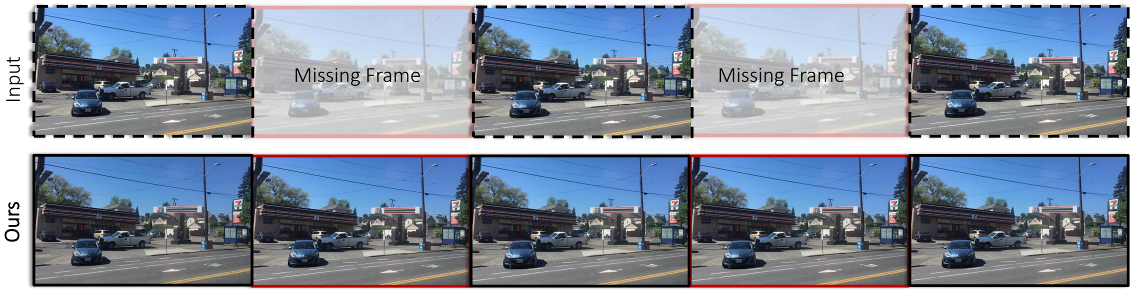


Figure 2.7: Representative video from Adobe240 dataset. The car at the bottom left has motion blur due to frame averaging as well as the motion of the car. Our proposed approach is able to enhance the frame by removing motion blur caused by frame averaging as well as object motion.

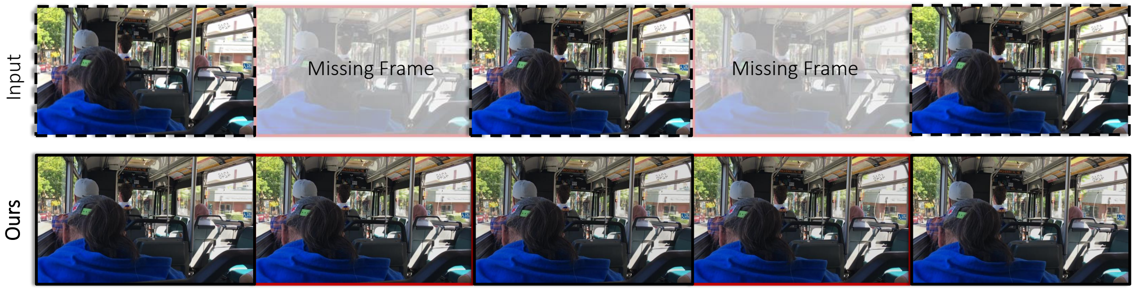


Figure 2.8: Representative video from Adobe240 dataset. Since the bus is moving there is blur towards the right window in the input frame. Our proposed approach is able to deblur specific parts of the video showing efficacy of Adaptive Latent Attention.

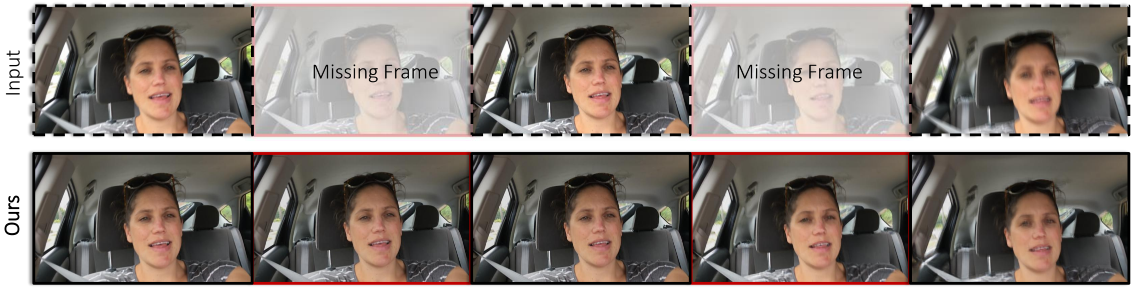


Figure 2.9: Representative video from Adobe240 dataset. In this video, the girl is holding the camera and there is lot of motion blur. Since we are leveraging information from neighbouring frames using cross-attention mechanism we are able to produce quality results even when very blurry input is given (observer last frame).

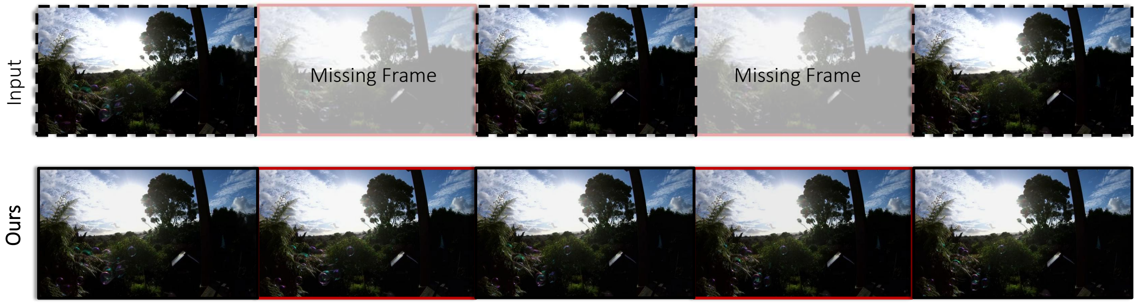


Figure 2.10: Representative video from crawled YouTube240 dataset. In this video, bubbles are moving objects. If observed carefully, it can be seen that our approach is able to generated sharp boundaries for each bubble while deblurring as well as interpolation.

2.7 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 a 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 the task of joint deblurring and interpolation over the state-of-the-art methods.

Chapter 3

Spatio-Temporal Video Super-Resolution

Most of the existing works in supervised spatio-temporal video super-resolution (STVSR) heavily rely on a large-scale external dataset consisting of paired low-resolution low-frame rate (LR-LFR) and high-resolution high-frame rate (HR-HFR) videos. Despite their remarkable performance, these methods make a prior assumption that the low-resolution video is obtained by down-scaling the high-resolution video using a known degradation kernel, which does not hold in practical settings. Another problem with these methods is that they cannot exploit instance-specific internal information of a video at testing time. Recently, deep internal learning approaches have gained attention due to their ability to utilize the instance-specific statistics of a video. However, these methods have a large inference time as they require thousands of gradient updates to learn the intrinsic structure of the data. In this work, we present **Ada**ptive **V**ideo **S**uper-**R**esolution (**Ada-VSR**) which leverages external as well as internal information through meta-transfer learning and internal learning, respectively. Specifically, meta-learning is employed to obtain adaptive parameters, using a large-scale external dataset, that can adapt quickly to the novel condition (degradation model) of the given test video during the internal learning task, thereby exploiting external and internal information of a video for super-resolution. The model trained using our approach can quickly adapt to a specific video condition with only a few gradient updates, which reduces the inference time significantly. Extensive experiments on standard datasets demonstrate that our method performs favorably against various state-of-the-art approaches.

3.1 Introduction

With the increasing popularity of high-performance higher resolution displays such as 4K Ultra HD (UHD), the demand for high-quality visual content is also increasing. However, professional video production and TV screen content are still at Full HD (1080p) resolution [40, 51, 50]. As rendering low-resolution content on higher resolution displays lowers perceptual quality, it calls for improving the resolution of the content to match that of the display. Enhancing the quality of video not only requires increasing the spatial resolution but also the temporal resolution for smooth rendering on high-performance displays. Therefore, it is critical to improve the spatial as well as the temporal resolution of videos to enhance the perceptual quality.

Most existing approaches have addressed the task of video spatial super-resolution (VSR) [47, 45, 119, 104, 10, 112] and temporal video super-resolution (TSR) [66, 130, 4,

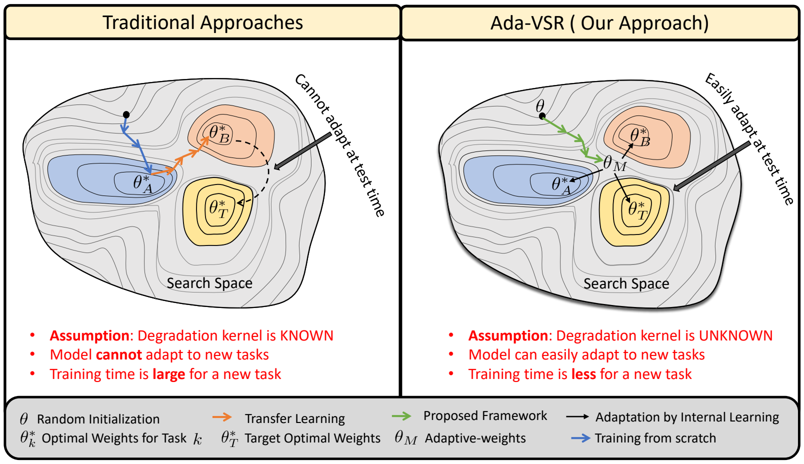

Traditional Approaches
Ada-VSR (Our Approach)
Cannot adapt at test time
θ^*_B
θ^*_A
θ^*_T
Search Space
Assumption: Degradation kernel is KNOWN
Model cannot adapt to new tasks
Training time is large for a new task
Easily adapt at test time
θ
θ_M
θ^*_B
θ^*_A
θ^*_T
Search Space
Assumption: Degradation kernel is UNKNOWN
Model can easily adapt to new tasks
Training time is less for a new task
θ Random Initialization
θ^*_k Optimal Weights for Task k
Transfer Learning
θ^*_T Target Optimal Weights
Proposed Framework
θ_M Adaptive-weights
Adaptation by Internal Learning
Training from scratch

Figure 3.1: Comparison of traditional approaches against Ada-VSR for the blind VSR task. Traditional supervised approaches train their model assuming the degradation kernel for task A is known (left; blue arrows). Transfer learning can be adopted to find optimal model parameters for Task B with a different degradation kernel(left; orange arrows). However, the model will not be able to generalize for the target task T when degradation is not known. On the other hand, our proposed approach tries to find weights that can easily adapt to the target task with only a few gradient updates via internal learning (right; green arrows. See sec. 3.3.2 for more details).

5, 43, 64], separately. A straightforward strategy to perform spatio-temporal video super-resolution (STVSR) is to cascade the VSR model and the TSR model to generate high-resolution high-frame rate (HR-HFR) video from low-resolution low-frame rate (LR-LFR) video. Nevertheless, this does not yield optimal results as it cannot fully utilize the available spatio-temporal information [118]. Recently, a few works [50, 51, 118, 117], studied the problem of joint spatio-temporal video super-resolution. Zooming Slow-Mo [117] proposed a one-stage STVSR framework using Deformable Convolutional LSTM. The authors in [118] utilize temporal profiles to exploit spatio-temporal information. FISR [51] proposes a multi-scale temporal loss for joint frame-interpolation and super-resolution. However, these approaches requires a large dataset of LR-HR pairs with the assumption that the down-sampling kernel to obtain LR frames from HR frames is known and fixed, which does not hold true in a real world setting (Figure 3.1, left). The problem of blind SR in images, where down-sampling kernel is unknown, is tackled either by estimating the down-sampling kernel [28, 7] or by exploiting the deep internal prior [94, 107] to learn the internal structure of the image. Consequently, these approaches achieve good performance at the expense of heavy computational time as it requires thousands of back-propagation gradient updates for each instance. Another shortcoming of such approaches is that they cannot take advantage of a pre-trained network learned using a large-scale external dataset [96].

Meta-learning has recently garnered much interest to tackle the aforementioned shortcomings. Meta-learning aims to adapt quickly and efficiently from a small set of data available at inference time. There are three common approaches to meta-learning: metric-based [95, 102, 111], model-based [88, 78, 70], and optimization-based [27, 25, 24]. Model-Agnostic Meta-Learning (MAML) [24] is a gradient-based method and has shown impressive performance by learning the optimal initial state of the model such that it can quickly adapt to a new task with a few gradient steps.

In this work, we introduce a novel framework **Ada**ptive **V**ideo **S**uper-**R**esolution (**Ada-VSR**) which aims to generate high resolution high-frame rate (HR-HFR) video from a low-resolution low-frame rate (LR-LFR) input. Inspired by meta-transfer learning, we utilize external knowledge as well as internal knowledge from videos for the task of joint spatio-temporal video super-resolution (STVSR). Our approach leverages knowledge from the external dataset and learns adaptive model parameters using meta-learning. As shown in Figure 3.1 (right), meta-learning is employed to learn initial model parameters that can quickly and efficiently adapt to the test video with unknown degradation. Specifically, we use different down-sampling kernels as different tasks for meta-learning in order to learn a model that can adapt to novel tasks easily. Finally, during meta-testing, internal learning is leveraged to learn video-instance specific knowledge with limited gradient steps. Since there are only a few gradient steps involved in the meta-testing for each video, our approach is significantly faster when compared to approaches that completely rely on internal learning and requiring thousands of gradient updates.

3.1.1 Approach Overview

An overview of our Adaptive Video Super-Resolution (**Ada-VSR**) training scheme is illustrated in Figure 3.2. Given a low-resolution low frame-rate input, our objective is to generate a high resolution high frame-rate video when the degradation kernel is unknown. Our **Ada-VSR** approach consists of two networks: the temporal super-resolution module (TSR) denoted by $\mathcal{F}_\theta$ and the spatial super-resolution module (SSR) as $\mathcal{S}_\phi$. We adopt a meta-learning framework to train both the networks jointly using an external dataset containing LR-LFR (obtained using dynamic task generator; refer Fig. 3.2) and HR-HFR video pairs. The objective of the meta-training is to learn model parameters that can be easily adapted to the test video. This meta-training is performed only once and the adaptive parameters are used to initialize the model in the next step. In a practical setting, we will only have access to the LR-LFR video. Hence, we exploit the internal structure within the test LR-LFR video using internal learning. During internal learning, we first downscale the LR-LFR further to obtain a super low-resolution low-frame rate video (SLR-LFR) video and try to reconstruct the LR-LFR video from SLR-LFR video. Note that we can obtain considerable results with only a few gradient steps during internal learning, as the adaptive parameters obtained by meta-training can quickly adapt to unseen tasks, which is unknown degradation in our case. Finally, we use the model trained with internal learning to infer HF-HFR video from LF-LFR video.

3.1.2 Contributions

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

- We propose a novel meta-learning framework **Ada-VSR** for the task of joint spatio-temporal super-resolution by leveraging external and internal learning.
- We employ a combination of various spatial and temporal down-sampling techniques during training to learn a model that can easily adapt to unknown degradation process.
- We significantly reduce the computational time by greatly reducing the gradient steps required during internal learning.

3.2 Related Work

Our work relates to research in spatial and temporal video super-resolution, internal learning and, meta-transfer learning. In this section, we discuss some methods closely related to our work. We provide a characteristic comparison of recent works in Table 3.1.

Image Super-Resolution. Deep learning approaches have shown remarkable performance on the task of image super-resolution [33, 20, 49, 55, 60, 94, 28, 121]. Recently, various Convolutional Neural Network (CNN) based approaches have been proposed for non-blind

Table 3.1: Categorization of prior works in video super-resolution. Different from the state-of-the-art approaches, we employ meta-learning to perform blind spatio-temporal video super-resolution.

Methods	Super-Resolution		Learning Method	
	Spatial	Temporal	Meta	Internal
DyaVSR [56]	✔	✗	✔	✔
Temporal Profiles [118]	✔	✔	✗	✗
Zooming Slow-Mo [117]	✔	✔	✗	✔
Ada-VSR (Ours)	✔	✔	✔	✔

image SR where the down-sampling kernel (e.g. bicubic), used to obtain low-resolution (LR) image, is known [33, 20, 49, 55, 60]. Despite the impressive performance of these methods, their efficacy deteriorates when the down-sampling kernel is different than the one used to train these models due to the domain gap. To overcome this issue, SRMD [121] incorporates multiple degradation kernels as input to their model along with the LR image. On the other hand, Zero-Shot Super-Resolution (ZSSR) [94] exploits the deep prior [107] to learn an image specific structure to obtain an SR image. Some approaches first try to estimate the degradation kernel and utilize the estimated kernel for image super-resolution. An iterative approach to correct inaccurate degradation kernels is introduced in [28]. Similar to ZSSR, the KernelGAN [7] utilizes the patch-recurrence property of a single image for super-resolution. However, these methods train the network from scratch for all the image instances, making them computationally heavy.

Video Spatial Super-Resolution. Earlier works in video super-resolution focused on developing effective priors on the HR frames to solve this problem [6, 11, 23, 91]. Motivated by the success of deep learning approaches in image super-resolution [94, 121, 28], several deep learning based methods have been proposed for video super-resolution [47, 45, 119]. A CNN based approach is proposed in [47], where the network is trained on both the spatial and temporal dimensions of videos to spatially enhance the frames. A SR draft-ensemble approach for fast video spatial super-resolution is proposed in [59]. In [104, 10], the authors incorporate optical flow estimation models to explicitly account for the motion between neighboring frames. However, accurate flow is difficult to obtain given occlusion and large motions. A computationally lighter flow estimation module (TOFlow) is proposed in [119] to account for motion information. DUF [45] overcomes this problem by implicit motion compensation using their proposed dynamic upsampling filter network. Pyramid, Cascading and Deformable convolution (PCD) alignment and the Temporal and Spatial Attention (TSA) modules are proposed in EDVR [112] to incorporate implicit motion compensation. However, these approaches assume the degradation kernel for down-sampling is known and/or require a large amount LR-HR pairs to train their models.

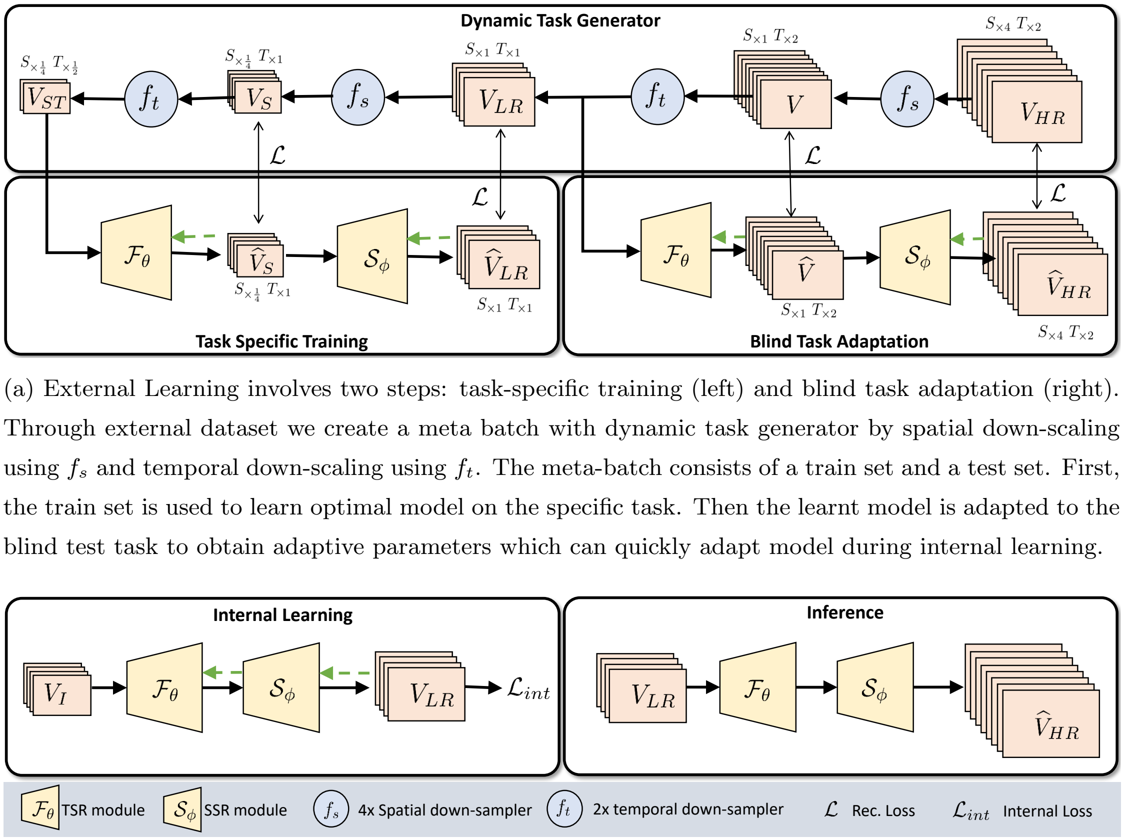
Video Temporal Super-Resolution. Video super-resolution can also be performed in the temporal dimension and is often termed as video interpolation. In temporal video super-resolution, the task is to generate a high-frame rate (HFR) video from a low-frame rate video (LFR). Existing approaches [66, 130, 4, 5, 43, 64] use optical flow estimation between input frames for temporal super-resolution. Thus, the quality of estimated optical flow governs the quality of frame interpolation. Deep learning approaches have demonstrated effectiveness in temporal super-resolution tasks. A straightforward application of

CNNs for intermediate frame synthesis is presented in [65]. Some methods [74, 76] apply CNNs to estimate space-varying and separable convolutional kernels for frame synthesis using neighbourhood pixels. [1] proposed a non-adversarial approach to generate videos by first learning optimized representation and then interpolating between the optimized latent representation of two frames to synthesize central frame. Joint video deblurring and interpolation to enhance and increase the frame-rate of a video is explored in [92, 31]. These approaches do not perform spatial super-resolution in their work and assume that the low-temporal resolution video is obtained by averaging 9 consecutive frames. Unlike these methods, we address the task of joint spatial and temporal super-resolution.

Meta-Learning. Recently, meta-learning algorithms have achieved impressive performance in various applications like few-shot learning [57, 41, 83, 101, 95], reinforcement learning [90, 24, 30, 72] and image super-resolution [96, 56]. Meta-learning aims to learn a model that can quickly and efficiently adapt to novel unseen tasks. There are three common approaches to meta-learning: metric-based [95, 102, 111], model-based [88, 78, 70], and optimization-based [27, 25, 24]. DynaVSR is proposed in [56], which utilizes meta-learning for spatial video super-resolution and has shown superior performance. Different from DynaVSR [56], we leverage meta-learning for the task of joint spatio-temporal video super-resolution.

3.3 Methodology

Given a low-resolution low frame-rate video our goal is to generate a high-resolution high frame-rate video in blind video super-resolution setting where the down-scaling kernel



(a) External Learning involves two steps: task-specific training (left) and blind task adaptation (right). Through external dataset we create a meta batch with dynamic task generator by spatial down-scaling using f_s and temporal down-scaling using f_t. The meta-batch consists of a train set and a test set. First, the train set is used to learn optimal model on the specific task. Then the learnt model is adapted to the blind test task to obtain adaptive parameters which can quickly adapt model during internal learning.

(b) Internal Learning and Inference. Left: We exploit the instance specific information during internal learning. Spatio-temporal down-scaled version ($\mathbf{V}_{ST}$) of LR-LFR video $\mathbf{V}_{LR}$ for reconstruction of $\mathbf{V}_{LR}$. Internal learning helps exploit the internal statistic of the input video. Right: HR-HFR video $\widehat{\mathbf{V}}_{HR}$ is obtained by passing $\mathbf{V}_{LR}$ video through the model trained via internal-learning.

Figure 3.2: Overview of Adaptive Video Super-Resolution (Ada-VSR) framework. Our framework consists of two modules External Learning and Internal Learning. (a) External learning leverages meta-training protocol and exploits the external dataset to learn parameters that can easily adapt to novel tasks. (b) Internal learning helps exploit internal structure of the given video and is used to generate a HF-HFR video from LR-LFR video.

is not known at the test time. Let the low-resolution low frame-rate video be denoted by $\mathbf{V}_{LR} = \begin{bmatrix} \mathsf{L}_1, \ \mathsf{L}_2, \cdots, \ \mathsf{L}_L \end{bmatrix}$, with M frames where $\mathsf{L}_t \in \mathbb{R}^{H \times W \times C}$ and t denotes the time step. We aim to generate a high-resolution high frame-rate video $\mathbf{V}_{HR} = \begin{bmatrix} \mathsf{S}_1, \ \mathsf{S}_2, \cdots, \ \mathsf{S}_N \end{bmatrix}$ with N frames, where $\mathsf{S}_t \in \mathbb{R}^{\mathbf{a}H \times \mathbf{a}W \times C}$ and $N = \mathbf{b}M$. Our objective is to increase the spatial resolution of the given input video $\mathbf{V}_{LR}$ by a factor of $\mathbf{a}$ and the temporal resolution by a factor of $\mathbf{b}$.

In this section, we describe the proposed **Ada-VSR** framework in detail. Our framework consists of two modules: the Temporal Super-Resolution (TSR) module $\mathcal{F}_\theta$ and the Spatial Super-Resolution (SSR) module $\mathcal{S}_\phi$. We use the TSR module to interpolate frames and increase the frame rate by a factor of 2. The SSR network uses the output of the temporal super-resolution module and increases the spatial resolution by a scaling factor of 4. The overall training scheme of our approach **Ada-VSR** is shown in Figure 3.2. Our framework consists of two training paradigms: external learning and internal learning.

3.3.1 External Learning

The external learning protocol leverages a large-scale external dataset to perform knowledge transfer and domain generalization using pre-training and meta-transfer learning respectively.

Large-Scale Training. In large-scale pre-training, we utilize a high-quality external dataset ($\mathcal{D}_{HR}$) to provide a warm start for meta-transfer learning. Since super-resolution tasks with different down-scaling kernels share similar parameter space, large-scale training helps to estimate the natural prior of high-resolution high-frame rate videos. The large-scale pre-training is also effective to stabilize the training of the MAML [24] algorithm.

For the SSR module, we apply bi-cubic spatial degradation to HR-HFR video $\mathbf{V}_{HR}$ to obtain low-resolution high-frame rate video LR-HFR $\widetilde{\mathbf{V}}_{LR}$. The videos $\mathbf{V}_{HR}$ and $\widetilde{\mathbf{V}}_{LR}$ form a synthetic dataset $\mathcal{D}_s$. We train the network $\mathcal{S}_\phi$ to learn spatial super-resolution task by minimizing the ℓ_1 pixel reconstruction loss between all the frames of the generated HR-HFR video $(\widehat{\mathbf{V}}_{HR})$ and corresponding ground truth HR-HFR video $\mathbf{V}_{HR}$. The objective function for large-scale training of the network $\mathcal{S}_\phi$ is defined as:

$$\mathcal{L}^{\mathcal{D}_s} = \sum_{(\widetilde{\mathbf{V}}_{LR}, \mathbf{V}_{HR}) \sim \mathcal{D}_s} \left\| \mathcal{S}_\phi(\widetilde{\mathbf{V}}_{LR}) - \mathbf{V}_{HR} \right\|_1 \tag{3.1}$$

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 [125].

The TSR module should be able to increase the frame-rate of a low-frame-rate (LFR) video. We can interpolate the frames to increase the frame-rate by factor of 2 and learn a residual by minimizing the reconstruction loss between the generated high-frame-rate $\widehat{\mathbf{V}}_{HR}$ and the ground truth video $\mathbf{V}_{LR}$. However, it may not be able to capture the temporal dynamics efficiently [118, 8]. Recently some works have addressed this by taking temporal profile or across dimension patches to leverage the patch recurrence in temporal dimension to train the network efficiently [118, 8]. We adopt the same strategy to train the TSR module $\mathcal{F}_\theta$. We define a temporal profile generator function f_r that takes a video input, performs the bi-cubic interpolation in temporal dimension and returns the temporal profile. To generate a dataset to train the TSR module we select alternate frames of the high-frame-rate (HFR) video $\mathbf{V}_{HR}$ to generate a LFR video $\overline{\mathbf{V}}_{LR}$. Then we apply the temporal profile generator function (f_r) to get the temporal profile $\mathbf{V}'_{LR}$ corresponding to the input $\overline{\mathbf{V}}_{LR}$ such that $\mathbf{V}'_{LR} = f_r(\overline{\mathbf{V}}_{LR})$, where $\mathbf{V}'_{HR}$ is the HFR temporal profile of the LFR input $\overline{\mathbf{V}}_{LR}$. We denote the paired data $(\overline{\mathbf{V}}_{LR}, \mathbf{V}_{HR})$ as $\mathcal{D}_t$. The loss function to update the TSR module $\mathcal{S}_\phi$ is given by the equation below.

$$\mathcal{L}^{\mathcal{D}_t} = \sum_{(\overline{\mathbf{V}}_{LR}, \mathbf{V}_{HR}) \sim \mathcal{D}_t} \left\| \mathcal{F}_\theta(\overline{\mathbf{V}}_{LR}) - \mathbf{V}_{HR} \right\|_1 \tag{3.2}$$

Dynamic Task Generator The Dynamic Task Generator (DTG; see fig. 3.2a) generates tasks for meta-training on-the-fly using diverse degradation settings. In our approach, the task distribution $p(\mathcal{T})$ is the combination of the spatial down-scaling kernel and temporal sub-sampling method. For spatial down-sampling by a factor of 4 we randomly apply the anisotropic Gaussian kernels using the function f_s. Temporal sub-sampling is performed with the function f_t by either selecting alternate frames or by averaging a window of size 3 to obtain a low-frame rate video.

Meta-Transfer Learning. We seek to find a set of transferable initial parameters where a few-gradient steps can adapt the model to the current video and achieve to large performance gain. Motivated by MAML [24] and [96], we employ meta-transfer learning strategy for spatio-temporal video super-resolution (STVSR) to learn adaptive weights. Unlike MAML, we use the external dataset for meta-training and leverage internal learning for meta-test step. Training with external dataset helps the meta-leaner to focus more on the down-scaling kernel-agnostic property, whereas internal learning helps to exploit the instance specific internal statistics.

Lines 10-19 in Algorithm 1 presents the meta-transfer learning optimization protocol. In our approach, we want to learn a generalized set of TSR parameters θ and SSR parameters ϕ such that the parameters can adapt to the test video quickly and efficiently in a blind video super-resolution setting. The meta-transfer learning achieves this using two steps: task-specific training and blind task adaptation. To learn generalized set of adaptive parameters, we first sample a task batch $\mathcal{D}_{meta}$ for the task $\mathcal{T}_i$ using Dynamic Task Generator. We then divide $\mathcal{D}_{meta}$ into two subsets: $\mathcal{D}_{tr}$ for task-specific training and $\mathcal{D}_{te}$ for blind task adaption.

Task-Specific Training. It is the inner loop of the MAML meta-learning algorithm where the meta-learner tries to learn a task-specific optimal parameters in one or more gradient descent steps. The inner loop is represented by Lines 12-16 in Algorithm 1. Given an external dataset $\mathcal{D}_{HR}$, we obtain a meta-task train batch $\mathcal{D}_{tr} = (\mathbf{V}_{LR}, \mathbf{V}_S, \mathbf{V}_{ST})$ for $\mathcal{T}_i \in p(\mathcal{T})$, where $\mathbf{V}_{LR}$ is LR-LFR video, $\mathbf{V}_S$ is 4x spatially down-scaled version $\mathbf{V}_{LR}$ and $\mathbf{V}_{ST}$ is 2x temporally down-scaled version of $\mathbf{V}_S$. We train the TSR model ($\mathcal{F}_\theta$) to generate a video $\widehat{\mathbf{V}}_S$ with temporal resolution twice to that of input video ($\mathbf{V}_{ST}$). The SSR model ($\mathcal{F}_\phi$) takes the output of TSR model ($\widehat{\mathbf{V}}_S$) and reconstruct the LR-LFR video $\widehat{\mathbf{V}}_{LR}$. The output of both the models are given by,

$$\widehat{\mathbf{V}}_S = \mathcal{F}_\theta(\mathbf{V}_{ST}) \qquad\qquad \widehat{\mathbf{V}}_{LR} = \mathcal{S}_\phi(\widehat{\mathbf{V}}_S) \tag{3.3}$$

We optimize both the network for n_i iteration to increase the resolution of a video for a task defined by random spatial and temporal down-scaling kernels. The loss function for the task-specific training is computed as follows:

$$\mathcal{L}^{tr}_{\mathcal{T}_i} = \sum_{\mathcal{D}_{tr}} \sum_{n_i} \left(\mathcal{L}\left(\widehat{\mathbf{V}}_S, \mathbf{V}_S\right) + \mathcal{L}\left(\widehat{\mathbf{V}}_{LR}, \mathbf{V}_{LR}\right) \right) \tag{3.4}$$

where $\mathcal{L}$ is reconstruction loss, n_i is number of inner loop iterations for task-specific training.

For one gradient update, new adapted parameters θ_i and ϕ_i are then obtained as

$$\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi) \tag{3.5}$$

$$\phi i = \phi - \alpha \nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi) \tag{3.6}$$

where α is the task-level learning rate.

Blind Task Adaptation. The blind task-adaptation is the outer loop of meta-learning which adapts the model parameters to the novel task. Here the test batch $\mathcal{D}_{te}$ is sampled from $\mathcal{D}_{meta}$ such that $\mathcal{D}_{te} = (\mathbf{V}_{HR}, \mathbf{V}, \mathbf{V}_{LR})$ for $\mathcal{T}_j \in p(\mathcal{T})$ where $\mathcal{T}_i \neq \mathcal{T}_j$, $\mathbf{V}_{HR}$ is HR-LFR video, $\mathbf{V}$ is 4x spatially down-scaled version of $\mathbf{V}_{HR}$ and $\mathbf{V}_{LR}$ is 2x temporally down-scaled version of $\mathbf{V}$. In order to adapt the models to new task $\mathcal{T}_j$, the model parameters θ and ϕ are optimized to achieve minimal test error on $\mathcal{D}_{meta}$ with respect to θ_i and ϕ_i. The meta-objective for blind task-adaptation is

$$\begin{aligned} &\arg\min_{\theta,\phi} \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \qquad (3.7) \\ =&\arg\min_{\theta,\phi} \sum_{\mathcal{T}_j} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta - \alpha \nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i}, \quad \phi - \alpha \nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}) \end{aligned}$$

Blind task adaptation using equation (3.8) learns the knowledge across tasks $\mathcal{T}_i$ and $\mathcal{T}_j$. The parameter update rule for for the above optimization can be expressed as:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \tag{3.8}$$

$$\phi \leftarrow \phi - \beta \nabla_\phi \sum_{\mathcal{T}_j \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_j}(\theta_i, \phi_i) \tag{3.9}$$

where β is the learning rate for blind task adaptation step.

Algorithm 1: Ada-VSR External Training

```
Input  : HF-HFR dataset $\mathcal{D}_{HR}$ and task distribution $p(\mathcal{T})$
Input  : $\alpha, \beta$: task-specific and adaptation learning rate
Output : Ada-VSR model parameters $\theta$ and $\phi$
/* Large-Scale training                                                    */
Randomly initialize $\theta$, $\phi$
Generate $\mathcal{D}_s$ using bi-cubic down-sampling kernel on $\mathcal{D}_{HR}$
while not done do
    /* Train SSR module                                                    */
    Sample LR-HR batch from $\mathcal{D}_s$
    Compute $\mathcal{L}^{\mathcal{D}_s}$ by Eq. (3.1)
    Update $\phi$ with respect to $\mathcal{L}^{\mathcal{D}_s}$
    /* Train TSR module                                                    */
    Sample LFR-HFR batch from $\mathcal{D}_t$
    Compute $\mathcal{L}^{\mathcal{D}_t}$ by Eq. (3.2)
    Update $\theta$ with respect to $\mathcal{L}^{\mathcal{D}_t}$
/* Meta-Transfer Learning                                                  */
while not done do
    Sample task batch $\mathcal{D}_{meta}$ for the task $\mathcal{T}_i \sim p(\mathcal{T})$
    /* Task-Specific Training (inner loop)                                 */
    for all $\mathcal{T}_i$ do
        Compute meta-training loss ($\mathcal{D}_{tr}$): $\mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi)$
        Adapt parameters with gradient descent:
        $\theta_i = \theta - \alpha \nabla_\theta \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi), \quad \phi_i = \phi - \beta \nabla_\phi \mathcal{L}^{tr}_{\mathcal{T}_i}(\theta, \phi)$
    /* Blind Task Adaptation (outer loop)                                  */
    Update $\theta$ and $\phi$ with respect to average test loss ($\mathcal{D}_{te}$):
    $\theta \leftarrow \theta - \alpha \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_i}(\theta_i, \phi_i)$
    $\phi \leftarrow \phi - \beta \nabla_\phi \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}^{te}_{\mathcal{T}_i}(\theta_i, \phi_i)$
```

```
Algorithm 2: Ada-VSR Internal Learning
Input  : LR-LFR test video $\mathbf{V}_{LR}$, meta-transfer trained model parameter
         $\theta, \phi$, number of gradient updates $n$ and learning rate $\gamma$
Output : High-resolution high-frame rate video $\widehat{\mathbf{V}}_{HR}$
1 Generate down-sampled video $\mathbf{V}_I$ by down-sampling $\mathbf{V}_{LR}$ with corresponding
   blur kernel.
  /* Internal Learning                                        */
2 for $n$ steps do
3     Evaluate loss $\mathcal{L}_{int}(\theta, \phi)$ using (3.10)
4     Update $\theta \leftarrow \theta - \gamma \nabla_{\theta} \mathcal{L}_{int}(\theta, \phi)$
5     Update $\phi \leftarrow \phi - \gamma \nabla_{\phi} \mathcal{L}_{int}(\theta, \phi)$
  /* Inference to generate HR-HFR                             */
6 return $\widehat{\mathbf{V}}_{HR} = \mathcal{S}_{\phi}\big(\mathcal{F}_{\theta}(\mathbf{V}_{LR})\big)$
```

3.3.2 Internal Learning and Inference

Algorithm 2 presents the internal learning and inference steps of our proposed approach. Given a LR-LFR video, we spatially down-sample it with corresponding down-sampling kernel by adopting the kernel estimation algorithms in [69, 79] for blind scenario and select alternate frames from the LR-LFR video to generate $\mathbf{V}_I$ and perform a few gradient updates with respect to the model parameter using a single pair of $\mathbf{V}_I$ as input and a given LR-LFR video $\mathbf{V}_{LR}$ as ground truth (Algorithm 2 Line 2-5). The aim here is to learn the internal statistics of the given video which can be utilized while generating HR-HFR video during inference. The objective function for internal learning is given:

$$\mathcal{L}_{int} = \left\| \mathcal{S}_{\phi}\big(\mathcal{F}_{\theta}(\mathbf{V}_I)\big) - \mathbf{V}_{LR} \right\|_1 \tag{3.10}$$

Then, we use the model trained with internal learning for inference. We feed the given LR-LFR input video $\mathbf{V}_{LR}$ to the model to generate a HR-HFR video $\widehat{\mathbf{V}}_{HR}$ as shown in Algorithm 2 Line 6.

3.4 Experiments

In this section, we first introduce the benchmark datasets and evaluation metrics. The qualitative and quantitative experiments are shown to demonstrate the effectiveness of our proposed approach in generating high-resolution high frame-rate videos.

3.4.1 Datasets and Metrics

We evaluate the performance of our approach using publicly available Vimeo-90K [119] and Vid4 [61] datasets which have been used in many prior spatial video super-resolution and temporal super-resolution works.

Vimeo-90K Dataset. The Vimeo-90K [119] dataset contains 91,707 short video clips, each containing 7 frames. The spatial resolution of each frame is (448×256). We use Vimeo-90K only for pre-training and meta-training, using the training split of 64,612 clips and use to the test set to compare against state-of-the-art approaches.

Vid4 Dataset. The Vid4 dataset [61],contains four video sequences: city, walk, calendar, and foliage. All the videos in Vid4 dataset contain at least 30 frames each, of spatial resolution 720×480. We evaluate a model trained on Vimeo-90K dataset on the Vid4 dataset and report the performance.

Metrics. For quantitative evaluation, we compare three metrics that evaluate different aspects of output image quality: Peak Signal-to-Noise Ratio (PSNR) [37], Structural Similarity Index Measure (SSIM) [113] and Naturalness Image Quality Evaluator (NIQE) [71].

3.4.2 Implementation Details

Our framework is implemented in PyTorch [80]. All the experiments are trained with a batch size of 32. We employ ADAM optimizer as the meta-optimizer in the meta-transfer learning step. The task-specific learning rate α is set to 0.01 and the adaptation learning rate β is set to 0.0001 for all our training experiments. The number of iterations in the task-specific training n_i is set to 10. We extracted training patches with a size of 64×64 for large-scale training. We utilize the Vimeo-90K dataset train split as the external dataset for large-scale training and meta-transfer learning. For internal learning the learning rate γ is set to 0.0001.

3.4.3 Qualitative Results.

Figure 3.3 compares a high-resolution high frame-rate videos generated using the proposed **Ada-VSR** approach with other state-of-the-art methods given a low-resolution low frame-rate video (left column). The low-resolution low-frame-rate input video is obtained by applying a non-bicubic degradation to the alternate frame of a high-resolution high-frame rate video. The skipped frames are faded in the Figure 3.3. It can be seen that the temporal profile based approach [118] does not perform well. It is due the fact that the model is trained with the assumption that there is a bi-cubic relationship between the LR-LFR and HR-HFR videos. This assumption is violated when we use an input video which was obtained by a non-bicubic down-sampling. Zooming Slow-Mo [117] produces slightly better video compared to the temporal profile approach as it is able to exploit the internal structure within the video. However, it is not able to exploit the external knowledge and

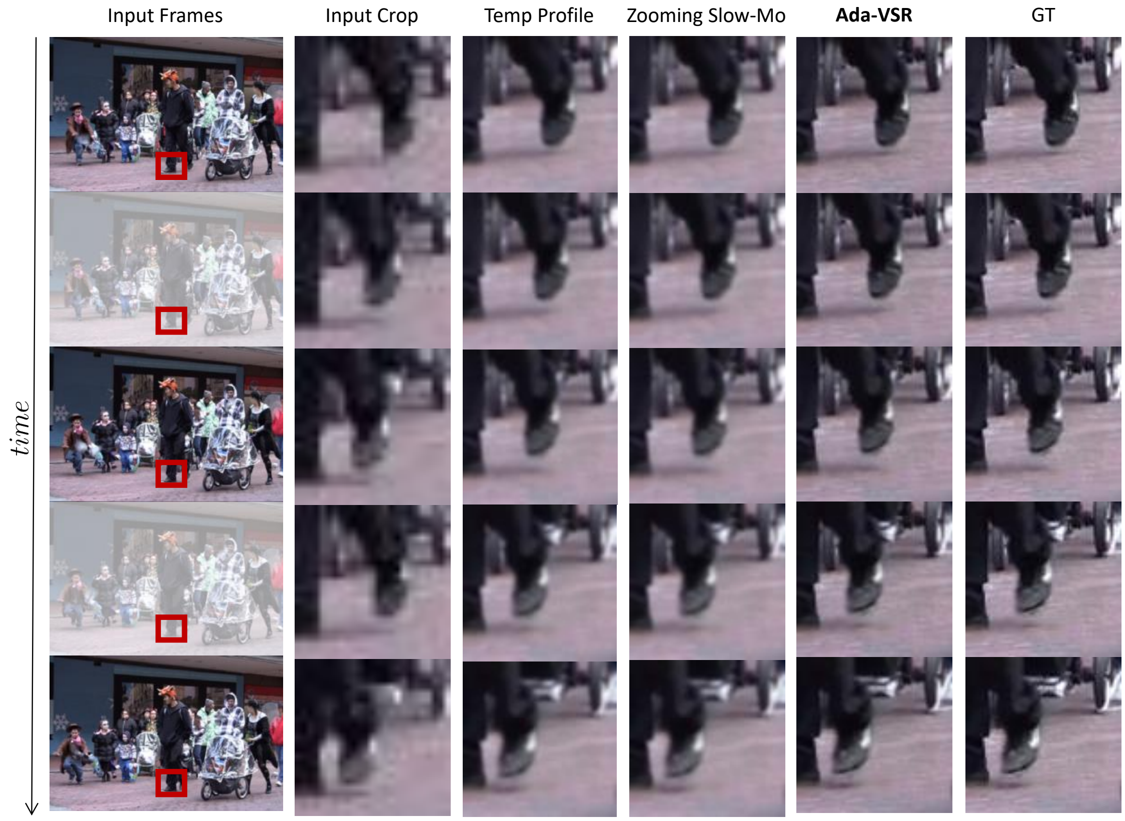


Figure 3.3: Comparison of qualitative results with the state-of-the-art methods. First column consists of the low-resolution low-frame-rate video input with non-bi-cubic degradation and the missing frames are faded. Second column and last column are the input and ground-truth crop of the input frame region marked in red. As opposed to temporal profile approach, Zooming Slow-Mo and Ada-VSR performs better as they can exploit internal structure of the input. Ada-VSR produces visually appealing results than Zooming Slow-Mo as the weights can easily adapt to novel tasks. Zoom-in for better visualization.

the output is still blurry. Our proposed approach **Ada-VSR** produces higher-quality and more visually appealing output as compared to both of these approach. The performance of our approach can be attributed to the adaptive parameters learned on external dataset with meta-learning. These parameters provide good initial parameters for internal training to learn instance specific characteristics.

Table 3.2: Quantitative results comparison on Vimeo-90K [119] and Vid4 [61] datasets. Our proposed approach is very competitive against various state-of-the-art approaches. Best scores are shown in bold and the second best are underlined. The state-of-the-art results are reported from [118].

Method		***Vimeo-90K Slow***			***Vimeo-90K Medium***			***Vimeo-90K Fast***			***Vid4***		
TSR	**SSR**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
SepConv [76]	IMDN [36]	31.75	0.88	7.68	33.13	0.90	7.78	34.31	0.92	8.55	24.87	0.72	6.34
SepConv [76]	SAN [19]	32.12	0.90	7.10	33.59	0.91	7.46	34.97	0.92	8.48	24.93	0.72	5.89
SepConv [76]	EDVR [112]	32.97	0.91	7.00	34.25	0.92	7.40	35.51	0.92	8.48	25.93	0.78	5.70
DAIN [4]	IMDN [36]	31.84	0.89	7.13	33.39	0.91	7.58	34.74	0.92	8.43	24.93	0.72	6.18
DAIN [4]	SAN [19]	32.26	0.90	7.05	33.82	0.92	7.45	35.27	0.92	8.48	25.14	0.73	5.78
DAIN [4]	EDVR [112]	33.21	0.91	7.06	34.73	0.93	7.39	35.71	0.93	8.47	26.12	0.79	5.62
SSR	**TSR**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
IMDN [36]	SepConv [76]	32.01	0.89	7.67	33.22	0.90	7.65	34.50	0.92	8.54	24.88	0.72	6.33
IMDN [36]	DAIN [4]	32.27	0.89	6.99	33.73	0.92	7.17	35.15	0.92	8.41	24.99	0.72	6.2
SAN [19]	SepConv [76]	32.32	0.90	6.99	33.73	0.92	7.32	35.33	0.92	8.42	25.01	0.73	5.87
SAN [19]	DAIN [4]	32.56	0.91	6.90	34.12	0.93	7.43	35.47	0.92	8.39	25.26	0.75	6.16
DynaVSR [56]	DAIN [4]	-	-	-	-	-	-	-	-	-	26.54	0.81	5.65
End-to-end Framework		**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**	**PSNR ↑**	**SSIM ↑**	**NIQE ↓**
Zooming Slow-Mo [117]		33.29	0.91	6.94	35.24	0.93	7.35	36.43	0.93	8.41	26.30	0.80	5.62
Temporal Profile [118]		**33.40**	**0.92**	6.17	35.55	0.94	6.37	36.29	0.93	7.13	26.50	0.82	5.48
Ada-VSR (Ours)		33.36	**0.92**	**6.12**	**35.91**	**0.95**	**6.33**	**36.52**	**0.95**	**6.99**	**26.98**	**0.84**	**5.40**

3.4.4 Quantitative Results

Our proposed method performs joint spatio-temporal video super-resolution. We compare **Ada-VSR** against representative STVSR approaches. We first compare our work with the two-stage solutions by cascading a video frame interpolation as temporal super-resolution module (TSR) and a video super-resolution as spatial super-resolution module (SSR). For temporal super-resolution module (TSR) we select SepConv [76] and DAIN [4] model, while SAN [19], IMDN [36], DynaVSR [56], and EDVR [112] are selected for spatial video super-resolution module (SSR). We also compare our work with recently proposed one-stage STVSR Zooming Slow-Mo [117] and Temporal profile based approach [118] where spatio-temporal super-resolution is performed jointly.

Table 3.2 presents the quantitative comparison on the test-set of Vimeo-90K [119] dataset and Vid4 [61] dataset. It is evident that our approach outperforms all the two-stage approaches by a significant margin against all the three metrics. When compared with the state-of-the-art one-stage approaches, temporal profile based [118] and Zooming Slow-Mo [117], our proposed **Ada-VSR** achieves superior performance on both the dataset except on Vimeo-90K Slow dataset in terms of PSNR where our performance is only 0.04db less than the temporal profile based approach.

In Table 3.3, we compare the average inference time of different state-of-the-art approaches. As our method learns adaptive weights that can easily adapt to novel tasks, only a few gradient updates during internal learning step are required to achieve visually compelling results. It can be observed from Table 3.3 that we outperform the Zooming Slow-Mo [117] by a margin of 0.66dB and the temporal profile approach by a significant margin of +1.18dB. As mentioned earlier, the temporal profile approach assumes that the degradation is bi-cubic hence it cannot generalize well on videos degraded using different blur kernels. It should also be noted that our approach is at about twice as fast as the temporal profile approach and at least 3 times faster when compared with Zooming Slow-Mo. Thus, significantly reducing the inference time during for new test videos with unknown degradation. Values for Zooming Slow-Mo [117] and temporal-profile are from [118].

Table 3.3: Average inference time (sec per frames) comparison of Ada-VSR with recent approaches for blind spatio-temporal video super-resolution.

Method	***Vid4 [61]***	
	PSNR↑	**Avg. time↓**
Zooming Slow-Mo [117]	26.30	0.1995
DynaVSR [56] + DAIN [4]	<u>26.54</u>	0.8940
Temporal profile [118]	25.78	<u>0.1328</u>
Ada-VSR (Ours)	**26.96**	**0.0680**

Table 3.4: Impact of large-scale training of TSR and SSR modules in Ada-VSR on the target performance.

External-Training		***Vid4 [61]***		
Spatial	**Temporal**	**PSNR ↑**	**SSIM↑**	**NIQE↓**
✔	✗	25.98	0.80	5.77
✗	✔	26.27	0.81	5.59
✔	✔	**26.98**	**0.84**	**5.40**

3.4.5 Ablation Study

. We investigate the contribution of large-scale training of the TSR module ($\mathcal{F}_\theta$) and the SSR module ($\mathcal{S}_\phi$). First, for the task-specific training in meta-learning, only the SSR module is initialized using the pre-trained weights obtained in the large-scale training step and the TSR module is initialized randomly. In second case, only the TSR module is initialized with the weights obtained from the large-scale training and the SSR module is randomly initialized. The quantitative results of impact of large-scale training using external data on Vid4 dataset are shown in Table 3.4. We can observe that the performance of **Ada-SVR** drops if large-scale pre-training is performed on only one of the SSR or TSR module. This is expected since the meta-learning algorithm MAML [24] has shown to be unstable when training without a warm model initialization. It is interesting to note that the model with large-scale training of only TSR module outperforms the one initialized with only the SSR module. We believe this could be due to the rich information available in temporal profiles used for large-scale training of the TSR module which provide stable initial parameters for task-specific training and blind-task adaptation during meta-training.

3.5 Conclusions

We present an Adaptive Video Super Resolution framework (**Ada-VSR**) for generating high resolution high frame-rate videos from low resolution low frame-rate input videos. We leverage external as well as internal learning to achieve spatio-temporal super-resolution. Specifically, external learning employ meta-learning to learn adaptive network parameters that can easily adapt unknown degradation and internal learning, on the other hand, helps to capture the underlying statistics of down-sampling and degradation specific to the input video by exploiting the internal structure, thereby making our approach more suited for practical, real-world data enhancement tasks. The proposed approach is able to achieve superior enhancement while adapting to unknown degradation models as shown inn our experiments. Experiments on standard datasets show not only the quantitative and qualitative efficacy of our proposed model in joint spatio-temporal video super-resolution, but also the improvement in computational time over various state-of-the-art methods.

Chapter 4

Joint Video Rolling Shutter Correction and Super-Resolution

With the prevalence of CMOS cameras in many computer vision applications, there is increase in appearance of rolling shutter (RS) artifacts in captured videos. However, existing video super-resolution algorithms assume that the motion is globally consistent in each video frame. and no rolling shutter effect is present. The problem of video super-resolution for video captured using RS cameras is challenging as the model needs to learn the row-wise local pixel displacements and the global structure of the objects for RS correction and super-resolution, respectively. In this work, we present Patch Attention Network (**PatchNet**) to address the problem of joint rolling shutter correction and super-resolution (RS-SR). Our hypothesis is that due to patch-recurrence property across different resolution scales, some neighbouring patches contain significant knowledge to super-resolve the input patch with finer details. Specifically, the Patch Attention Network leverages bi-directional motion information in feature space to extract relevant information from neighbouring patches using attention mechanism, and deformable fields using deformable convolution layers to extract local pixel-level information. Experiments on real as well as synthetic datasets demonstrate that our model performs favourably against various state-of-the-art approaches.

4.1 Introduction

CMOS (Complementary Metal–Oxide–Semiconductor) camera sensors are predominantly used in mobile devices largely due to their low cost, reduced power consumption and compact light weight design [42]. Motion blur and rolling shutter (RS) artifacts are often commonplace in videos captured using rolling shutter CMOS cameras. Various factors, including low shutter frequency, long exposure times, and the movement of the device [44, 73] can cause motion blur and rolling shutter artifacts. RS cameras capture each frame by sequentially scanning pixels row by row as opposed to the global shutter (GS) cameras that capture all the frame pixels at once. This causes rolling shutter artifacts such as skew, wobble or smear if the camera is moving during video capture. Subsequently, as the sensor gets higher in resolution, the potential for rolling shutter artifacts increases due to increase in readout time of pixels in a row [122]. With increasing popularity of CMOS cameras in various computer vision applications [100, 68, 99], which require high-quality high-resolution imaging, it calls for jointly addressing the task of rolling shutter rectification and spatial super-resolution.

Early works [81, 85] studied the problem of multi-image super-resolution for images captured from rolling shutter cameras. In [81], authors assume that one of the images is

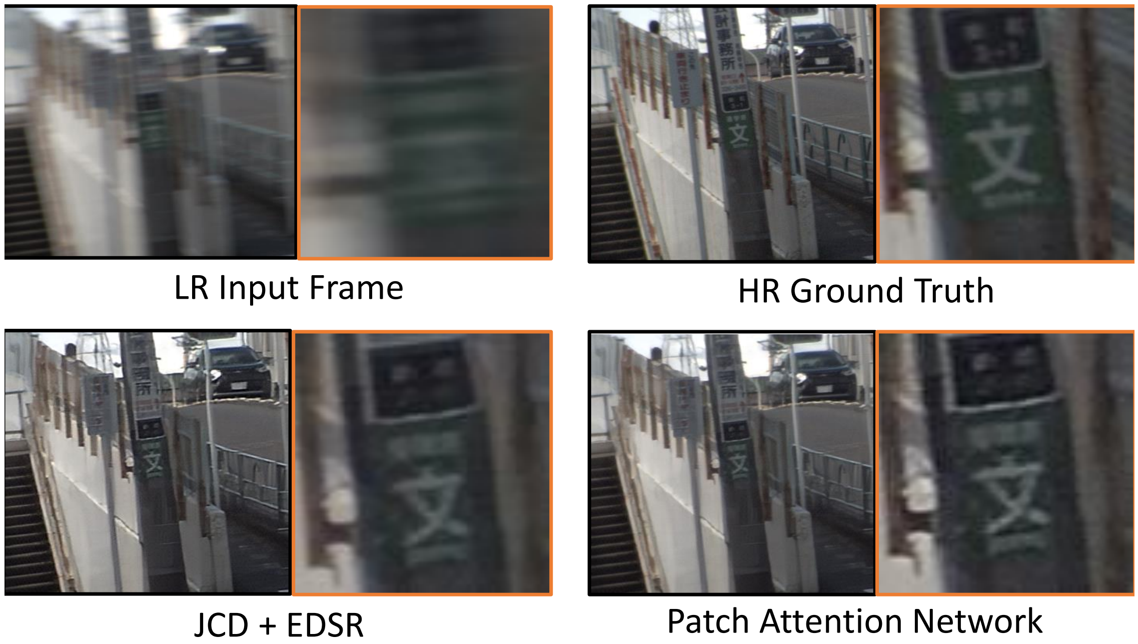


Figure 4.1: **Patch Attention Network.** We show frame on the left and zoomed in patch on the right. **Top Left:** Input LR rolling shutter frame (resized to HR frame resolution). **Top Right:** Global shutter ground truth HR frame. **Bottom Left:** Output of combination of state-of-the-art RSC method (JCD [127]) and SR method (EDSR [60]). **Bottom Right:** Predicted image by Patch Attention Network. It can be observed that the **PatchNet** model generates superior results as compared to state-of-the-art rolling shutter correction and super-resolution method.

free from rolling shutter distortion, and use this image as reference to estimate the row-wise motion of the other images for task of super-resolution. In contrast, an approach to recover high-resolution (HR) image when all the images, captured using burst mode, have rolling shutter artifacts is presented in [85]. However, these multi-image based approaches rely on geometric constraints from multiple views formulating a computationally expensive optimization problem for 6 DoF camera motions.

The success of deep learning methods and the availability large-scale datasets [97, 43, 119, 61] have greatly facilitated the research in video restoration techniques. Video super-resolution (VSR) approaches [47, 45, 119, 104, 10, 112] assume that the camera is global shutter and there are no rolling shutter artifacts. Consequently, the lack of realistic high-resolution datasets with RS effect has restricted the development of learning-based RS correction. Recently, with prevalence of CMOS sensors, rolling shutter correction has received renewed research interest [127, 63]. Authors in [63] proposed a synthetic dataset (FastecRS) for rolling shutter correction by sequentially copying a row of pixels from GS images to obtain RS images. However, it is challenging to obtain RS distorted image and corresponding GS image.

Addressing this issue, a realistic dataset for rolling shutter correction and deblurring (RSCD) was proposed in [127] which includes the GS images and their corresponding RS images for learning based approaches. This new dataset opens avenue for further research towards a realistic and more challenging problem of rolling shutter correction and super-resolution. Authors in [127] also propose a joint correction and deblurring model (JCD) to rectify the rolling shutter correction along with deblurring by utilizing deformable convolutional attention layers. The deformable convolution layers can easily learn geometric variations in object scale, pose, viewpoint and deformations due to their flexible kernel operation as opposed to the fixed kernel operations (size and stride) in traditional convolutional layers. The deformable attention in JCD relies on flow features to learn the displacement field to correct the rolling shutter effect and deblur simultaneously. However, deformable convolution can only obtain local pixel-level information and do not take into account the global information available in neighbouring patches.

To incorporate the global information in features space, we introduce a novel architecture Patch Attention Network (**PatchNet**) which aims to jointly rectify rolling shutter

(RS) artifacts and generate high-resolution frames from a low-resolution video acquired using RS camera. Specifically, we utilize the patch recurrence in the feature space to exploit the patch-level information for the task of rolling shutter correction and super-resolution. Our approach is motivated by the observation that small image patches tend to recur in a captured image [69, 21, 131] and using the combination of patch-level features can span a superior space for super-resolution as compared to bi-linear interpolation or convolution operations alone [18]. Convolution layers have a fixed kernel size so they cannot leverage the information beyond their receptive field [18]. Our Patch Attention network leverages the forward and backward flow information for patches in the feature space to obtain relevant information from correlated neighbours and then performs super-resolution in the feature space to generate high-resolution global shutter frames. Using information from neighbouring patches, **PatchNet** is able to generate superior results as shown in Figure 4.1.

An overview of Patch Attention Network (**PatchNet**) is illustrated in the Figure 4.2. Given a low-resolution rolling shutter (LR-RS) video input, our objective is to recover a high-resolution global shutter (HR-GS) video. Our framework consists of current frame feature encoder $\mathcal{E}$, two flow feature networks ($\mathcal{F}_p$, $\mathcal{F}_f$) with respect to previous and future frame, Patch Attention Module $\mathcal{M}$ and a decoder model $\mathcal{G}$. We utilize the encoder model $\mathcal{E}$ to obtain current frame features X. The flow estimation networks generate forward flow features F_f and backward flow features F_p. The network $\mathcal{M}$ then utilizes deformable convolution followed by patch-level attention to obtain features suitable for RS correction and super-resolution. The flow features guide the generation of HR features by providing attention weights to the unfolded patch tensor at feature level as shown in Figure 4.3.

The deformable convolution provides local information at pixel-level, whereas the patch attention tries to extract global information from neighbouring patches for super-resolution, thereby performing the task of rolling shutter correction and super-resolution. Our Patch Attention Network generates superior global shutter high-resolution image when compared with a model trained by combining the state-of-the-art rolling shutter correction method (JCD) and the image super-resolution method (EDVR) together as shown in Figure 4.1.

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

- We introduce a novel framework **PatchNet**, Patch Attention Network, designed to recover high-resolution global shutter frames form low-resolution rolling shutter video. Unlike prior related work, we *jointly* optimize our model for rolling shutter correction and super-resolution in feature space.
- This is the first work to leverage the combination of local information, using deformable convolution, and optical-flow driven global patch-level information from neighbouring patches to recover a high-resolution global shutter video.
- Our framework demonstrates consistently effective results on two datasets, RSCD [127] and FastecRS [63] with better performance over state-of-the-art approaches due to the joint optimization framework and patch recurrence property, thereby also producing finer visual results.

4.2 Related Work

In this section, we review some recent methods pertaining to video super resolution, rolling shutter correction, and later discuss different attention mechanisms in various computer vision tasks. We provide a characteristic comparison of recent works in Table 4.1.

Video Super-Resolution. Several learning-based approaches have been proposed for video super-resolution [47, 45, 119, 32] for video with no rolling shutter distortions. A deep learning based approach is presented in [47], where the network is trained using the information in the spatial and temporal dimensions of videos for super-resolution. For fast video super-resolution, a draft-ensemble approach is proposed in [59]. The authors in [104, 10] incorporate optical flow estimation models to explicitly account for the motion between neighboring frames. However, accurate flow is difficult to obtain given occlusion and large motions. To account for the motion information, a computationally lighter flow estimation module (TOFlow) is proposed in [119]. DUF [45] overcomes the problem of estimating accurate optical flow by implicit motion compensation using their proposed dynamic upsampling filter network. Pyramid, Cascading and Deformable convolution (PCD) alignment and the Temporal and Spatial Attention (TSA) modules are proposed in EDVR [112] to incorporate implicit motion compensation. Though these approaches leverage optical flow with deformable convolution, they do not leverage the internal patch recurrence across space and time for super-resolution.

Rolling Shutter Correction. Classical works rely on motion estimation to rectify a rolling shutter image. For instance, block matching based approach to estimate global and local motion is presented in [58]. Another approach [26] parameterised the camera motion as a continuous curve and estimated the curve parameters by minimizing non-linear least squares over inter-frame correspondences obtained from a KLT tracker. Extension of the work [26] using inertial measurement is proposed in [48]. Their framework calibrates gyroscope and camera outputs from a single video capture to effectively correct rolling shutter artifacts and to stabilize the video. Authors in [86] utilize prominent curves from the RS image to decipher the varying row-wise motion. They enforce line desirability costs for camera motion estimation as lines can be rendered as curves due to the row-wise scanning in rolling shutter cameras. For two consecutive RS images, one approach [129] proposes to estimate depth-map and motion, by solving Structure-for-Motion (SfM) problem from dense correspondence, to rectify rolling shutter effect. The problem of occlusion aware rolling shutter correction problem for 3D scene is addressed using multiple consecutive frames by authors in [109]. They model the 3D geometry as a layer of planar scenes. First the depth, camera motion, latent layer mask and latent layer intensities are estimated jointly. Then an image formation model is designed using the estimated values to recover the global shutter image. Recently, a configuration with two cameras with different rolling shutter directions is utilized to undo the rolling shutter correction [2].

With success of deep neural network, some learning based approaches are proposed to address the problem of rolling shutter correction and have shown impressive results [129, 84, 63, 127]. A CNN based model with long rectangular kernels is proposed in [84] to estimate the row-wise camera motion from a single rolling shutter image. The camera motion is estimated assuming a simple affine motion model and is used to recover the global shutter image. Authors in [129] address the depth aware rolling shutter correction

Table 4.1: Characteristic comparison of prior works in rolling shutter correction (RSC) and super-resolution (SR). Different from the state-of-the-art approaches, **PatchNet** demonstrates patch-level attention in latent space to exploit internal patch recurrence and global information along with pixel-level attention using deformable convolution.

Methods	Task		Attention	
	RSC?	SR?	Pixel?	Patch?
DUN [63]	✓	✗	✗	✗
VSR-T [12]	✗	✓	✗	✓
JCD [127]	✓	✗	✓	✗
PatchNet (Ours)	✓	✓	✓	✓

using two independent neural networks to estimate dense depth map and camera motion. The rolling shutter correction is performed as a post-processing step, given the estimated dense depth map and camera motion.

More recently, an end-to-end deep learning approach for rolling shutter correction is presented in Deep Unrolling Network [63] trained using synthetic datasets (FastecRS) obtained by sequentially copying a row of pixels from GS images to obtain corresponding RS images. Though these approaches show impressive performance, one major shortcoming is that they have limited performance for the data in realistic setting. It is challenging to obtain RS distorted image and corresponding GS image. Another realistic dataset for joint rolling shutter correction and deblurring (RSCD) is presented in [127]. The dataset is captured using a beam-splitter acquisition system. An RS camera and a GS camera are physically aligned to capture RS distorted blur video as well as GS sharp video pairs, simultaneously. Both of these methods leverage optical flow to address the issue of rolling shutter correction. Joint Rolling Shutter Correction and Deblurring [127] (JCD) utilizes bi-directional optical flow as compared to Deep Unrolling Network [63]. Additionally, JCD leverages deformable convolution with attention for hierarchical features for task of joint rolling shutter correction and deblurring. Deformable convolution [18] greatly enhances capability of modeling geometric transformation at pixel level. This property of deformable convolution layers makes it suitable for RS correction problem. However, for any super-resolution modeling, local (pixel-level) as well as global (patch-level) geometric transformation is necessary. In this work, we leverage the global information, available in the neighbouring patches using our Patch Attention Network, along with the local pixel-level information using deformable convolutions for the task of joint rolling shutter correction and super-resolution.

Attention Modelling. Attention mechanism has garnered a lot of research interest in computer vision tasks due to their learnable guidance ability. Various adaptations of attention mechanism have shown promising results in object recognition [3, 13], image generation [120] and image super-resolution [124]. Recently, different attention models are proposed for video deblurring [115], video super-resolution [31] and video interpolation [16]. In [16], attention is applied channel-wise on concatenated down-shuffled frames for video interpolation. Authors in [31] explore attention in latent space for the task of video deblurring and interpolation. A patch-wise attention network (Patchwork) is presented in [13] for object detection and segmentation. Patchwork processes only a portion of the features for further processing thereby reducing the computational cost and achieving superior performance. Transformer based attention at block-level is also utilized in [12] to generate high-resolution video. The spatio-temporal convolutional self-attention is leveraged followed by bi-directional optical-flow based feed-forward network for feature learning and then a reconstruction model is used for super-resolution. Unlike this approach, which only tackles video super-resolution, our patch-level attention is guided by the flow-features and utilizes deformable convolution to address rolling shutter correction as well as video super-resolution. Our approach performs super-resolution in the feature space followed by video reconstruction as opposed to the approach in [12] where a generator model is used to perform super-resolution from a low-resolution feature.

4.3 Approach

Problem Statement. Given a low-resolution rolling shutter (LR-RS) video, our goal is to rectify the rolling shutter artifacts and generate a high-resolution global shutter (HR-GS) image. We propose to recover a high-resolution global shutter video by modelling attention in the feature space at patch-level. Our hypothesis is that the neighbouring patches in the latent space can help project more informative patches for the task of rolling shutter correction and super-resolution. The combination of neighbouring patches along with their respective optical flow representations can help synthesize patches in a larger space as compared to bi-linear interpolation or convolutional layers which has a fixed geometric structure (fixed kernel shape).

Notations. Let the low-resolution RS video be denoted by $\mathbf{V}_{LR} = \Big[\mathsf{V}_1,\ \mathsf{V}_2, \cdots,\ \mathsf{V}_N\Big]$, with N number of frames, where $\mathsf{V}_t \in \mathbb{R}^{H_I \times W_I \times 3}$ and t denotes the time step. Let $\mathcal{E}$ be the feature encoder for the i^{th} frame, $\mathcal{F}_p$ and $\mathcal{F}_f$ be the branches corresponding to the

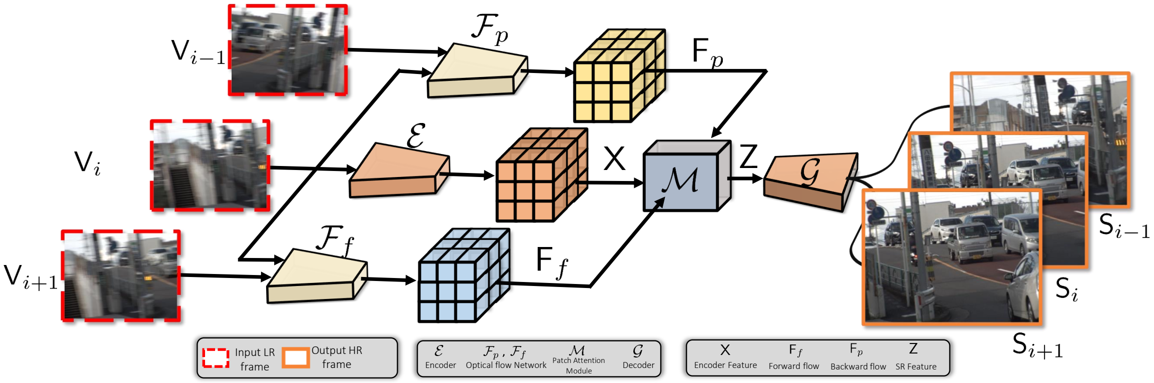


Figure 4.2: Overview of the proposed approach. Given a low-resolution input video frames V_{i-1}, V_i and V_{i+1}, we extract the feature representation X corresponding to frame V_i using the encoder network $\mathcal{E}$ and the flow features F_p and F_f with respect to the past frame V_{i-1} and future frame V_{i+1}, respectively. Patch Attention Network $\mathcal{M}$ utilizes deformable convolution and patch-level attention to obtain high-resolution features Z that can recover global shutter image (see sec. 4.3.2). The high-resolution feature Z is then used by the decoder network $\mathcal{G}$ to produce high-resolution global shutter frames S_{i-1}, S_i and S_{i+1}.

optical flow of current frame (V_i) with respect to previous frame (V_{i-1}) and future frames (V_{i+1}), respectively. The output of each network $\mathcal{E}$, $\mathcal{F}_p$ and $\mathcal{F}_f$ is a feature at different scales extracted from different layers of the network. Let the encoder representation of the the i^{th} frame (V_i) be denoted by X such that $X \in \mathbb{R}^{H \times W \times C}$. Similarly, let the optical flow features obtained from the forward and backward flow networks $\mathcal{F}_p$ and $\mathcal{F}_f$, be denoted by F_f and F_p, respectively.

We aim to generate a high-resolution GS video $\mathbf{V}_{HR} = \Big[S_1,\ S_2, \cdots ,\ S_L\Big]$, where $S_t \in \mathbb{R}^{aH_I \times aW_I \times 3}$ using the Patch Attention Network $\mathcal{M}$. Patch Attention Network leverages the encoder features X, the optical flow features F_f and F_p by unfolding them into $P \times P$ patches and finding correlation between the forward flow and backward flow patches as attention maps. To leverage the patch-recurrence property, we need to obtain correlated neighbouring patches for each input patch-level feature. This can be achieved by representing the problem of finding correlated patches as mapping a query to a set of key-value pairs in a retrieval problem [22]. In key-value based retrieval problem, key acts as an unique identifier for different values and query is matched with various keys to obtain respective values. In our case, we assume that that forward flow acts as the key representation ($\mathcal{K}$) for different patch values of encoder features ($\mathcal{V}$), and utilize the backward flow as query ($\mathcal{Q}$) to retrieve correlated encoder features value ($\mathcal{V}$). The resultant informative patch representation is Z which is obtained using key-query attention similarity computed with the help of its neighbouring patches. This representation is the input to the reconstruction model $\mathcal{G}$ to generate high-resolution global shutter video.

4.3.1 Features Extraction

The encoder $\mathcal{E}$ is a trainable convolutional neural network which projects the current RS-LR input frame (V_i) into a latent space such that $\mathsf{X} = \mathcal{E}(\mathsf{V}_i)$, where $\mathsf{X} \in \mathbb{R}^{H \times W \times C}$. The forward flow network ($\mathcal{F}_f$) takes the current frame (V_i) and the future frame (V_{i+1}) to generate forward warped feature, whereas the backward flow network ($\mathcal{F}_p$) generates the backward warped feature using the current frame (V_i) and the past frame (V_{i-1}). The forward and backward warped features are given by equations below.

$$\mathsf{F}_f = \mathcal{F}_f(\mathsf{V}_{i+1}, \mathsf{V}_i) \tag{4.1}$$

$$\mathsf{F}_p = \mathcal{F}_p(\mathsf{V}_i, \mathsf{V}_{i-1}) \tag{4.2}$$

The frame representation generated by the encoder $\mathcal{E}$ and the forward and backward warped flow features generated by $\mathcal{F}_f$ and $\mathcal{F}_p$ are then used by the Patch Attention Network $\mathcal{M}$ to generate features that can rectify rolling shutter effect and are utilized to synthesize high-resolution frame.

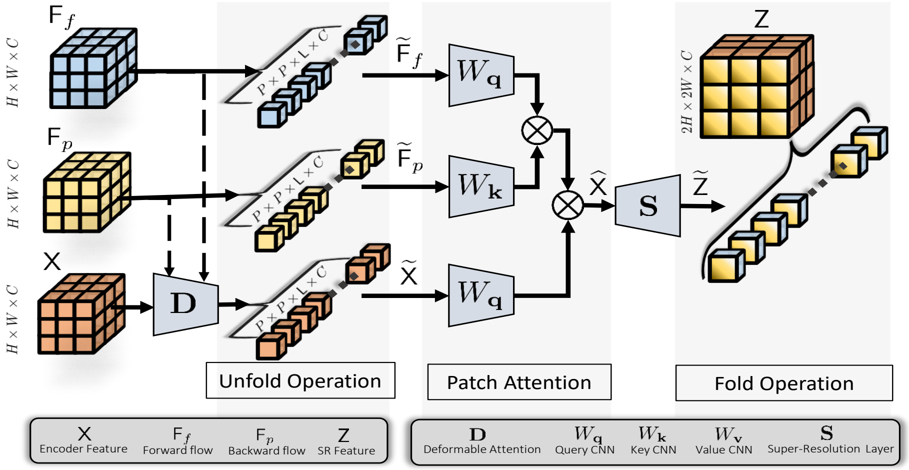


Figure 4.3: Overview of Patch Attention Network. Given the encoder feature X and the motion features F_p and F_f, we first utilize the deformable attention network $\mathbf{D}$ [127] for to incorporate motion information at pixel-level and unfolded into $P \times P$ patches to obtain the patch-level encoder feature $\widetilde{\mathsf{X}}$. Similarly, the motion features F_p and F_f are unfolded into patches of size $P \times P$, represented by $\widetilde{\mathsf{F}}_p$ and $\widetilde{\mathsf{F}}_f$, respectively. The patch-level features $\widetilde{\mathsf{F}}_p$ and the encoder feature $\widetilde{\mathsf{X}}$ form input to the key-value networks $\mathbf{W}_k$ and $\mathbf{W}_v$, respectively. The patch-level flow feature $\widetilde{\mathsf{F}}_f$ acts as query input to $\mathbf{W}_q$ to find the correlated features ($\widehat{\mathsf{X}}$) from the key-value pair $\widetilde{\mathsf{F}}_p$ and $\widetilde{\mathsf{X}}$. Finally, a super-resolution layer is used to generate high-resolution features at patch-level $\widetilde{\mathsf{Z}}$, followed by folding operation to obtain the high-resolution features Z, which is used to generate high-resolution global shutter frames.

4.3.2 Patch Attention Network

We aim to obtain enhanced features to generate a high-resolution global shutter image. In order to effectively integrate the information from the flow features $(\mathsf{F}_p, \mathsf{F}_f)$ and encoder feature X, we propose a patch-level attention based module Patch Attention Network (**PatchNet**). The PatchNet module $\mathcal{M}$ utilizes deformable convolution and patch-level attention to extract correlated information from neighbouring patches. Then a super-resolution model S is utilized to produce high-resolution features. Figure 2.3 presents the overview of the patch attention used in the **PatchNet**. First, we employ a deformable convolution attention module $\mathbf{D}$ to incorporate the bi-directional motion information at pixel-level. The deformable attention module fuses the bi-directional motion information with the encoder feature and then applies unfolding operation to extract $P \times P$ patches resulting in the feature $\widetilde{\mathsf{X}}$ of shape $P \times P \times L \times C$ using the unfolding operation, where L is the total number of patches such that $L = H * W/P * P$. The output feature $\widetilde{\mathsf{X}}$ is given by the equation 4.3.

$$\widetilde{\mathsf{X}} = \mathbf{D}(\mathsf{X}, \mathsf{F}_p, \mathsf{F}_f) \tag{4.3}$$

Similar to the unfolding operation in module $\mathbf{D}$, we also divide bi-directional flow features in $P \times P$ patches. The patch-level feature representation of the forward flow feature and the backward flow feature are represented by $\widetilde{\mathsf{F}_f}$ and $\widetilde{\mathsf{F}_p}$, respectively. To extract patch-level information, we use three convolutional networks $\mathbf{W}_q$, $\mathbf{W}_k$ and $\mathbf{W}_v$ to capture patch-level information with help of bi-directional flow features. We then generate the query, key and value using the patch-level encoder features and bi-directional flow features. Since, we want to generate high-resolution patches of the patch-level encoder feature ($\widetilde{\mathsf{X}}$), we assume the patch-level backward flow feature F_p and the patch-level encoder feature $\widetilde{\mathsf{X}}$ forms key-value pair. Hence, we use the network $\mathbf{W}_v$ to compute the value representation ($\mathcal{V}$) using $\widetilde{\mathsf{X}}$ and the network $\mathbf{W}_k$ to compute the key representation ($\mathcal{K}$) using $\widetilde{\mathsf{F}_p}$. We extract the query representation ($\mathcal{Q}$) using the network $\mathbf{W}_q$ with the forward flow features $\widetilde{\mathsf{F}_f}$ as input. The query, key and value representation are computed using the equations below.

$$\mathcal{Q} = \mathbf{W}_q\Big(\widetilde{\mathsf{F}}_f\Big), \quad \mathcal{K} = \mathbf{W}_k\Big(\widetilde{\mathsf{F}}_p\Big), \quad \mathcal{V} = \mathbf{W}_v\Big(\widetilde{\mathsf{X}}\Big) \tag{4.4}$$

The patch-level attention is computed by first calculating the attention maps $\sigma(\mathcal{Q}^T\mathcal{K})$, where σ is a ReLU activation function. Then the weighted patch-level features are extracted by multiplying the attention maps with the value representation $\mathcal{V}$. The feature obtained after this operation is denoted by $\widehat{\mathsf{X}}$ and given by the following equation.

$$\widehat{\mathsf{X}} = \sigma\big(\mathcal{Q}^T\mathcal{K}\big)\mathcal{V} \tag{4.5}$$

We then utilize a super-resolution layer $\mathbf{S}$ to obtain high-resolution patch representation $\widetilde{\mathsf{Z}}$ using the equation 4.6.

$$\widetilde{\mathsf{Z}} = \mathbf{S}\Big(\widehat{\mathsf{X}}\Big) = \mathbf{S}\big(\sigma\big(\mathcal{Q}^T\mathcal{K}\big)\mathcal{V}\big) \tag{4.6}$$

Then, unfolding operation is applied to the high-resolution patch features $\widetilde{\mathsf{Z}}$ to obtain high-resolution reconstruction features, Z. These high-resolution reconstruction features are then utilized to recover the high-resolution global shutter frame corresponding to low-resolution rolling shutter frame V_i

4.3.3 GS-HR Video Generation

To recover the global shutter frame and perform super-resolution from a given low-resolution rolling shutter frame V_i, we employ a generative neural network $\mathcal{G}$ that transforms the high-resolution features to high-resolution global shutter frame. The aggregated reconstruction features Z generated from **PatchNet** are then utilized to obtain high-resolution global shutter frame (S_i) corresponding to V_i.

4.3.4 Loss Function

Our objective function is composed of Charbonnier loss ($\mathcal{L}_c$) [14] as it helps to preserve edges, perceptual loss ($\mathcal{L}_p$) for the predicted results $\mathbf{V}_{HR}$ to improve perceptual quality and a total variational loss ($\mathcal{L}_v$) applied to the estimated displacement fields to smooth the forward and backward warping processes in bi-directional flow networks. The total loss function ($\mathcal{L}$) is given by:

$$\mathcal{L} = \lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_v \mathcal{L}_v \tag{4.7}$$

where, λ_c, λ_p and λ_v are weights for the loss terms $\mathcal{L}_c, \mathcal{L}_p,$ and $\mathcal{L}_v$, respectively.

4.4 Experiments

4.4.1 Datasets and Evaluation Metrics

We evaluate the performance of our approach using publicly available RSCD [127] and synthetic Fastec-RS [63] datasets which have been used in prior RS correction works.

BS-RSCD Dataset. BS-RSCD [127] is a dynamic urban environment dataset which includes both ego-motion and object-motion. There are total of 80 short video sequences with 50 frames each in this dataset. The training set includes 50 video sequences (2500 image pairs), the validation set includes 15 sequences with 750 image pairs and the test set contains 750 image pairs. This dataset is composed of RS frames along with their corresponding GS frames. All frames in video sequence are of 640 × 480 resolution. We down-sample the RS frames, in the dataset to 320 × 240 to generate low-resolution rolling shutter frames for training in all our experiments.

Fastec-RS Dataset. The Fastec-RS dataset [63] is a synthetic dataset captured using a professional Fastec TS51 high speed global shutter camera. Total of 76 image sequence are captured at 2400 fps with a resolution of 640 × 480 pixels in mainly urban environment. Each sequence synthesizes 34 rolling shutter images to obtain 2584 image pairs. To synthesize the rolling shutter image, pixels in each row are copied sequentially from the captured GS images and down-sampled to the RS frames at 320 × 240 resolution to generate a low-resolution rolling shutter frames.

Evaluation Metrics. For quantitative evaluation, we compare three metrics that evaluate different aspects of output image quality: Peak Signal-to-Noise Ratio (PSNR) [37], Structural Similarity Index Measure (SSIM) [113] and Learned Perceptual Metric (LPIPS) [123].

4.4.2 Implementation Details

Our framework is implemented in PyTorch [80]. All the experiments are trained for 400 epochs with a batch size of 8. We use ADAM [54] optimizer with initial learning rate of 0.0001 with cosine annealing scheduler. The loss weights λ_c, λ_p and λ_v are set to 10, 1 and 0.1 , respectively. The deformable convolution attention layer is adopted from JCD approach [127] with deformable groups as 8.

4.4.3 Qualitative Results

We compare our work with combination of state-of-the-art rolling shutter correction (JCD [127] and super-resolution works such as bi-linear interpolation and EDSR [60]. Figure 4.4 and Figure 4.5 show some examples of our proposed **PatchNet** against various baselines. For combination of bi-linear interpolation and JCD approach, it can be noticed that the quality of output image is poor. It is due to the fact that the bi-linear interpolation is not learnable when compared to other approaches and hence cannot learn the pixel displacement for super-resolution. From Figure 4.4, it can be observed that our

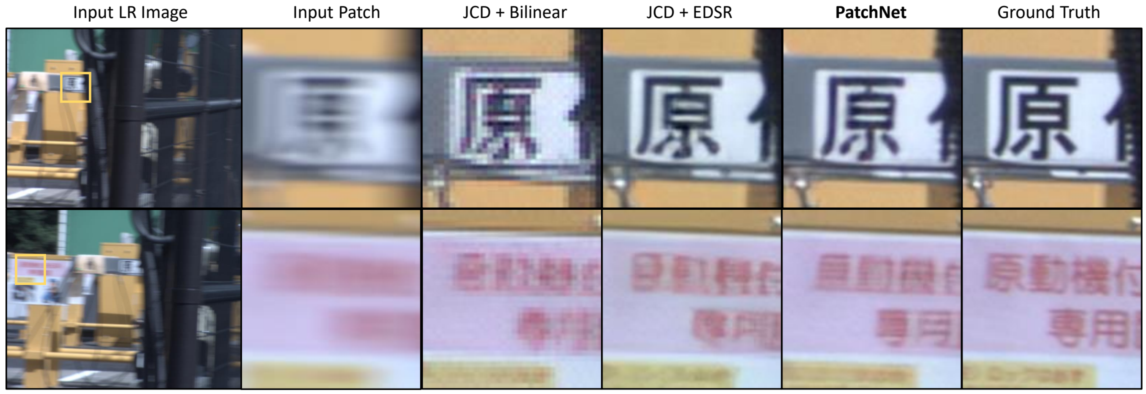


Figure 4.4: **Qualitative results comparison on BS-RSCD dataset.** First column consists of three low-resolution rolling shutter input frames . Second column and last column are the input and ground-truth crop of the input frame region marked in gold. As opposed to JCD [127] + bi-linear interpolation, JCD [127] + EDSR [60] and **PatchNet** performs better as they utilize learnable module for super-resolution. **PatchNet** produces visually sharper results as it learns RS correction and super-resolution jointly unlike JCD [127] + EDSR [60].

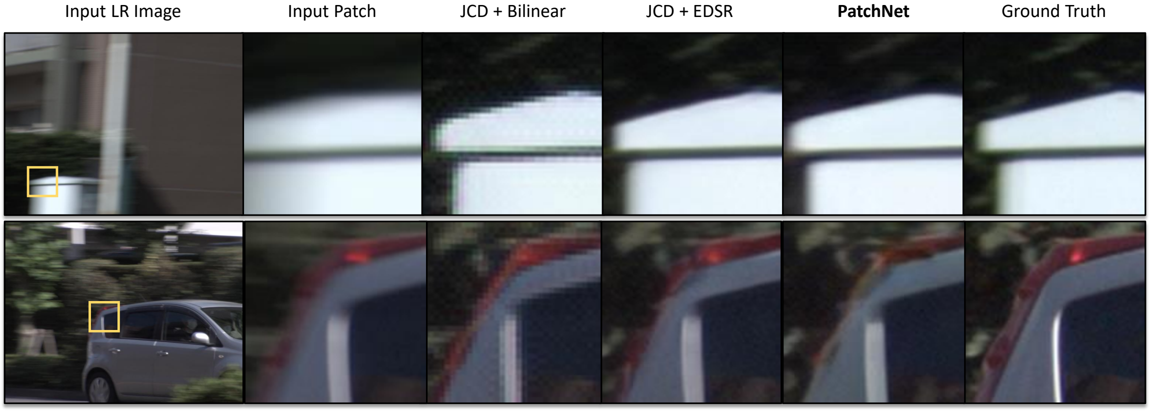


Figure 4.5: **Additional qualitative results on BS-RSCD dataset.** First column consists of the low-resolution rolling shutter input frame for two videos. Second column and last column are the input and ground-truth crop of the input frame region marked in gold. As opposed to JCD [127] + bi-linear interpolation, JCD [127] + EDSR [60] and **PatchNet** performs better as it utilizes patch-recurrence property along with deformable convolution. **PatchNet** produces visually sharper results as it can extract available information from neighbouring patches as opposed to JCD [127] + EDSR [60].

Table 4.2: Quantitative results comparison of **PatchNet** with the state-of-the-art baselines on RSCD and Fastec-RS datasets.

Methods		**RSCD**			**FastecRS**		
RSC	**SR**	**PSNR ↑**	**SSIM ↑**	**LPIPS↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
JCD	Bi-linear Interpolation	22.74	0.581	0.463	23.87	0.655	0.339
JCD	Transposed Conv	24.15	0.628	0.328	24.12	0.632	0.262
JCD	SAN	24.37	0.633	0.305	24.07	0.643	0.281
JCD	EDSR	24.94	0.650	0.263	24.67	0.713	0.187
Deep Unrolling Net	Bi-linear Interpolation	21.64	0.552	0.489	25.34	0.792	0.185
Deep Unrolling Net	Transposed Conv	24.02	0.602	0.342	25.88	0.801	0.179
Deep Unrolling Net	SAN	24.16	0.621	0.322	26.10	0.807	0.165
Deep Unrolling Net	EDSR	24.58	0.634	0.286	26.43	0.810	0.147
Deep Unrolling Net (SR Input)		25.14	0.729	0.159	27.00	**0.825**	0.108
JCD (SR Input)		26.42	0.757	**0.122**	24.84	0.778	0.107
PatchNet (LR Input)		**27.38**	**0.793**	0.144	**27.12**	0.811	**0.103**

approach is able to produce sharp frames with fine details in text (top and bottom row) and in objects (middle row). Other approaches tackle the problem of rolling shutter correction and super-resolution separately and hence cannot exploit the information available completely when compared it **PatchNet**. Also, as our approach is extracting information by leveraging the neighbouring patch information in feature space, along with deformable attention, it produces visually more appealing videos. Additional results on frames from two other videos are shown in Figure 4.5. It can be observed that the combination of JCD and EDSR generates blurry results as super-resolution is performed after rolling shutter correction. Our approach overcomes this issue by jointly learning rolling shutter correction and super-resolution in feature space, thereby producing finer visual results.

4.4.4 Quantitative Results

We compare our proposed approach, with patch size of 8 for patch-attention, against different combinations of the state-of-the-art approaches for rolling shutter correction and super-resolution. Quantitative results comparison with the state-of-the-art baselines are shown in Table 4.2.

For the task of joint rolling shutter correction and super-resolution in BS-RSCD dataset, the proposed method achieves improvement of 2.32dB in average PSNR when compared with the best combination of RS correction and super-resolution approaches (JCD + EDSSR). It can also be observed that **PatchNet**, which takes LR rolling shutter video input, even outperforms JCD and Deep Unrolling Net methods which only perform RS correction using high-resolution input by a margin of 0.98dB and 2.24dB, respectively. It can be attributed to the patch information used for the task of joint learning.

Similar trends can be observed for the performance of our proposed approach on the synthetic Fastec-RS dataset. The state-of-the-art rolling shutter correction approach, JCD, which works better on RSCD dataset, doesn't outperform Deep Unrolling Network [63] even though JCD relies on deformable attention. It could be due to the use of bi-directional motion estimation used in JCD which may not be best to model rolling shutter effect in synthetic dataset. Compared to these methods, our approach uses lower resolution input and still outperforms them by generating high-resolution global shutter frames. It is due to the use of patch-level attention utilized by the **PatchNet** along with the deformable attention, which help learn the motion model better even in the Fastec-RS dataset.

Table 4.3: Impact of patch size on performance of **PatchNet** on the benchmark BS-RSCD dataset.

Patch Size	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
2×2	25.29	0.734	0.165
4×4	27.24	0.778	0.157
8×8	**27.38**	**0.793**	**0.144**

Ablation study on impact of patch-size on performance of **PatchNet** is presented in Table 4.3. The performance with patch size 2×2 is poor as it is not able to extract global information from neighbouring patches as the size is too small. We observed the performance of model with patch sizes 8×8 and 4×4 shows a significant improvement over the baselines (2dB gain). We see that the performance increases with size of the patch size. However, the performance for the patch size 8×8 is only slightly higher than that of 4×4. This suggest that our model can perform well with patch size of 4×4 without significant drop in the performance.

4.5 Conclusion

We present Patch Attention Network (**PatchNet**) to recover high-resolution global shutter frames from low-resolution rolling shutter video. The proposed approach employs patch-level attention in feature space to extract information from neighbouring patches using the key-query similarity and deformable convolution, respectively. Specifically, the patch attention module obtains correlation maps between neighbouring patches to extract information relevant for *simultaneous* rolling shutter correction and super-resolution. Our main contribution over existing approaches is in jointly learning how to do rolling shutter correction and super-resolution, which have been treated separately in the past and leveraging patch-recurrence property through attention mechanism. Experiments on standard datasets show the efficacy of our proposed approach over state-of-the-art methods.

Chapter 5

Application in Biomedical Image Enhancement

While machine learning approaches have shown remarkable performance in biomedical image analysis, most of these methods rely on high-quality and accurate imaging data. However, collecting such data requires intensive and careful manual effort. One of the major challenges in imaging the Shoot Apical Meristem (SAM) of Arabidopsis thaliana, is that the deeper slices in the $z-$stack suffer from different perpetual quality related problems like poor contrast and blurring. These quality related issues often lead to the disposal of the painstakingly collected data with little to no control on quality while collecting the data. Therefore, it becomes necessary to employ and design techniques that can enhance the images to make them more suitable for further analysis. In this paper, we propose a data-driven Deep Quantized Latent Representation (DQLR) methodology for high-quality image reconstruction in the Shoot Apical Meristem (SAM) of Arabidopsis thaliana. Our proposed

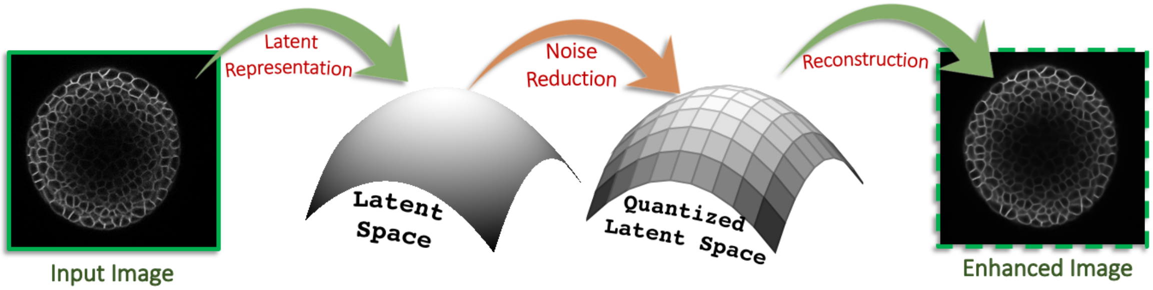


Figure 5.1: Conceptual Overview of DQLR. The latent representation of the collected image is quantized using $k-$means over the entire dataset [108]. This quantized representation is then used to reconstruct the enhanced image.

framework utilizes multiple consecutive slices in the z-stack to learn a low dimensional latent space, quantize it and subsequently perform reconstruction using the quantized representation to obtain sharper images. Experiments on a publicly available dataset validate our methodology showing promising results.

5.1 Introduction

Automated analysis in biomedical research is critical to provide researchers with concrete evidence to prove any proposed hypothesis without any bias. However, automated image analysis requires high-quality imaging data. Image quality-related problems are often encountered while imaging deeper layers of the Shoot Apical Meristem (SAM) of arabidopsis thaliana [62]. These quality-related problems hinder automated analysis and often lead to disposal of painstakingly collected data. To this end, we propose a data driven Deep Quantized Latent Representation (DQLR) framework for high-quality imaging data reconstruction of the $z-$stack of the SAM. In this work, we propose to project noisy stack in a latent space, quantize the latent representations and utilize the quantized latent representations for reconstruction of enhanced $z-$stack (see Fig. 5.1 for conceptual overview).

Overview. An architectural overview of our approach is illustrated in Fig. 5.2. During training, the encoder E compresses i^{th} input slice image to a latent representation x_i. The consecutive slices in the z-stack are correlated which implies that they must be correlated in the latent space as well. We employ a recurrent neural network (RNN) R to learn this correlated representation $\{y_i, y_{i+1}, \cdots, y_{i+n}\}$ by passing the latent vector $\{x_1, x_2, \cdots, x_n\}$ through R. The compressed representation x_i is processed through R_i to learn the inter-correlation between this latent representation of the consecutive slices $\{x_i, x_{i+1}, \cdots, x_{i+n}\}$ during training. RNN generated latent codes $\{y_i, y_{i+1}, \cdots, y_{i+n}\}$ are then used as input to quantization module Q_i. Q_i learns a vector dictionary for quantized representation of the network and generates a quantized latent code $\{y_i^q, y_{i+1}^q, \cdots, y_{i+n}^q\}$. In our proposed method, the quantization of the latent code will remove the noisy component of $\{y_i\}$ and the reconstructed/predicted images using the quantized latent codes by generator G should be enhanced. During testing, we pass one slice at a time from the $z-$stack, compress it using the encoder, predict the correlated latent codes using the RNN, and finally quantize it using the quantization dictionary learned during the training stage using Q. This quantized code is then used to reconstruct and predict enhanced consecutive slices from the given $z-$stack.

5.2 Related Work

In this section we describe prior works closely related to the our proposed method. Our method closely relates to reconstruction using auto-encoders [89] and enhancement in the compressed domain [108, 103, 116].

Auto-Encoders. Variations of auto-encoders are extensively used in reconstruction tasks by compressing the input to a latent representation and using the latent representation to retrieve the input as close as possible [108, 89, 67]. However, often the reconstructed images are blurry due to inherent nature of Mean Square Error (MSE) loss to produce blurry results. In the proposed approach we also include Structural Similarity Index (SSIM) [126] loss to enhance the visual results.

Compressed Domain Enhancement. Some works have tried to enhance the images in the compressed domain. In [103] a method based on a contrast measure defined within the discrete cosine transform (DCT) domain is proposed to enhance the image. Attention based video enhancement is proposed in [31]. Authors in [108] propose a vector quantized variations auto-encoder for reconstruction of various media input. We adopt their approach of vector quantization in our framework. However, we exploit the input data correlation using RNN for enhancement task as opposed to reconstruction in [108] where ground truth data was available.

5.3 Methodology

We propose a Deep Quantized Latent Representation (DQLR) framework for enhancing $z-$stack imaging in SAM of Arabidopsis thaliana. We apply quantization in the latent space of the noisy $z-$stack for enhanced reconstruction. In this section, we first formulate the problem statement and then explain our proposed approach in details.

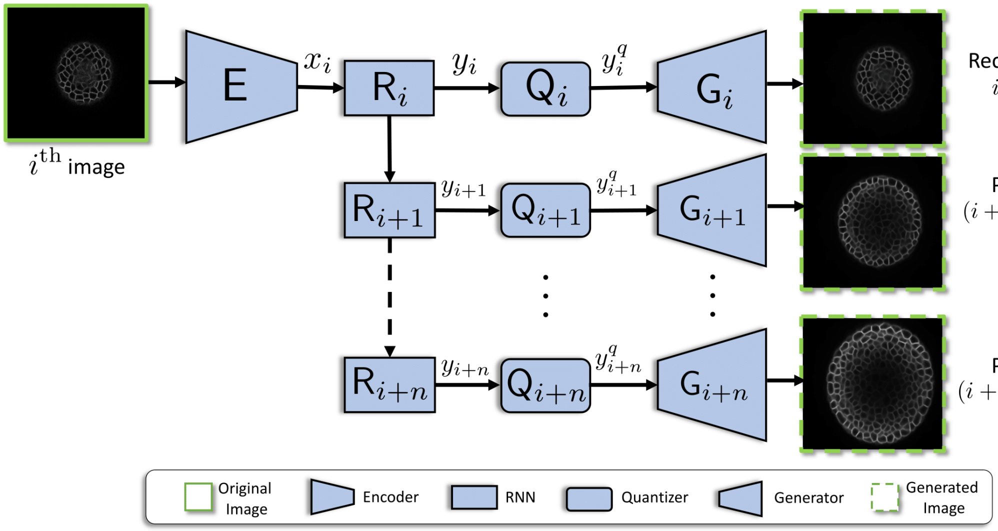


Figure 5.2: **Architectural Overview of DQLR (for one slice of the stack).** Encoder E encodes input image to x_i. Recurrent Neural Network (RNN) module generates correlated codes for reconstruction (y_i) and prediction ($\{y_i, y_{i+1}, \cdots, y_{i+n}\}$). Quantizer module Q$_i$ quantizes the latent codes and Generator G reconstructs/predicts the images.

5.3.1 Problem Formulation

Given a $z-$stack $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$, with $\mathbf{z}_i$ being the i^{th} slice in the stack from the top, we aim to reconstruct $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{z}}_1, \widehat{\mathbf{z}}_2, \cdots, \widehat{\mathbf{z}}_n\}$ such that $\widehat{\mathbf{z}}_i$ is the visually enhanced slice compared to $\mathbf{z}_i, \forall i = 1, 2, \cdots, n$. Let there be a latent representation of input noisy $z-$stack $\mathbf{X_Z} = \{x_1, x_2, \cdots, x_n\}$ where x_i is the latent representation corresponding to the i^{th} slice $\mathbf{z}_i$. Since the slices in $z-$stack are correlated in the pixel space, their latent representations should inherit the same property in the latent space. Therefore, corresponding to each latent representation $\mathbf{X_Z}$ let there be a latent representation $\mathbf{Y_Z} = \{y_1, y_2, \cdots, y_n\}$ such that all $\{y_i\}$ are correlated.

We propose to generate visually enhanced $z-$stack by quantizing the latent representation of the noisy input stack. Our hypothesis is that each correlated latent representation y_i of a slice in the $z-$stack consists of two components; the quantized representation y_i^q

and the noise representation y_i^{noise} of y_i, such that $y_i = y_i^q + y_i^{\text{noise}}$. Hence, noise component y_i^{noise} can be removed by applying quantization on the correlated latent codes leaving the representation y_i^q required to generate the enhanced image $\widehat{\mathbf{z}}_i \; \forall \; i = 1, 2, \cdots, n$.

5.3.2 Proposed Approach

Our proposed framework is shown in Figure 5.2. It consists of four components: the encoder network E, the recurrent neural network R, the quantization module Q and the generator network G. The encoder network is used to extract latent representation for each slice in the noisy input stack. The recurrent neural network utilizes the latent representations to generate correlated latent representations. These correlated representations are quantized to reduce noise in the latent space by the quantization module. Finally, the quantized representations are used to generate an enhanced $z-$stack.

Input Latent Representation. We employ a convolutional neural network as an encoder E which extracts the latent representation for each slice in a given noisy $z-$stack such that

$$\mathsf{E}(\mathbf{Z}) = \mathsf{E}(\{\ \mathbf{z}_1,\ \mathbf{z}_2, \cdots,\ \mathbf{z}_n\}) = \{\ x_0,\ x_1, \cdots,\ x_n\} \tag{5.1}$$

where x_i is latent representation corresponding to slice $\mathbf{z}_i$. A set of correlated representations is generated by the recurrent neural network for the latent representations extracted from the encoder E to incorporate the z-resolution dynamics of the $z-$stack in the latent representations.

Recurrent Neural Network (RNN). The consecutive slices in a $z-$stack capture 3D-structure of any cell in the plant. Thus, there must be a correlation between the consecutive slices. The latent representation $\mathbf{X_Z}$ of the noisy input $\mathbf{Z}$ should also be correlated in some space $\mathbf{Y_Z}$. Thus, we employ a recurrent neural network R_i to transform the i^{th} noisy latent representation to the correlated latent representation as RNN can capture dynamics of the sequence given by

$$y_{i+1} = \mathsf{R}_i(y_i, h_i) \tag{5.2}$$

where h_0 is the hidden state sampled randomly from a Gaussian distribution and $h_i = x_{i-1}\ \forall\ i > 0$. Here, we aim to capture the $z-$resolution dynamics of the stack unlike traditional recurrent neural network where temporal dynamics of the sequence is captured.

Deep Quantized Latent Representation. We propose that a data driven quantization of the latent representation can reduce the average noise in the stack and enhance it visually. In order to quantize the latent representation, we employ vector quantization dictionary learning algorithm as proposed in [108], represented as $\mathsf{Q_i}$ in our framework.

Enhanced Stack Generation. We employ a generative model G to transform the quantized representations into an enhanced stack $\widehat{\mathbf{Z}}$. The quantized representations $\mathbf{Y}^q_\mathbf{Z}$ are used by the generator G to synthesize enhanced stack $\widehat{\mathbf{Z}} = \{\widehat{\mathbf{z}}_1, \widehat{\mathbf{z}}_2, \cdots, \widehat{\mathbf{z}}_n\}$ such that $\widehat{\mathbf{z}_i}$ is the visually enhanced image of the slice $\mathbf{z}_i$ in the noisy stack $\mathbf{Z}$.

5.3.3 Optimization

Our optimization function consists of the Mean Squared Error (MSE) pixel reconstruction loss, the Structural Similarity (SSIM) loss [126] and quantization loss as defined in [108]. Please note that we do not have de-noised image as ground truth. We assume that the quantized latent codes should reduce noise when it is used by generator G to reconstruct the stack. Results in section 5.4 demonstrate the validity of this assumption.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} + \lambda_{\text{s}} \mathcal{L}_{\text{ssim}} + \lambda_{\text{q}} \mathcal{L}_{\text{quant}} \tag{5.3}$$

We briefly describe the loss functions below. Define P as the total number of non-overlapping patches in a given image, N as total number of pixels in P, and α and β as the generated and ground truth image, respectively.

$$\mathcal{L}_{\text{mse}}(P) = \frac{1}{N} \sum_{p \in P} \|\alpha(p) - \beta(p)\|_2$$

$$\mathcal{L}_{\text{ssim}}(P) = \frac{1}{N} \sum_{p \in P} 1 - \text{SSIM}(p),$$

$$\text{with, } \text{SSIM}(p) = \left(\frac{2\mu_\alpha \mu_\beta + C_1}{\mu_\alpha^2 + \mu_\beta^2 + C_1} \right) \left(\frac{2\sigma_\alpha \sigma_\beta + C_2}{\sigma_\alpha^2 + \sigma_\beta^2 + C_2} \right)$$

where, $\mu_{(\cdot)}$ and $\sigma_{(\cdot)}$ are computed with a Gaussian filter with standard deviation σ_G, $C_1 < 1$ and $C_2 < 1$ are constants introduced to handle division by zero issue, λ_{s} and λ_{q} weights for SSIM and quantization loss, respectively. For $\mathcal{L}_{\text{quant}}$, we use the loss function as proposed in [108] on the correlated latent space $\mathbf{Y_Z}$ and dictionary $\mathbf{D} = \{d^1, d^2, \cdots, d^k\}$, where $k = 128$ is length of dictionary to learn for quantization. 0

5.4 Experimentation and Results

Datasets. We used the publicly available Confocal Membrane dataset [114] consisting of six plants. We train our model using four plant stacks, and use one plant stack each for validation and testing.

Qualitative Results. Fig. 5.3 shows few examples of the reconstructed slices from the $z-$stack using the our approach along with the input slice. It can be observed that our pro-

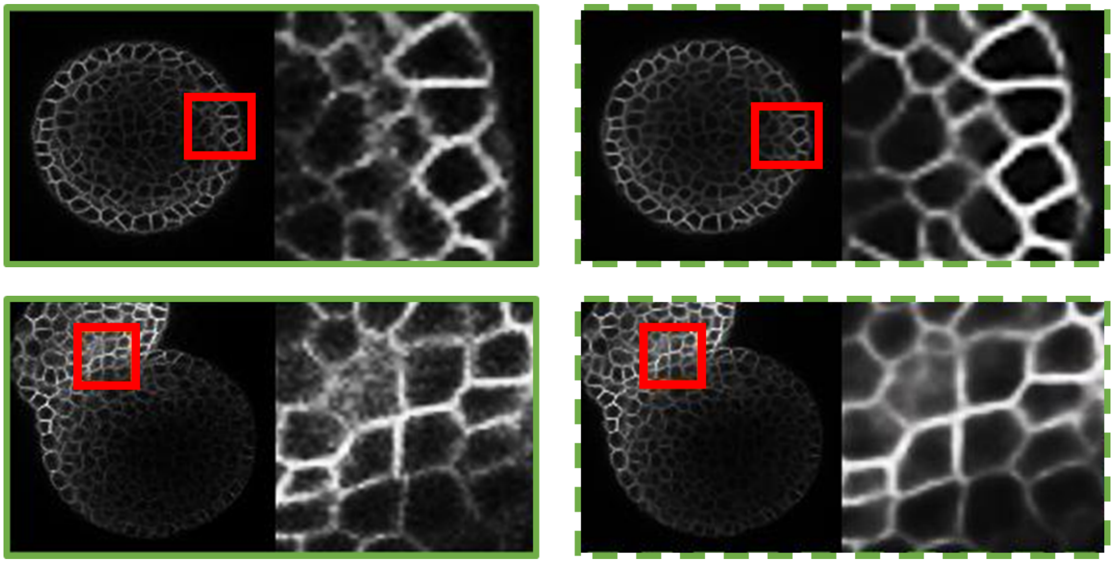


Figure 5.3: **Qualitative Results of Proposed Method.** Original image (*left*) and Reconstructed image (*right*) with corresponding zoomed parts are presented here. The proposed method is able to generate sharper images from the given blurry image slices.

posed method is able to generate sharper cell boundaries. Since we learn the quantization dictionary using all the slices in various $z-$stacks, our method is able to generate cleaner images. Deconvolution is a standard technique used by many researchers to enhance microscopy images. We compare our proposed approach with deconvolution operation used to denoise microscopy images using ImageJ [17]. It is performed on 2D slices using Gaussian Point Spread Function (PSF) with standard values. It can be seen from Fig. 5.5 that our proposed approach reconstructs visually enhanced slices compared to deconvolution operation in ImageJ. A key reason that deconvolution doesn't work well is due to the selection of PSF which highly depends on the capturing instrument. This demonstrates the advantage of our approach with respect to existing algorithms. Note that in Fig. 5.3, Fig. 5.4, and Fig. 5.5, input slice is shown inside solid green bounding box and the reconstructed slice using the proposed approach is shown inside dotted green bounding box.

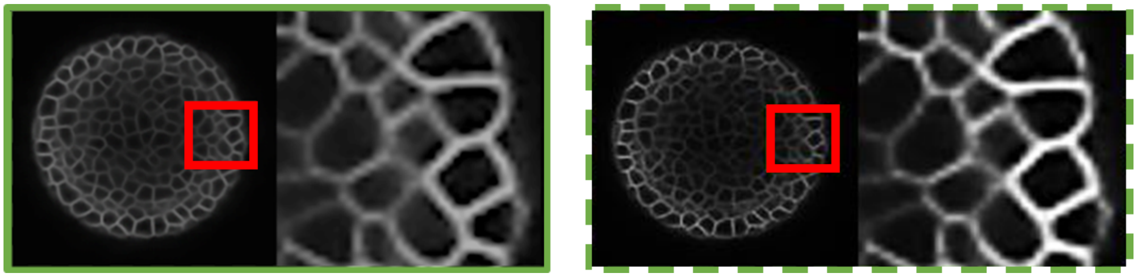


Figure 5.4: **Reconstruction Results without Quantization.** Reconstructed image without quantization (*left*) and Reconstructed image (*right*) with quantization with corresponding zoomed parts are presented here. This demonstrates that the quantization module in our proposed approach is effective in deblurring the data.

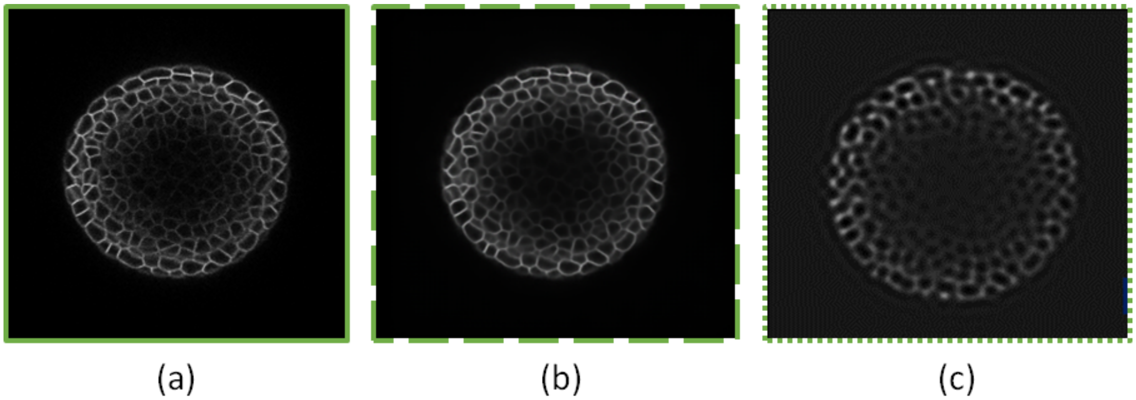


Figure 5.5: **Comparison of Reconstructed Results with ImageJ** [17].**(a)** Original Image, **(b)** Reconstructed using DQLR (**ours**) and **(c)** Reconstructed using deconvolution by ImageJ.

Qualitative Ablation. To evaluate the impact of quantization in the latent space, we perform an experiment without applying quantization keeping all other parameters same in the proposed method. Fig. 5.4 qualitatively shows the contribution on quantization in latent space. The image generated without quantization is less sharp than with quantization. This is due to inherent property of mean square loss to produce blurry results which dominates the reconstruction in absence of latent representation quantization loss.

5.5 Conclusion

Micro-imaging data collected for various bio-medical research suffers from inherent blurriness and using this data for further analysis is a challenging task. We present an approach for enhanced reconstruction of microscopic sequential data by leveraging the information from consecutive image slices and using quantization of their latent representation to alleviate blurriness. Our data driven approach demonstrates visually superior results on a publicly available benchmark. The proposed approach would be useful for bio-medical researchers to enhance images where data is scarce and consequently, avoid unwanted laborious efforts for re-imaging the data.

Chapter 6

Conclusions

6.1 Thesis Summary

In this thesis, we focused on the problem of recovering high-quality videos from input videos that are affected by different degradation models. We explored different attention mechanisms to extract internal structure from neighbouring frames and external information available from external datasets for various video enhancement tasks. Unlike other state-of-the-art methods, we do not make unrealistic assumption on image formation model. We address the video enhancement problem in a blind input prior settings such as deblurring problem with no assumption that all input frames are blurry, video super-resolution problem with unknown degradation kernel and recovering the global shutter image from a rolling shutter camera.

In Chapter 2, we addressed the problem of generating high frame-rate sharp videos, with no knowledge that either an input frame is blurry or not, for the task of deblurring and interpolation. Our proposed framework employs self-attention and cross-attention mechanisms in the latent representations to extract local information within the frame and additional information available in neighbouring frames for the task of joint video deblurring and interpolation.

In Chapter 3, we presented a novel meta-learning framework for blind spatio-temporal super-resolution where degradation kernel is not known. We leveraged meta-training using an external dataset to learn a model that can easily adapt to unseen degradation models. Furthermore, we exploit the internal structure of the test video to adapt the model, trained using external learning, specific to the given video. We demonstrated that the proposed approach is able to achieve superior enhancement while adapting to unknown degradation models and improved computational time as shown in our experiments.

In Chapter 4, we exploited the patch-recurrence property in frames to recover high-resolution global shutter frames from low-resolution rolling shutter video. Our proposed approach employs patch-level attention to compute correlation maps between neighbouring patches to extract information relevant for rolling shutter correction and super-resolution. Extensive experiments on different benchmark datasets demonstrate efficacy of our proposed approaches over various baselines and state-of-the-art approaches.

Finally in Chapter 5, we extend the process of quantization in the feature space and show that quantization, that is usually utilized to denoise any image in pixel space, can also be applied in the internal feature space for the task of unsupervised denoising. Our data driven approach demonstrates visually superior results on a publicly available benchmark. The proposed approach would be useful for bio-medical researchers to enhance images where data is scarce and consequently, avoid unwanted laborious efforts for re-imaging the data.

6.2 Future Research Directions

6.2.1 Learning Video compression through Super-Resolution

We address the problem of spatio-temporal video super-resolution and spatial super-resolution along with rolling shutter correction in Chapters 3 and 4, respectively. In terms of storage space, super-resolution problem can also be presented as video compression. In this case, the target is to learn a compression model to generate a low-resolution video and then learn a super-resolution model. Concepts of patch-recurrence, correlation and interpolation are explored in Chapters 3 and 4, which can also be leveraged for video compression. One simple baseline solution is to use the patch attention presented in Chapter 4 to learn a model to down-sample the image instead of super-resolution. With growing need of large-scale datasets for computer vision application that requires huge storage memory, video compression is an interesting future direction of our work.

6.2.2 Video Stabilization using Motion Estimation

Video stabilization requires accurate estimation of the displacement field of the camera. We have shown that attention mechanism can model motion for the task of deblurring and interpolation in Chapter 2 and for the task of rolling shutter correction in Chapter 4. However, existing works do not leverage the attention mechanism to estimate displacement field for video stabilization. Our proposed approach tries to learn the motion in the latent space. Hence, our approaches can help learn dataset specific motion cues, which can be utilized along with global and local attentions mechanism, to learn a motion to rectify shaky video.

6.2.3 Generation of 3D Dynamic Scenes using Videos

Virtual reality (VR) has huge potential for technological innovation. However, realistic content creation for VR devices is a challenging task which requires capturing the 3D view of the scene. With the abundance of videos available online, combination of our video interpolation technique, presented in Chapter 2, and a 3D reconstruction model can help generate high-resolution 3D dynamic scenes. Furthermore, our approaches for video super-resolution and deblurring can also be extended for 3D scene dataset to improve the user experience. With numerous applications of virtual reality in computer vision, developing algorithms for VR content creation can be a very interesting and challenging future research direction.

Bibliography

[1] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M Salman Asif, and Amit K Roy-Chowdhury. Non-adversarial video synthesis with learned priors. *arXiv preprint arXiv:2003.09565*, 2020.

[2] Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, and Konrad Schindler. From two rolling shutters to one global shutter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2513, 2020.

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

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

[5] 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*, 2019.

[6] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion deblurring and super-resolution from an image sequence. In *European conference on computer vision*, pages 571–582. Springer, 1996.

[7] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint arXiv:1909.06581*, 2019.

[8] Gordon J Berman. Measuring behavior across scales. *BMC biology*, 16(1):1–11, 2018.

[9] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017.

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4778–4787, 2017.

[11] Stephen C Cain, Majeed M Hayat, and Ernest E Armstrong. Projection-based image registration in the presence of fixed-pattern noise. *IEEE transactions on image processing*, 10(12):1860–1872, 2001.

[12] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021.

[13] Yuning Chai. Patchwork: A patch-wise attention network for efficient object detection and segmentation in video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3415–3424, 2019.

[14] Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pages 168–172. IEEE, 1994.

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

[16] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020.

[17] Tony J Collins. ImageJ for microscopy. *Biotechniques*, 43(S1):S25–S30, 2007.

[18] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[19] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[20] 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*, 38(2):295–307, 2015.

[21] Mehran Ebrahimi and Edward R Vrscay. Solving the inverse problem of image zooming using "self-examples". In *International Conference Image Analysis and Recognition*, pages 117–130. Springer, 2007.

[22] Mihail Eric and Christopher D Manning. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*, 2017.

[23] Sina Farsiu, M Dirk Robinson, Michael Elad, and Peyman Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.

[24] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[25] Chelsea Finn and Sergey Levine. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. *arXiv preprint arXiv:1710.11622*, 2017.

[26] Per-Erik Forssén and Erik Ringaby. Rectifying rolling shutter video from hand-held devices. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 507–514. IEEE, 2010.

[27] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[28] Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1604–1613, 2019.

[29] Yunhui Guo, Noel CF Codella, Leonid Karlinsky, John R Smith, Tajana Rosing, and Rogerio Feris. A new benchmark for evaluation of cross-domain few-shot learning. *arXiv preprint arXiv:1912.07200*, 2019.

[30] Abhishek Gupta, Benjamin Eysenbach, Chelsea Finn, and Sergey Levine. Unsupervised meta-learning for reinforcement learning. *arXiv preprint arXiv:1806.04640*, 2018.

[31] Akash Gupta, Abhishek Aich, and Amit K Roy-Chowdhury. Alanet: Adaptive latent attention network for joint video deblurring and interpolation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 256–264, 2020.

[32] Akash Gupta, Padmaja Jonnalagedda, Bir Bhanu, and Amit K Roy-Chowdhury. Ada-vsr: Adaptive video super-resolution with meta-learning. *arXiv preprint arXiv:2108.02832*, 2021.

[33] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1664–1673, 2018.

[34] Yedid Hoshen, Ke Li, and Jitendra Malik. Non-adversarial image synthesis with generative latent nearest neighbors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5811–5819, 2019.

[35] Yan Huang, Wei Wang, and Liang Wang. Bidirectional recurrent convolutional networks for multi-frame super-resolution. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 235–243, 2015.

[36] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2024–2032, 2019.

[37] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

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

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

[40] Posted in Video Animation, Feb 2018. Accessed: 2021-04-04.

[41] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.

[42] James Janesick, Jeff Pinter, Robert Potter, Tom Elliott, James Andrews, John Tower, John Cheng, and Jeanne Bishop. Fundamental performance differences between cmos and ccd imagers: part iii. In *Astronomical and Space Optical Systems*, volume 7439, page 743907. International Society for Optics and Photonics, 2009.

[43] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018.

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

[45] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3224–3232, 2018.

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

[47] Armin Kappeler, Seunghwan Yoo, Qiqin Dai, and Aggelos K Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.

[48] Alexandre Karpenko, David Jacobs, Jongmin Baek, and Marc Levoy. Digital video stabilization and rolling shutter correction using gyroscopes. *CSTR*, 1(2):13, 2011.

[49] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016.

[50] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3116–3125, 2019.

[51] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Fisr: deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11278–11286, 2020.

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

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

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

[55] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[56] Suyoung Lee, Myungsub Choi, and Kyoung Mu Lee. Dynavsr: Dynamic adaptive blind video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2093–2102, 2021.

[57] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.

[58] Chia-Kai Liang, Li-Wen Chang, and Homer H Chen. Analysis and compensation of rolling shutter effect. *IEEE Transactions on Image Processing*, 17(8):1323–1330, 2008.

[59] Renjie Liao, Xin Tao, Ruiyu Li, Ziyang Ma, and Jiaya Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–539, 2015.

[60] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

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

[62] Min Liu, Anirban Chakraborty, et al. Adaptive cell segmentation and tracking for volumetric confocal microscopy images of a developing plant meristem. *Molecular Plant*, 4(5):922–931, 2011.

[63] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5949, 2020.

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

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

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

[67] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. In *Adversarial Autoencoders*, volume abs/1511.05644, 2015.

[68] Peter N McMahon-Crabtree and David G Monet. Commercial-off-the-shelf event-based cameras for space surveillance applications. *Applied Optics*, 60(25):G144–G153, 2021.

[69] Tomer Michaeli and Michal Irani. Nonparametric blind super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 945–952, 2013.

[70] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *arXiv preprint arXiv:1707.03141*, 2017.

[71] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[72] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *arXiv preprint arXiv:1803.11347*, 2018.

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

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

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

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

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

[78] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

[79] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1628–1636, 2016.

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

[81] Abhijith Punnappurath, Vijay Rengarajan, and AN Rajagopalan. Rolling shutter super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 558–566, 2015.

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

[83] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[84] Vijay Rengarajan, Yogesh Balaji, and AN Rajagopalan. Unrolling the shutter: Cnn to correct motion distortions. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 2291–2299, 2017.

[85] Vijay Rengarajan, Abhijith Punnappurath, AN Rajagopalan, and Gunasekaran Seetharaman. Rolling shutter super-resolution in burst mode. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2807–2811. IEEE, 2016.

[86] Vijay Rengarajan, Ambasamudram N Rajagopalan, and Rangarajan Aravind. From bows to arrows: Rolling shutter rectification of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2773–2781, 2016.

[87] 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*, pages 234–241. Springer, 2015.

[88] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

[89] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[90] Nicolas Schweighofer and Kenji Doya. Meta-learning in reinforcement learning. *Neural Networks*, 16(1):5–9, 2003.

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

[92] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation, 2020.

[93] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.

[94] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018.

[95] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[96] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3516–3525, 2020.

[97] Sanghyun Son, Suyoung Lee, Seungjun Nah, Radu Timofte, and Kyoung Mu Lee. Ntire 2021 challenge on video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 166–181, 2021.

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

[99] Susrutha Babu Sukhavasi and Suparshya Babu Sukhavasi. Role of cmos image sensors based surveillance systems in demanding fields. *Sensors*, 2021.

[100] Susrutha Babu Sukhavasi, Suparshya Babu Sukhavasi, Khaled Elleithy, Shakour Abuzneid, and Abdelrahman Elleithy. Cmos image sensors in surveillance system applications. *Sensors*, 21(2):488, 2021.

[101] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019.

[102] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[103] Jinshan Tang, Eli Peli, and Scott Acton. Image enhancement using a contrast measure in the compressed domain. *IEEE Signal Processing Letters*, 10(10):289–292, 2003.

[104] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4472–4480, 2017.

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

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

[107] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018.

[108] Aaron van den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[109] Subeesh Vasu, AN Rajagopalan, et al. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–645, 2018.

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

[111] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

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

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

[114] Lisa Willis et al. Cell size and growth regulation in the arabidopsis thaliana apical stem cell niche. *Proceedings of the National Academy of Sciences*, pages E8238–E8246, 2016.

[115] Junru Wu, Xiang Yu, Ding Liu, Manmohan Chandraker, and Zhangyang Wang. David: Dual-attentional video deblurring. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2376–2385, 2020.

[116] Yan Wu, Mihaela Rosca, and Timothy Lillicrap. Deep compressed sensing. In *International Conference on Machine Learning*, pages 6850–6860, 2019.

[117] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020.

[118] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. Space-time video super-resolution using temporal profiles. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 664–672, 2020.

[119] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019.

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

[121] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3262–3271, 2018.

[122] Ke Zhang, Cankun Yang, Xiaojuan Li, Chunping Zhou, and Ruofei Zhong. High-efficiency microsatellite-using super-resolution algorithm based on the multi-modality super-cmos sensor. *Sensors*, 20(14):4019, 2020.

[123] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

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

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

[126] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2016.

[127] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021.

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

[129] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 948–956, 2017.

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

[131] Liad Pollak Zuckerman, Eyal Naor, George Pisha, Shai Bagon, and Michal Irani. Across scales and across dimensions: Temporal super-resolution using deep internal learning. In *European Conference on Computer Vision*, pages 52–68. Springer, 2020.