Joint Video Rolling Shutter Correction and Super-Resolution

Akash Gupta
Vimaan
akash.gupta@vimaan.ai

Sudhir Kumar Singh
Vimaan
sudhir.singh@vimaan.ai

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

Abstract

*With the prevalence of CMOS cameras in many computer vision applications, there is an increase in the 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 frame for RS correction and super-resolution, simultaneously. Different from existing works, we address a more realistic problem of joint rolling shutter correction and super-resolution (RS-SR). We introduce a novel architecture, deformable Patch Attention Network (**PatchNet**), that utilizes patch-recurrence property along with deformable receptive fields to learn the global and local structure of the video. Specifically, **PatchNet** leverages bi-directional motion field in the feature space to extract relevant information from neighboring patches using attention mechanism, and deformable fields using deformable convolutions to extract local pixel-level information for joint rolling shutter correction and super-resolution. Our work is the first to tackle the task of RS correction and super-resolution on the recently released BS-RSCD dataset. Experiments on the BS-RSCD and FastecRS datasets demonstrate that our model performs favorably against various state-of-the-art approaches. Project details are available at* `https://akashagupta.com/publication/wacv23_patchnet/project.html`

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 [18]. 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

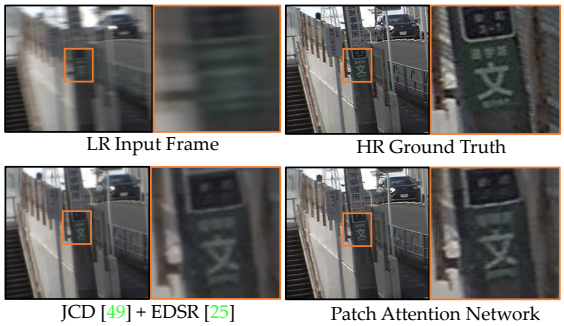


Figure 1: **Patch Attention Network.** We show frame on the left and zoomed in patch on the right. **Top Left:** Input LR rolling shutter frame (resized to HR frame resolution). **Top Right:** Global shutter ground truth HR frame. **Bottom Left:** Output of combination of state-of-the-art RSC method (JCD [52]) and SR method (EDSR [28]). **Bottom Right:** Predicted image by Patch Attention Network. It can be observed that the **PatchNet** model generates superior results as compared to the cascade approach using state-of-the-art rolling shutter correction and super-resolution method.

times, and the movement of the device [20, 33] 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 or object 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 [49]. With increasing popularity of CMOS cameras in various computer vision applications [40, 31, 39, 14], which require high-quality high-resolution imaging, it calls for jointly addressing the task of rolling shutter rectification and spatial super-resolution.

Early works [35, 36] studied the problem of multi-image super-resolution for images captured from rolling shutter cameras. In [35], authors assume that one of the images is 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 ap-

proach to recover high-resolution (HR) image when all the images, captured using burst mode, have rolling shutter artifacts is presented by the authors in [36]. However, these multi-image based approaches rely on geometric constraints from multiple views formulating a computationally expensive optimization problem for 6 DoF camera motions.

Availability of large-scale datasets [38, 19, 47, 29, 41, 1] have greatly facilitated the research in learning based video restoration techniques. Existing video super-resolution (VSR) approaches [23, 21, 47, 42, 4, 44] 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 [52, 30]. Authors in [30] 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 rolling shutter (RS) distorted image and its corresponding global shutter image.

Addressing this issue, a realistic dataset for rolling shutter correction and deblurring (RSCD) was proposed in [52] 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 video enhancement problem such as joint rolling shutter correction and super-resolution. Authors in [52] 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 does not take into account the global information available in neighbouring patches.

Motivation. We introduce a novel architecture Patch Attention Network (**PatchNet**) to utilize global as well as local information to jointly rectify rolling shutter (RS) artifacts and generate high-resolution frames from a low-resolution video acquired using RS camera. Specifically, we leverage the patch recurrence property in the feature space to exploit the information available in neighbouring patches 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 frames [32, 11, 54, 22] 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 [9]. Convolution layers have a fixed kernel size so they cannot leverage the information beyond their receptive field [9]. Unlike convolutional layers, which cannot extract the information beyond their receptive field [9], Patch Attention Network relies on deformable convolutions and attention mechanism to extract pixel-level and patch-level information to generate high-resolution global shutter frames. The attention mechanism utilizes bi-directional motion field to extract correlated information from neighbouring patches. Since Patch Attention Network is jointly learning for rolling shutter correction and super-resolution, with the help of deformable fields and correlated neighbouring patches, it is able to generate superior results as shown in Figure 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 the feature space.
- This is the first work to leverage the combination of local information, using deformable convolution, and motion field driven global patch-level information from neighbouring patches to recover a high-resolution GS video.
- Our framework demonstrates consistently effective results on two datasets, BS-RSCD [52] and FastecRS [30] 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.

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 vision tasks. We show characteristic comparison of our approach against prior works in RS correction and super-resolution Table 1.

Video Super-Resolution. Several learning-based approaches have been proposed for video super-resolution [23, 21, 47, 16] for video with no rolling shutter distortions. A deep learning based approach is presented in [23], 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 [27]. The authors in [42, 4] 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 [47]. DUF [21] overcomes the problem of estimating accurate optical flow by implicit motion compen-

sation 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 [44] 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 [26]. Another approach [13] 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 [13] using inertial measurement is proposed in [24]. 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 [37] 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 [53] 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 [43]. They model 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 effect and recover GS image [2].

More recently, an end-to-end deep learning approach for rolling shutter correction is presented in Deep Unrolling Network [30] 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 [52]. 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 [52] (JCD) utilizes bi-directional optical flow as compared to Deep Unrolling Network [30]. Additionally, JCD leverages deformable convolution for hierarchical features for task of joint rolling shutter correction and deblurring. Deformable convolution [9] 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, 6], image generation [48] and image super-resolution [51]. Recently, different attention models are proposed for video deblurring [46], video super-resolution [15] and video interpolation [8]. In [8], attention is applied channel-wise on concatenated down-shuffled frames for video interpolation. Authors in [15] explore attention in latent space for the task of video deblurring and interpolation. A patch-wise attention network (Patchwork) is presented in [6] 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 [5] 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

Table 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 [30]	✓	✗	✗	✗
VSR-T [5]	✗	✓	✗	✓
JCD [52]	✓	✗	✓	✗
PatchNet	✓	✓	✓	✓

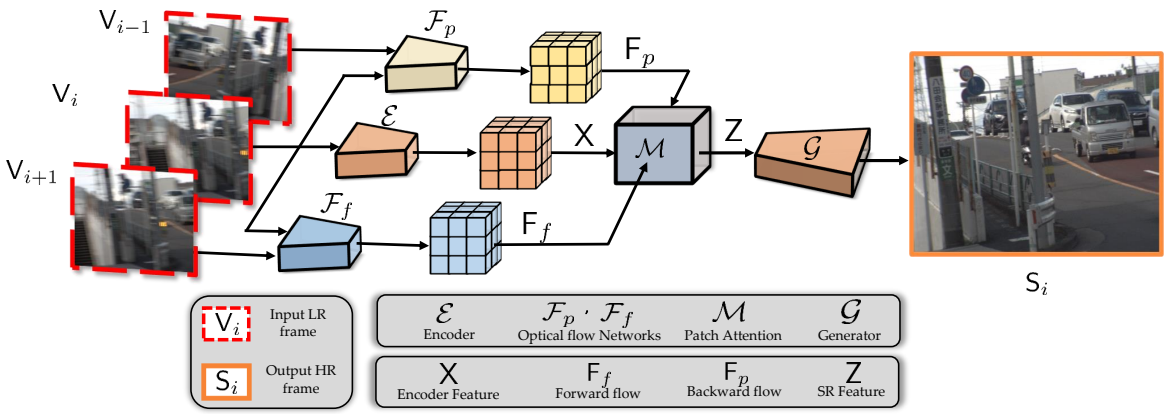


Figure 2: **Overview of the proposed approach.** Given a low-resolution input video frames V_{i-1}, V_i and V_{i+1}, we extract the feature representation X corresponding to frame V_i using the encoder network $\mathcal{E}$ and the flow features F_p and F_f with respect to the past frame V_{i-1} and 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. 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.

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 in addition to video super-resolution. Our approach specifically performs super-resolution first in the feature space and then employs generator model to obtain HR-GS video.

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 due to the fixed kernel shape.

Notations. Let the low-resolution rolling shutter video be denoted by $\mathbf{V}_{LR} = \left[\mathsf{V}_1, \mathsf{V}_2, \cdots, \mathsf{V}_N\right]$, 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 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 $\mathsf{X} \in \mathbb{R}^{H \times W \times C}$. 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.

We aim to generate a high-resolution global shutter video denoted by $\mathbf{V}_{HR} = \left[\mathsf{S}_1, \mathsf{S}_2, \cdots, \mathsf{S}_L\right]$, where $\mathsf{S}_t \in \mathbb{R}^{aH_I \times aW_I \times 3}$ using the Patch Attention Network $\mathcal{M}$ as shown in Figure 2. 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 encoder features patches by utilizing motion fields from forward flow and backward flow patches. 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 [12]. 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 backward flow acts as the key representation ($\mathcal{K}$) for different patch values ($\mathcal{V}$) of encoder features, and utilize the forward 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 utilized by the reconstruction model $\mathcal{G}$ to generate HR global shutter video.

3.1. Feature 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

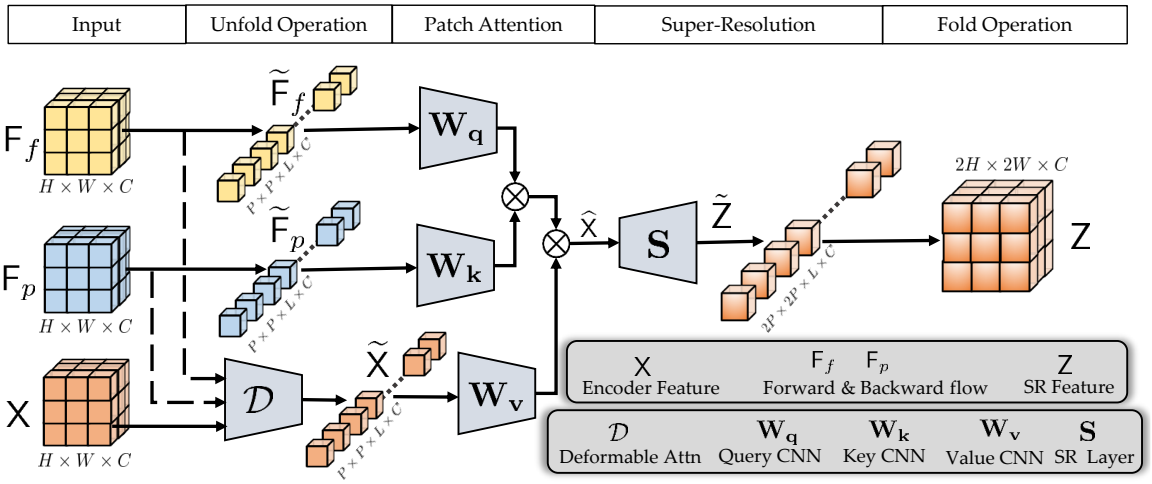


Figure 3: **Overview of Patch Attention Network.** Given the encoder feature X and the motion features F_p and F_f, we first utilize the deformable attention network $\mathcal{D}$ [52] to incorporate motion information at pixel-level and unfold it into $P \times P$ patches to obtain the patch-level encoder 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 flow features $\widetilde{\mathsf{F}}_p$ and the patch-level 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 using generator $\mathcal{G}$ as shown in Figure 2.

($\mathcal{F}_p$) 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.

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

$$\mathsf{F}_p = \mathcal{F}_p(\mathsf{V}_i, \mathsf{V}_{i-1}) \tag{2}$$

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.

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 $\mathbf{S}$ is utilized to produce high-resolution features. Figure 3 presents the overview of the patch attention used in the **PatchNet**. First, we employ a deformable convolution attention module $\mathcal{D}$ 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 3.

$$\widetilde{\mathsf{X}} = \mathcal{D}(\mathsf{X}, \mathsf{F}_p, \mathsf{F}_f) \tag{3}$$

Similar to the unfolding operation in module $\mathcal{D}$, we also divide bi-directional flow features in $P \times P$ patches. The patch-level feature representations 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.

$$\mathcal{Q} = \mathbf{W}_q\left(\widetilde{\mathsf{F}}_f\right), \quad \mathcal{K} = \mathbf{W}_k\left(\widetilde{\mathsf{F}}_p\right), \quad \mathcal{V} = \mathbf{W}_v\left(\widetilde{\mathsf{X}}\right) \tag{4}$$

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.

$$\widehat{\mathsf{X}} = \sigma\big(\mathcal{Q}^T\mathcal{K}\big)\mathcal{V} \tag{5}$$

We then utilize a super-resolution layer $\mathbf{S}$ to obtain high-resolution patch-representation $\widetilde{\mathsf{Z}}$ using the equation 6.

$$\widetilde{\mathsf{Z}} = \mathbf{S}\Big(\widehat{\mathsf{X}}\Big) = \mathbf{S}\big(\sigma\big(\mathcal{Q}^T\mathcal{K}\big)\mathcal{V}\big) \tag{6}$$

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 S_i corresponding to the low-resolution rolling shutter frame V_i

3.3. GS-HR Video Generation

Our task is to recover the global shutter frame and perform super-resolution from a given low-resolution rolling shutter frame V_i. To this end, we employ a generative neural network $\mathcal{G}$ that transforms the high-resolution features obtained from the **PatchNet** to high-resolution global shutter frame. The aggregated reconstruction features Z is a high-resolution features obtained using **PatchNet** which are then utilized to generate high-resolution global shutter frame (S_i) corresponding to V_i using the generator model $\mathcal{G}$, such that $\mathsf{S}_i = \mathcal{G}(\mathsf{Z})$.

3.4. Loss Function

Our objective function is composed of Charbonnier loss ($\mathcal{L}_c$) [7] 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:

$$\mathcal{L} = \lambda_c\mathcal{L}_c + \lambda_p\mathcal{L}_p + \lambda_v\mathcal{L}_v \tag{7}$$

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

4. Experimental Setup and Results

In this section, we first introduce the benchmark datasets, evaluation metrics and provide implementation details. Then extensive experiments are shown to demonstrate the effectiveness of our proposed approach in recovering high resolution global shutter videos from low-resolution rolling shutter frames.

4.1. Datasets and Evaluation Metrics

We evaluate the performance of our approach using publicly available BS-RSCD [52] and synthetic Fastec-RS [30] datasets which have been used in prior rolling shutter correction works.

BS-RSCD Dataset. BS-RSCD [52] 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 RS frames for training in all our experiments.

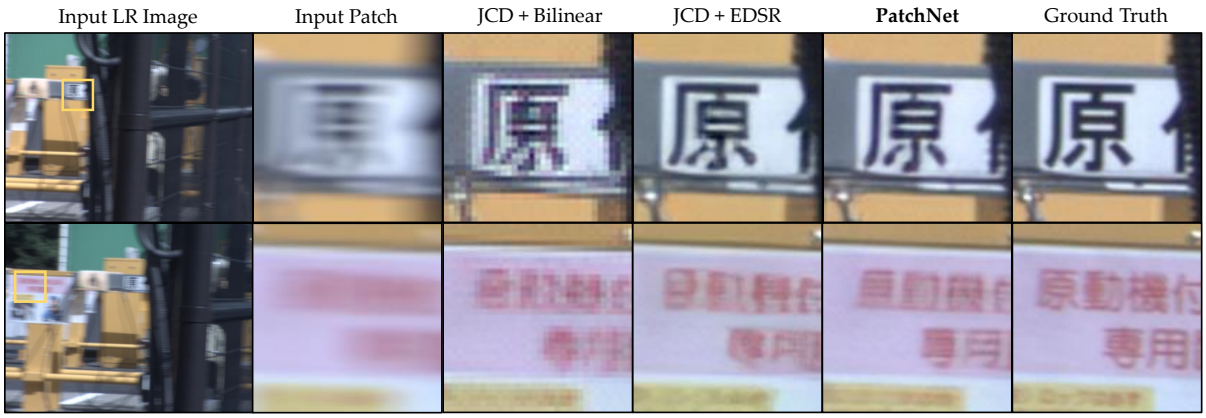
Fastec-RS Dataset. The Fastec-RS dataset [30] is a synthetic dataset captured using a professional Fastec TS51 high speed global shutter camera. Total of 76 image sequence are captured at 2400 fps with a resolution of 640 × 480 pixels in mainly urban environment. Each sequence synthesizes 34 rolling 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) [17], Structural Similarity Index Measure (SSIM) [45] and Learned Perceptual Metric (LPIPS) [50].

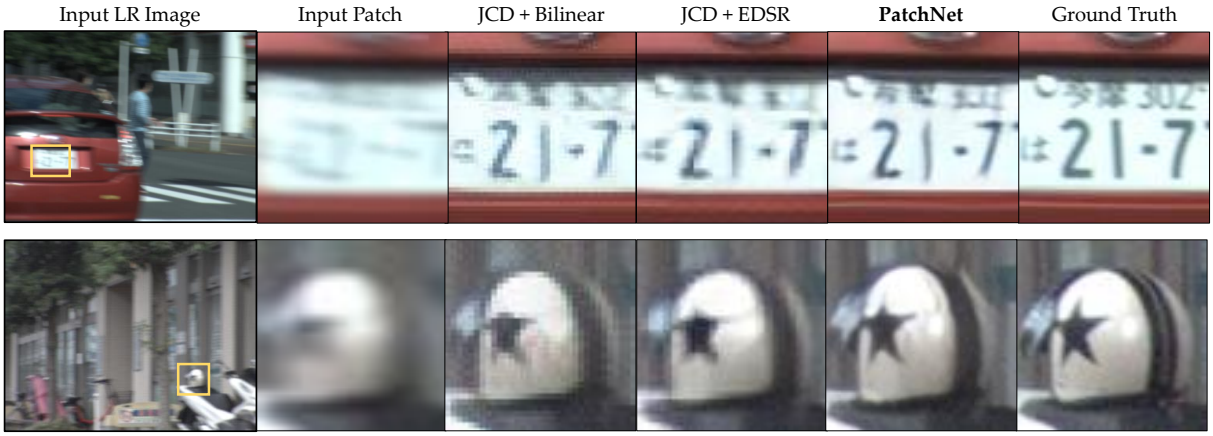
Implementation Details. Our framework is implemented in PyTorch [34]. All the experiments are trained for 400 epochs with a batch size of 8. We use ADAM [25] optimizer with initial learning rate of 0.0001 with cosine annealing scheduler. The loss weights λ_c, λ_p and λ_v are set to 10, 1 and 0.1 , respectively. The deformable convolution attention layer is adopted from JCD approach [52] with deformable groups as 8. For details on network architecture, please refer the supplemental material.

4.2. Qualitative Results

We compare our work with combination of state-of-the-art rolling shutter correction (JCD [52]) and super-resolution works such as bi-linear interpolation and EDSR [28] . Figure 4(a) and Figure 4(b) 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



(a) First column consists of two consecutive 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 [52] + bi-linear interpolation, JCD [52] + EDSR [28] and **PatchNet** performs better as they utilize learnable module for super-resolution. **PatchNet** produces visually sharper results as it can extract available information from neighbouring patches as opposed to JCD [52] + EDSR [28].



(b) 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 [52] + bi-linear interpolation, JCD [52] + EDSR [28] and **PatchNet** performs better as it utilizes patch-recurrence property along with deformable convolution. **PatchNet** produces visually sharper results as it learns RS correction and SR jointly.

Figure 4: Qualitative results on BS-RSCD. Consecutive frames of a video is shown in (a) and two different videos in (b).

cannot learn the pixel displacement for super-resolution task. From Figure 4(a), it can be observed that our approach is able to produce sharp frames with fine details in text using consecutive frames of a video. 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(b). 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 high quality visual frames. Please refer to supplemental materials for some video examples generated using **PatchNet**.

4.3. 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 these baselines are shown in Table 2.

• **BS-RSCD dataset.** 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

Table 2: **Quantitative results comparison of PatchNet with the state-of-the-art baselines.** We compare our approach with various combination of RSC model followed by an SR model. We demonstrate that **PatchNet** is able to generate HR global frames, with LR input, compared to approaches which only perform RSC with HR input (highlighted in red).

Methods		BS-RSCD			FastecRS		
RSC	**SR**	**PSNR** ↑	**SSIM** ↑	**LPIPS**↓	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
JCD [52] with LR Input	Bi-linear Interpolation	22.74	0.581	0.463	23.87	0.655	0.339
	Transposed Convolution	24.15	0.628	0.328	24.12	0.632	0.262
	SAN [10]	24.37	0.633	0.305	24.07	0.643	0.281
	EDSR [28]	24.94	0.650	0.263	24.67	0.713	0.187
DUN [30] with LR Input	Bi-linear Interpolation	21.64	0.552	0.489	25.34	0.792	0.185
	Transposed Convolution	24.02	0.602	0.342	25.88	0.801	0.179
	SAN [10]	24.16	0.621	0.322	26.10	0.807	0.165
	EDSR [28]	24.58	0.634	0.286	26.43	0.810	0.147
JCD [52] with HR Input		<u>26.42</u>	<u>0.757</u>	**0.122**	24.84	0.778	<u>0.107</u>
DUN [30] with HR Input		<u>25.14</u>	0.729	0.159	<u>27.00</u>	**0.825**	0.108
PatchNet with LR Input		**27.38**	**0.793**	<u>0.144</u>	**27.12**	<u>0.811</u>	**0.103**

Table 3: **Impact of patch-size on performance of PatchNet on BS-RSCD dataset.** It can be observed that the performance of the proposed **PatchNet** improves with increase in patch-size with best results for 8×8 patch size.

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**

of RS correction and super-resolution approaches (JCD + EDSR). 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.

• **Fastec-RS dataset.** 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 [30] 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.

4.4. Ablation Analysis

For the ablation, we study on impact of patch-size on performance of **PatchNet** and present our findings in Table 3 using BS-RSCD dataset. 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 ($\sim$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 even with patch size of 4×4 without significant drop in the performance.

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 global information from neighbouring patches using the key-query similarity and local information using deformable convolution. Specifically, the patch attention module obtains correlation maps between neighbouring patches for *simultaneous* rolling shutter correction and super-resolution. Our main contribution over existing approaches is in learning the rolling shutter correction and super-resolution jointly, 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.

Limitations and Future work. Until recently, the research in RS correction and Super-Resolution was limited due to unavailability of a realistic supervised dataset. With the BS-RSCD [52] dataset, where images are captured using specific camera configuration, our model is suitable for images captured in similar camera settings. One direction would be to develop models agnostic to camera configuration by incorporating specific settings, such as readout and exposure time, while training the model. Incorporating camera configuration requires additional supervision, along with data acquisition, and calls for methods to address this problem in semi-supervised, unsupervised or self-supervised setups.

References

[1] Abhishek Aich, Akash Gupta, Rameswar Panda, Rakib Hyder, M. Salman Asif, and Amit K. Roy-Chowdhury. Non-adversarial video synthesis with learned priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 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] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] Akash Gupta, Abhishek Aich, Kevin Rodriguez, G. Venugopala Reddy, and Amit K. Roy-Chowdhury. Deep Quantized Representation For Enhanced Reconstruction. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*, pages 1–4. IEEE, Apr. 2020.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Padmaja Jonnalagedda, Daniel Schmolze, and Bir Bhanu. [regular paper] mvpnets: Multi-viewing path deep learning neural networks for magnification invariant diagnosis in breast cancer. In *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 189–194, 2018.

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] Calvin-Khang Ta, Abhishek Aich, Akash Gupta, and Amit K. Roy-Chowdhury. Poisson2Sparse: Self-Supervised Poisson Denoising From a Single Image. *arXiv*, June 2022.

[42] 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.

[43] 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.

[44] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[45] 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.

[46] 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.

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.