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

1. Network architectures

Overview The overall architecture of our Patch Attention Network (**PatchNet**) is shown in Figure 1. It is comprised of of five main modules: a feature encoder network, a forward flow network, a backward flow network, our proposed Patch Attention network and a generator network.

• **Encoder Network.** The encoder network $\mathcal{E}$ outputs features at three different levels. We utilize the output of third level (X) for our Patch Attention Network to obtain high-resolution feature patches. Each level is composed of a convolutional layer followed by a ReLU activation function and three ResNet blocks as in [4]. The first level features are extracted using a convolutional layer with a kernel size 7×7, 32 filters and uses a stride size 1. The convolutional layers of the second level and the third level feature extractors have a kernel size 3×3 and use a stride size 2 to down-sample the learned feature representations. The number of convolutional filters used in the second level and third level feature extractors are 64 and 128, respectively.

• **Forward Flow and Backward Flow Network.** Both the forward and backward flow networks consists of five DenseNet blocks [3] and a deconvolution layer to learn the bi-directional dense motion field using second level and third level feature maps as in [4]. We first pass the second and the third level feature from two different DenseNet blocks (consisting of 5 blocks) and utilize the deconvolutional layer to fuse the second level encoder feature with the output of DenseNet blocks and generate forward flow features $\mathcal{F}_f$ and backward flow features $\mathcal{F}_b$. Each DenseNet block consists of a convolutional layer followed by a ReLU activation function. The convolutional layer has a 3×3 kernel size and a stride size 1. The number of filters for the five DenseNet blocks are 128, 128, 96, 64 and 32 for the third level. Similarly, we have 64, 64, 48, 32 and 16 for the second level. The deconvolutional layers have a kernel size 4×4, a stride size 2, a padding size 1 and 2 filters.

• **Patch Attention Network.** The **PatchNet** consists of a deformable convolutional attention model $\mathcal{D}$ to fuse the motion information with encoder features, three independent convolutional neural networks to capture the inter and intra-patch recurrence property in each frame and a feature super-resolution layer $\mathbf{S}$ as shown in Figure 2. The deformable attention model $\mathcal{D}$ concatenates the encoder feature X, the forward motion feature F_f and the backward motion features F_p from different streams via a channel attention module [2], to assign channel-wise weight to the concatenated features, followed by a deformable convolutional layer [1].

Each of the Query $\mathbf{W_q}$, Key $\mathbf{W_k}$ and Value $\mathbf{W_v}$ CNN take 64 channel input feature and apply convolutional filters with kernel size of 3×3, stride size 1 and padding size 1. The number of output channels for each of these CNN is 256. The super-resolution layer $\mathbf{S}$ is a PixelShuffle layer with upscale factor 2.

• **Generator Network.** The generator model $\mathcal{G}$ consists of 20 residual blocks and each residual block is composed of convolutional layer with skip-connection followed by a Leaky ReLu activation. The convolutional layer takes 64 channel input from **PatchNet** and apply convolution operation with kernel size 3×3, stride size and padding size 1 to synthesis high resolution global shutter image.

2. Qualitative Results

Please refer to the attached video files for additional qualitative results. We show the input video and compare output of **PatchNet** with ground truth video.

References

[1] 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.

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[4] 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.

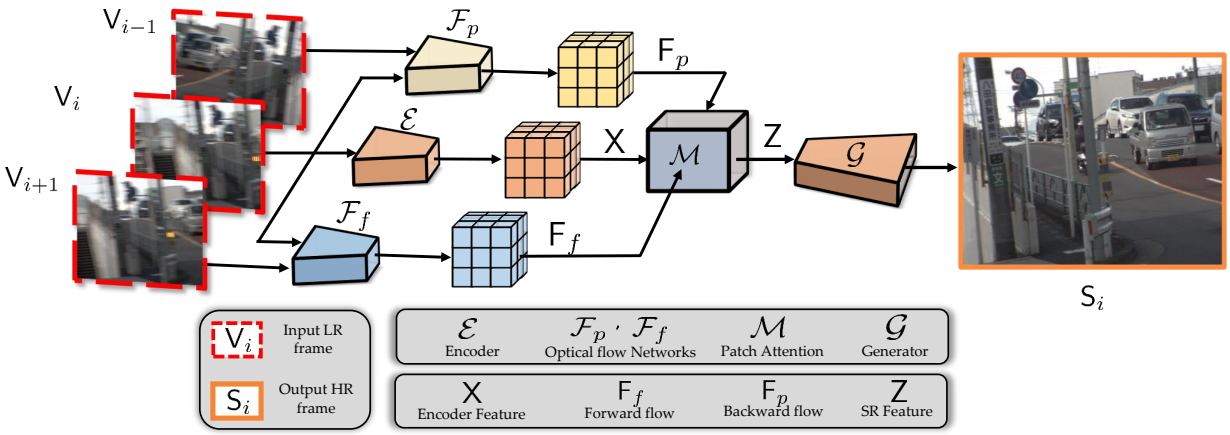

Figure 1: **Overview of the proposed approach.** Given a low-resolution input video frames $\mathsf{V}_{i-1}, \mathsf{V}_i$ and V_{i+1}, we extract the feature representation X corresponding to frame V_i using the encoder network $\mathcal{E}$ and the flow features F_p and F_f with respect to the past 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 in main paper). The high-resolution feature Z is then used by the decoder network $\mathcal{G}$ to produce high-resolution global shutter frames $\mathsf{S}_{i-1}, \mathsf{S}_i$ and S_{i+1}.

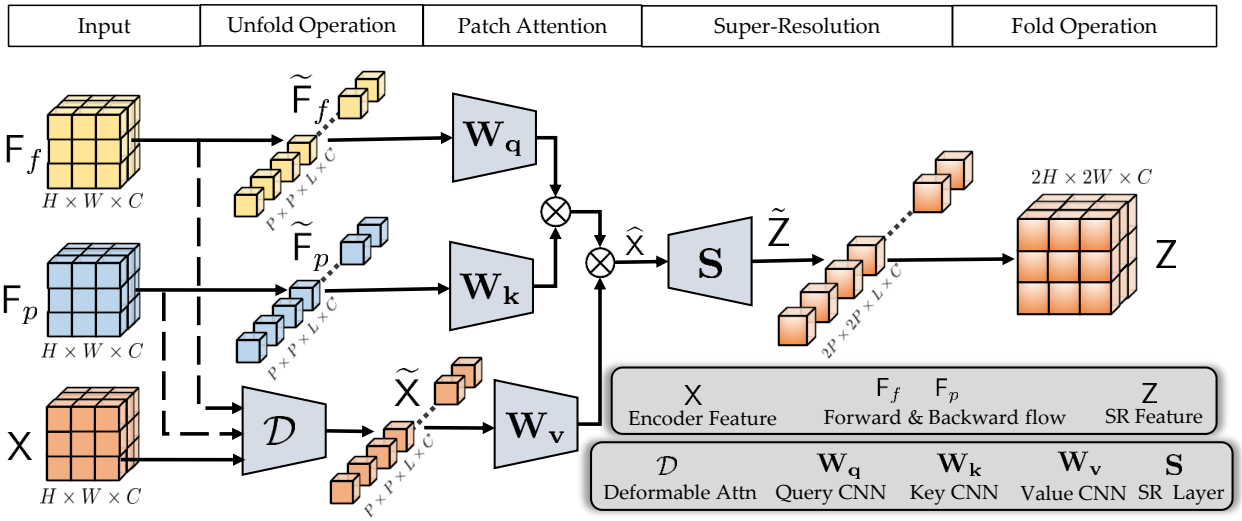

Figure 2: **Overview of Patch Attention Network.** Given the encoder feature X and the motion features F_p and F_f, we first utilize the deformable attention network $\mathcal{D}$ [5] to incorporate motion information at pixel-level and unfold it into $P \times P$ patches to obtain the patch-level encoder 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 1.

[5] 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.