Single Image Deblurring with Row-dependent Blur Magnitude

Xiang Ji[1] Zhixiang Wang[1,2] Shin'ichi Satoh[2,1] Yinqiang Zheng[1†]
[1]The University of Tokyo, Japan [2]National Institute of Informatics, Japan
{jixiang,wangzhixiang}@g.ecc.u-tokyo.ac.jp, satoh@nii.ac.jp, yqzheng@ai.u-tokyo.ac.jp

Abstract

Image degradation often occurs during fast camera or object movements, regardless of the exposure modes: global shutter (GS) or rolling shutter (RS). Since these two exposure modes give rise to intrinsically different degradations, two restoration threads have been explored separately, i.e. motion deblurring of GS images and distortion correction of RS images, both of which are challenging restoration tasks, especially in the presence of a single input image. In this paper, we explore a novel in-between exposure mode, called global reset release (GRR) shutter, which produces GS-like blur but with row-dependent blur magnitude. We take advantage of this unique characteristic of GRR to explore the latent frames within a single image and restore a clear counterpart by only relying on these latent contexts. Specifically, we propose a residual spatially-compensated and spectrally-enhanced Transformer (RSS-T) block for row-dependent deblurring of a single GRR image. Its hierarchical positional encoding compensates global positional context of windows and enables order-awareness of the local pixel's position, along with a novel feed-forward network that simultaneously uses spatial and spectral information for gaining mixed global context. Extensive experimental results demonstrate that our method outperforms the state-of-the-art GS deblurring and RS correction methods on single GRR input.

1. Introduction

As a fundamental research task of computer vision, image restoration has been explored for many years, aiming to recover a high-quality clean image from its corrupted counterpart by removing undesired degradation. Fast or even extreme motion, which is a major inducement to image degeneration and highly correlated to the camera's exposure mode: *global* shutter (GS) and *rolling* shutter (RS).

In a GS image sensor, all pixels are reset simultaneously and immediately charged with exposure, then stored for sequential readout (Figure 1a). In the presence of fast and sudden motion, blurring effects will happen. [25, 39] depict this motion-blurred output $\mathbf{I}^B$ as an average of N successive latent sharp frames $\mathbf{I}_i^G$:

$$\mathbf{I}^B = \frac{1}{N}\sum_{i=0}^{N-1}\mathbf{I}_i^G\,,\ \ i = 0, 1, \cdots, N-1\,.$$

In contrast, RS cameras expose image pixels row by row (including reset and readout), leading to an invariant time delay between consecutive scanlines (Figure 1b). Therefore, RS distortion effects, also known as jello effects, will appear if the camera is moving during the image acquisition, which is formulated in [23] as $[\mathbf{I}^R]_i = [\mathbf{I}_i^G]_i$. $[\mathbf{I}_i^G]_i$ is an operator that extracts the i^{th} row from the i^{th} latent GS frame $\mathbf{I}_i^G$ and N is also the height of RS image.

Although image deblurring itself is highly involved, it is believed that restoring clear images from single degraded captures of GS cameras is much more tractable than that of RS cameras as mentioned in [23]. Intuitively, motion-blur removal process aims to disentangle target GS frame from redundant input without sophisticated displacement estimation or speculation of unknown context. On the contrary, RS correction needs to shift pixels of each scanline to one virtual GS canvas by building pixel-level correspondences. Moreover, the rectified output usually suffers from missing boundaries not directly captured during RS exposure. Methodologically, the coarse-to-fine network design has gained remarkable performance for single image deblurring [25, 30, 33, 45, 10, 46]. On the contrary, due to the ill-posed nature, single image RS correction relies heavily on strong prior assumptions, explicitly [35, 17, 32, 31] or implicitly [34, 58], which limits their applicability to real scenarios. So, recent research has moved onto multi-image RS correction to bypass the ill-posedness. But both classical [57, 1] and learning-based methods [23, 9] still cannot work well under complex dynamic scenes or drastic camera motions because of nontrivial pixel alignment.

It is largely overlooked in computer vision that there exists an in-between shutter mode, called global reset release (GRR) shutter, which can be easily enabled in some image sensors that originally work in GS mode (*e.g.*, EV76C560 in EO-1312C Camera) or in RS mode (*e.g.*, IMX178 in BFS-U3-63S4C camera). As shown in Figure 1c, GRR

†Corresponding author

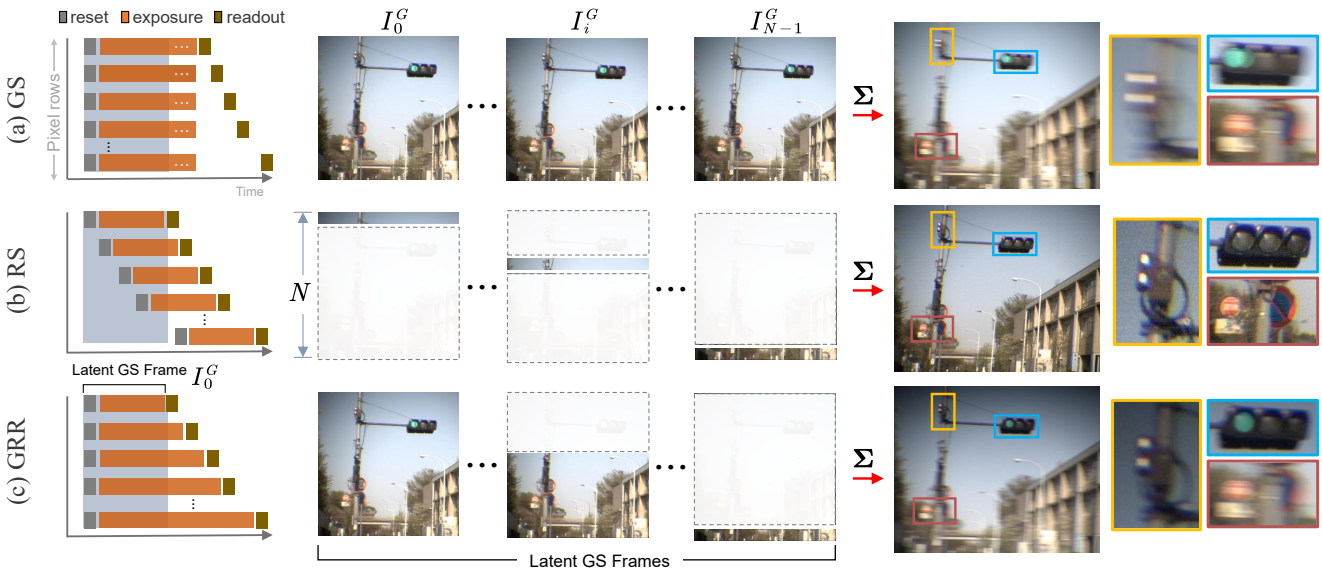

Figure 1: **Different exposure modes and corresponding degrading processes induced by fast motions.** (a) GS exposes all pixels simultaneously, and it causes the output with *row-independent* blur that can be interpreted as an average of the latent sharp frames. (b) RS captures pixels scanline by scanline, featured in higher frame rate, and lower cost than GS. However, RS sensors are prone to *row-dependent* distortion, which can be more challenging to correct than deblurring. (c) GRR begins exposure of all pixels simultaneously but ends row by row, causing blurring effects with *row-dependent* blur magnitude. A specialized deblurring task is required but without the need for RS distortion correction. GRR also has the advantage of reducing the flickering effect of artificial illuminants, such as green traffic lights.

starts to expose all pixels simultaneously as GS does but reads out signals scanline by scanline like conventional RS does. Apparently, GRR leads to blurring effects. However, different from GS blur, the blur magnitude of GRR (the brightness level and Signal-Noise-Ratio as a consequence) is row-dependent. Very recently, Wang *et al.* [47] proposed to enable this global reset feature in RS sensors that turns RS correction into a more tractable deblurring problem and reported superior results of video deblurring in GRR mode. We note the cost to pay for this performance gain by switching from RS to GRR mode includes **1)** the superiority of RS mode over GS mode in capture speed is sacrificed since GRR does not allow early scanning of the next frame before the end of the current frame; **2)** scanning rows in the bottom are more likely to be saturated, and the standard automatic exposure mechanism should be adjusted accordingly. Therefore, in the presence of multiple consecutive frames (e.g. 8 frames are used in [47]), the authentic benefits of using GRR mode are slightly questionable.

In this paper, we deal with a more challenging scenario with a single input image, which is practically meaningful in avoiding well-known flaws of using consecutive frames, including **1)** Larger buffer capacity is needed; **2)** Sensitivity to video recording settings, like frame rate and deadtime between frames, leading to additional generalization issues; **3)** Temporal alignment with severe degradation is still challenging.

By fully considering its unique exposure characteristic of GRR mode, we bring single GRR image deblurring into the community. Mathematically, row-dependent blur and brightness caused by varying exposure time of different scanlines could be formulated as:

$$\mathbf{I}^{RG} = \mathbf{I}_0^G + \delta \sum_{i=1}^{N-1} [\mathbf{I}_i^G]_{i:} \,,$$

where $[\mathbf{I}_i^G]_{i:}$ is an operator that extracts image patch of i^{th} row to the end from the i^{th} latent GS frame $\mathbf{I}_i^G$ and δ denotes the ratio between readout time and the first scanline's exposure duration. As a result, directly applying existing GS deblurring algorithms on GRR deblurring may give rise to an adaptation issue. On the other hand, Transformers have shown significant performance gains on image restoration task [22, 50] by mitigating the shortcomings of CNNs (*i.e.*, limited receptive field and content-independent kernel). Although shifted widow strategy [22, 50, 46, 8, 24] has largely reduced computational loads and made it possible to process high dimensional input, the introduced severe corruption to image spatial information has not been addressed, especially for spatially-sensitive tasks (*e.g.*, GRR deblurring). Besides,

according to the spectral convolution theorem [14], updating a single value in the spectral domain globally affects all original data, which has been implemented for non-local receptive field and proven to be effective in capturing long-range context from frequency domain [4, 5, 55].

We aim to leverage the capability of self-attention and U-net [36] structure with multi-scale input/output to tackle the single GRR deblurring task. To this end, we propose shifted horizontal window-based Transformer block with *spatial compensation* and *spectral enhancement*, which consists of three main components: **1**) Shifted Window-based Multi-head Self-Attention(SW-MSA) with horizontal partition strategy and hierarchical positional encoding. We update original squared window as rectangular one to better adapt row-dependent blur and brightness. Meanwhile, hierarchical positional encoding compensates global positional context of windows lost in window partition and enables order-awareness of local pixel's position within each window. **2**) Spectrally-enhanced Feed-Forward Network (Se-FFN) based on depth-wise convolutional layer. The spectral branch captures the discrepancy between blurry and sharp image pairs from the perspective of frequency domain, as a complementary part of self-attention. **3**) Cross-scale Feature Fusion module based on Squeeze and Excitation block (CFF-SE) [11], which enables us to actively emphasize or suppress the features from encoders with different scales. The main contributions of this paper are summarized as follows:

- We propose an original Transformer block with *spatial compensation* and *spectral enhancement* to address single GRR deblurring by fully exploring the latent frames, which further validates the advantages of GRR exposure mode over its RS counterpart.
- Horizontal partition strategy and hierarchical positional encoding are used to compensate global positional context of windows and enable order-awareness of local pixel's position within each window. The spectrally-enhanced feed-forward network simultaneously uses spatial and spectral information for gaining mixed global context.
- To facilitate the development and evaluation of GRR deblurring, we take paired GRR/Sharp images to offer a new dataset captured under real scenes named GRR-real. Furthermore, by simulating data captured by three shutter modes in a comparable way, the benefits of using GRR for single image restoration are verified.

2. Related Work

Deep Image Deblurring Deep learning methods have achieved significant success in multi [12, 26, 41, 54, 20, 41] or single [15, 18, 25, 27, 40, 45, 48, 51, 52, 44, 19, 29] image deblurring. Initially, researchers try to find the spatially-varying kernels of motion blur before estimating latent image and no exception for deep CNN methods [37, 43]. However, blur kernel is normally computed for all pixels leading to huge demands for memory and computation. Besides, kernel estimation process is overly sensitive to noise and saturation, which is not practical for real scenes. Later, DeepDeblur [25] firstly exploits kernel-free learning to directly construct the relation between blurry and sharp images in an end-to-end manner. And the coarse-to-fine strategy is also taken by following SRN [45] uses an encoder-decoder network with skip-connections for three scale levels to significantly reduce training difficulty and introduce obvious stability benefits. Adversarial training has also been extensively studied [15, 16]. Recently, Cho *et al.* [6] revisit the coarse-to-fine scheme and present a novel deblurring network (MIMO U-net) that can handle multi-scale blur efficiently. These methods have been proven to be effective and achieved remarkable performance in single image deblurring.

Rolling Shutter Correction RS correction methods with single image input could be divided into classical and learning based. Classical approaches heavily rely on strong prior assumptions, such as the scene is static, and the movement of the camera is limited to pure rotation or in-plane translation [35, 17, 32, 31]. Rengarajan *et al.* [35] propose to estimate motion by converting transformed curves to be straight based on the assumption that "straight lines must remain straight" and purely rotation of camera. [32] simplifies the real scene as Manhattan world to rectify the monocular RS image. These classical methods are barely applicable in real situations because of the restrictive prior assumptions. For learning-based methods, Rengarajan *et al.* [34] proposed a new CNN architecture based on long rectangular kernels to correct rolling shutter distortions by simply modeling the camera motion as translation+rotation polynomials. Zhuang *et al.* [58] extend [34] for learning to predict both the camera velocity and depth from a single RS image. The global shutter image is then recovered as a post-processing step. Recent research mainly focuses on multi-image RS correction [57, 1, 23, 9]. But those methods still cannot work under complex dynamic scenes or large camera motions.

Vision Transformers As an alternative to CNN, the Transformer model has been adapted to numerous vision tasks such as image restoration [22, 50, 46], segmentation [49, 53], objection detection [3, 24]. The pioneering work of ViT [8] directly conducts self-attention on flattened patches by decomposing an image into a sequence of tokens and gets excellent results on image classification. To further overcome the quadratic complexity of original self-attention, Liu *et al.* [24] perform local windows with fixed size or window shift to help cross-window interaction. Liang *et al.* proposes an image restoration model, SwinIR [22] based on Swin Transformer [24]. Similarly, Uformer [46] imple-

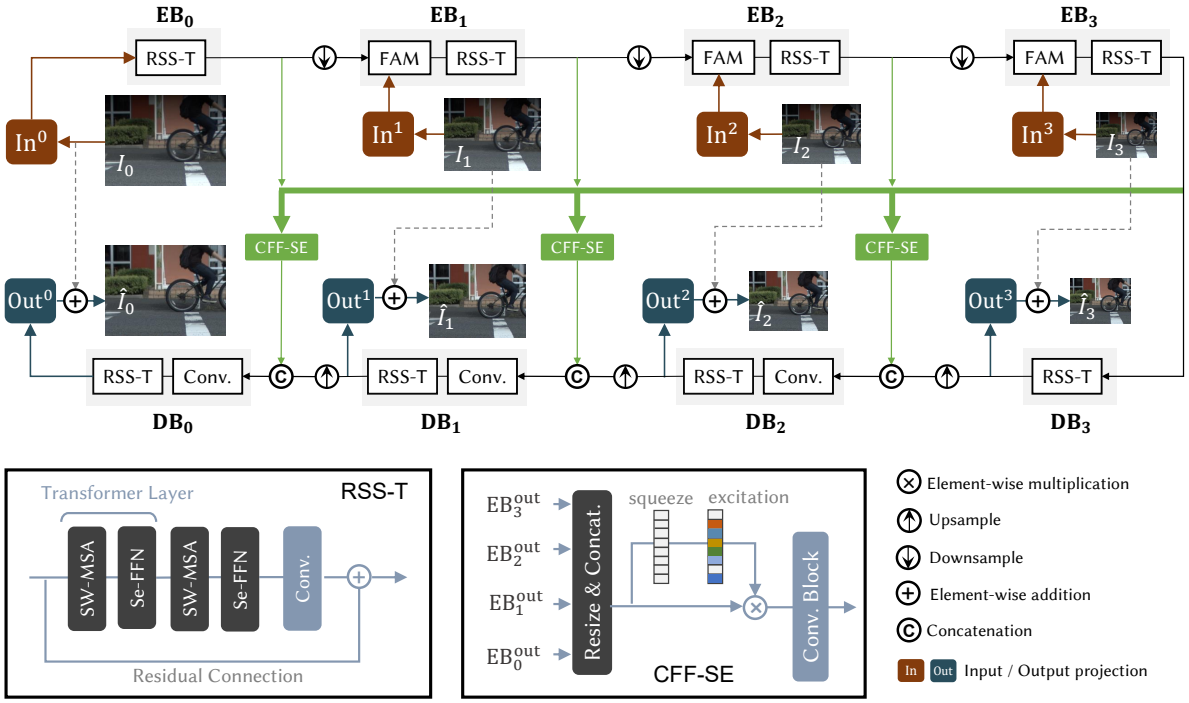

Figure 2: **Overview of our model architecture.** The model is implemented in an encoder-decoder manner with muli-scale input and output. Each encoder/decoder (except for $\texttt{EB}_0$ and $\texttt{EB}_3$) consists of (a) specially designed residual spatially-compensated and spectrally-enhanced transformer block (RSS-T); (b) feature attention module (FAM)/convolution layer (Conv.). Cross-scale feature fusion based on squeeze and excitation block (CFF-SE) is explored to actively emphasize or suppress the features from encoders.

ments the Locally enhanced Window Transformer block in Unet structure and introduces a learnable multi-scale restoration modulator. All these methods have gained remarkable performance on image restoration tasks by mitigating the shortcomings of CNNs. But no one has addressed the *corruption* to image spatial information introduced by the widow partition strategy. And spectral feature, which is proven to capture long-range context effectively, has also been ignored in vision transformers.

3. Method

Overview The overall structure is in the manner of encoder-decoder with multi-scale inputs and outputs (Figure 2). Except for $\texttt{EB}_0$ and $\texttt{DB}_3$, all encoder/decoder blocks contain a Feature Attention Module (FAM) or Convolution layer (Conv.) to refine aggregated features. Cross-scale feature fusion based on squeeze and excitation block (CFF-SE) is exploited to propagate information flow between encoders and decoders. To be specific, given a degraded image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, multi-scale input set $\mathcal{I} = [\mathbf{I}_0, \mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3], \mathbf{I}_k \in \mathbb{R}^{3\times \frac{H}{2^k}\times \frac{W}{2^k}} (k = 0, 1, 2, 3)$ is obtained by down-sampling operations. Then low-level feature $\mathbf{X}_k \in \mathbb{R}^{2^k C\times \frac{H}{2^k}\times \frac{W}{2^k}}$ of input $\mathbf{I}_k$ is extracted by its own input projection module In^k and will be fed into $\texttt{EB}_k$. Similarly, the first three encoder blocks are followed by a down-sampling operation to guarantee the shape of feature map from $\texttt{EB}_{k-1}$ is consistent with that of $\mathbf{X}_k$. For feature reconstruction, an up-sampling operation is inserted between each two decoders. And multi-scale output set $\mathcal{O} = [\hat{\mathbf{I}}_0, \hat{\mathbf{I}}_1, \hat{\mathbf{I}}_2, \hat{\mathbf{I}}_3], \hat{\mathbf{I}}_l \in \mathbb{R}^{3\times \frac{H}{2^l}\times \frac{W}{2^l}} (l = 0, 1, 2, 3)$ is generated by corresponding output projection module Out^l and residual connection from input image $\mathbf{I}_l$. Figure 2 illustrates the detailed structure of our RSS-T block.

3.1. SW-MSA with Spatial Compensation

Two main challenges exist to directly extend the existing Transformer architecture to GRR deblurring. First, because of the characteristic of GRR exposure mode, generated spatial-variant blur and brightness are closely correlated to image rows, while normal Transformer blocks do not consider this. They divide the image into squared window [22, 50, 46], leading to large blur and brightness variance within each window, which makes the rectification process more challenging. Second, standard vision Transformer architecture cuts the computational cost due to the usage of self-attention conducted on non-overlapping windows, which enables high dimensional input processing.

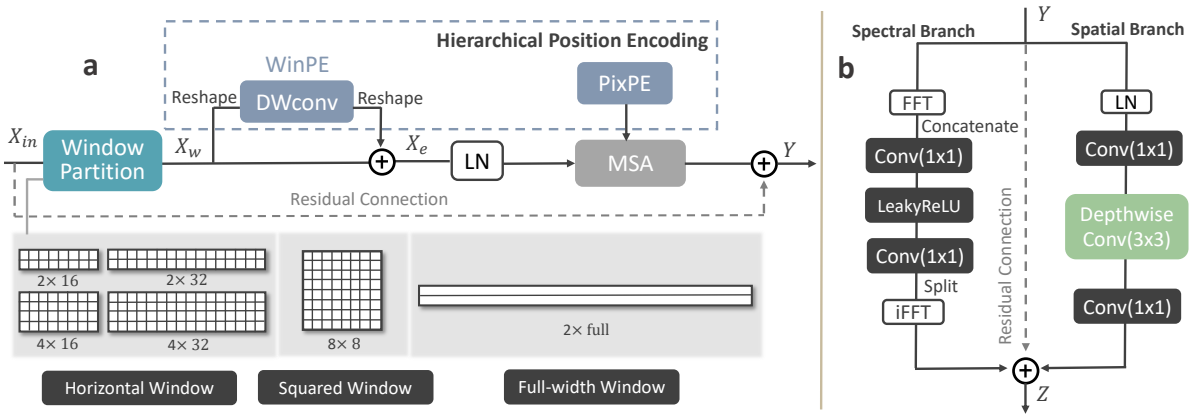

Figure 3: **Two main components of our proposed RSS-T.** (a) The proposed SW-MSA with Spatial Compensation includes two core designs: horizontal window partition strategy and hierarchical position encoding comprising window positional encoding (WinPE) and pixel positional encoding (PixPE). (b) Se-FFN.

But during window partition, image spatial information has been corrupted. That absolute spatial location is crucial for spatially-sensitive tasks. To address the aforementioned issues, we propose SW-MSA with spatial compensation, as shown in Figure 3a, which has two core designs: horizontal window partition and hierarchical positional encoding.

Horizontal Window Partition As discussed above, common squared windows without distinguishing blur and brightness variance along the row and column dimensions make rectification more challenging. We update the original squared window as rectangular to better adapt row-dependent blur and brightness. These long and narrow horizontal windows will mitigate large variances within each widow. Bhojanapalli *et al.* [2] has depicted a kind of low-rank bottleneck, *i.e.*, while increasing the number of heads seemingly gives the model more expressive power, at the same time, the head size is reduced, which can decrease the expressive power. When input token number n is larger than head size d_h, it will create a low-rank bottleneck and lose its ability to represent arbitrary context vectors. Thus, by following this rule and considering computational cost, we set the horizontal window size as $(2, 16), (4, 16), (2, 32)$, and $(4, 32)$. For better comparison, we also provide a normal squared window size of $(8, 8)$ and an extreme case with a window size of $(2, full)$. The 'full' means the window width equals the feature map width. Given input feature $\mathbf{X}_{in} \in \mathbb{R}^{B\times H\times W\times C'}$, the partition process is represented as:

$$\mathbf{X}_w = \text{WinPart}(\mathbf{X}_{in}), \tag{1}$$

where $\mathbf{X}_w \in \mathbb{R}^{BN_w\times H_w\times W_w\times C'}$, H_w, W_w denote height and width of windows and $N_w = {}^{H}/_{H_w} \times {}^{W}/_{W_w}$, which is the total number of divided windows.

Hierarchical Positional Encoding Window partition largely reduces computational cost but corrupts the image's global positional information crucial to GRR deblurring. So, we propose hierarchical positional encoding: **1)** Window Position Encoding (WinPE) compensates for the absolute global location lost in the partition process. **2)** Pixel Position Encoding (PixPE) enables context of the local pixel's relative position within each window. Chu *et al.* [7] proposes a positional encoding generator to obtain Conditional Position Encodings (CPE), which can be efficiently implemented with a 2D depth convolution with kernel $k\,(k \geq 3)$ and $\frac{k-1}{2}$ zero paddings. They have proven zero paddings here are important to make the model aware of the absolute positions of each token. Li *et al.* [21] use a similar positional encoding strategy to capture temporal position for videos. Here, we extend CPE to the 3D situation and propose our WinPE:

$$\text{WinPE}(\mathcal{R}(\mathbf{X}_w)) = \text{DWConv}(\mathcal{R}(\mathbf{X}_w)), \tag{2}$$

where $\mathcal{R}(\cdot)$ is reshape operation and DWConv means 3D depthwise convolution with zero paddings. Each window's absolute position will be encoded by repeatedly applying depthwise convolution to the 3D feature space. Moreover, this encoding is dynamic and allows us to eliminate the constraint to the token's length when testing on different datasets, which perfectly matches the requirement of our window position encoding. The fixed size of the window will lead to the dynamic length of the window sequence once the image size changes. We also apply the relative position encoding [24, 38] for PixPE through a learnable parameter table $\mathcal{B} \in \mathbb{R}^{(2H_w-1)\times(2W_w-1)}$. Overall, the formulation is:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}_{in} + \text{MSA}(\text{LN}(\mathbf{X}_e)), \\ \mathbf{X}_e &= \mathbf{X}_w + \mathcal{R}(\text{DWConv}(\mathcal{R}(\mathbf{X}_w))), \\ \text{MSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{SoftMax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_h} + \mathbf{B})\mathbf{V}, \end{aligned} \tag{3}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are projected query, key and value from input. $\mathbf{B}$ is the relative position bias, whose values are taken

from $\mathcal{B}$. LN$(\cdot)$ means layer norm and d_h denotes head size.

3.2. Spectrally-enhanced FFN

Researchers have proven the spectral feature effectively captures *long-range* dependencies. The Fast Fourier Convolution (FFC) operator in [4] could efficiently implement a non-local receptive field. Inspired by this work, we devised our Se-FFN based on depth-wise convolutional layer. The spectral branch captures the discrepancy between blurry and sharp image pairs as a supplemental part of self-attention for effective long-range dependency modeling.

Se-FFN includes two parts: spectral branch and spatial branch (Figure 3b). The spatial part first applies layer norm to the input feature followed by two convolutions with 1×1 kernel and a 3×3 depth-wise convolution to capture local information. The spectral branch is formulated as (a) convert the input feature map $\mathbf{Y}\in\mathbb{R}^{B\times H\times W\times C'}$ into the frequency domain to obtain $\mathcal{F}(\mathbf{Y})$. (b) Concatenate the real part and imaginary part of $\mathcal{F}(\mathbf{Y})$ along channel dimension to get the spectral feature $\tilde{\mathbf{Y}}\in\mathbb{R}^{B\times H\times W/2\times 2C'}$. (c) Exploit two 1×1 convolutional layers with the LeakyReLU function in between to process the spectral feature. (d) Split the processed spectral feature as real and imaginary parts, then convert them back to spatial domain $\mathbf{Y}_f\in\mathbb{R}^{B\times H\times W\times C'}$ by inverse Fast Fourier Transform (iFFT). Finally, the output of Se-FFN is $\mathbf{Z}=\mathbf{Y}+\mathbf{Y}_f+\mathbf{Y}_s$, where $\mathbf{Y}_s$ is the output of spatial branch.

3.3. Cross-scale Feature Fusion

In most conventional U-net structures, the corresponding encoder and decoder are related by a skip connection to allow information flow between the same scale level. Later in [6], authors present an Asymmetric Feature Fusion (AFF) module to enable cross-scale interaction but without any control. So, we present our CFF-SE block. As shown in Figure 2, before the convolutional block, the input firstly goes through a SE module to control the information flow from the respective scale level of all encoders, which enables us to actively emphasize or suppress the features. Thus, each decoder can selectively exploit multi-scale features, resulting in improved deblurring performance as will be demonstrated in the experimental part.

4. Experiments

Learning-based methods for debluring or RS correction are usually trained on synthetic data, which has limitations of generalization. We, therefore, built an image acquisition system to collect a real dataset named GRR-real. The details about the system, training implementation, and more experimental results are offered in *supplemental materials*.

4.1. Comparison on GRR Deblurring

Comparison with Other Models Since single GRR deblurring is addressed for the first time, there is no specially devised algorithm. So, we compare our model with previous state-of-the-art approaches from:

- GS image-based deblurring by exploiting coarse-to-fine strategy: basic U-net [36], SRN [45], MIMO-UNet[6] and Uformer[46].
- GS video-based deblurring by fusing neighboring temporal information: STR-CNN [12], DBN [42] and IFIRNN[26].
- RS video-based correction through predicting pixel-wise displacement: JCD [56] and DSUR [23].[1]
- GRR video deblurring using a spatial-aware encoder with long-short term temporal information aggregator: NGS [47]. For fair comparisons, we not only provide the original setting with 8 frames as input (NGS_8) but also update the model with a single input (NGS_1).

After retraining all models, Table 1 depicts quantitative results on the GRR-real test dataset. Overall, it demonstrates our specially devised algorithm significantly outperforms models from RS correction and GS deblurring, even though they take multi frames as input and exploit temporal information. As discussed in Section 1, RS correction usually resorts to nontrivial motion estimation and warping process. When directly used for GRR, they performed worse than deblurring models. Furthermore, we observe that, for deblurring models, those with video clips as input may degrade sharply, even are no match for some approaches of a single input. A key reason is that these models are struggling to align the content of different video frames. If failed, the redundant inputs would confuse learning process instead.

Although the tailored GRR video deblurring method NGS_8 [47] achieved higher metrics than ours (30.03/0.90 *vs.* 28.64/0.90), it heavily relies on temporal correlation of 8 frames and the performance decreased drastically when given single input (NGS_1). In contrast to RSS-T, the demand of NGS_8 for 8 consecutive frames in training or testing can be difficult to be satisfied, sometimes even impossible due to buffer restrictions. Besides, sensitivity to video capturing settings (*e.g.,* frame rate and deadtime between frames) will lead to additional generalization issues. We also computed the complexity of all algorithms, which shows that our model is at medium level and outperforms SOTA RS correction or Transformer-based models. From qualitative

[1]Noting that single RS image correction methods either heavily rely on strong assumption [35, 17, 32, 31], which is apparently inapplicable to our natural scene dataset or require actual velocity or depth to train the model [34, 58] that are barely accessible in common cameras. So, we ignore them and only choose RS video correction methods to compare.

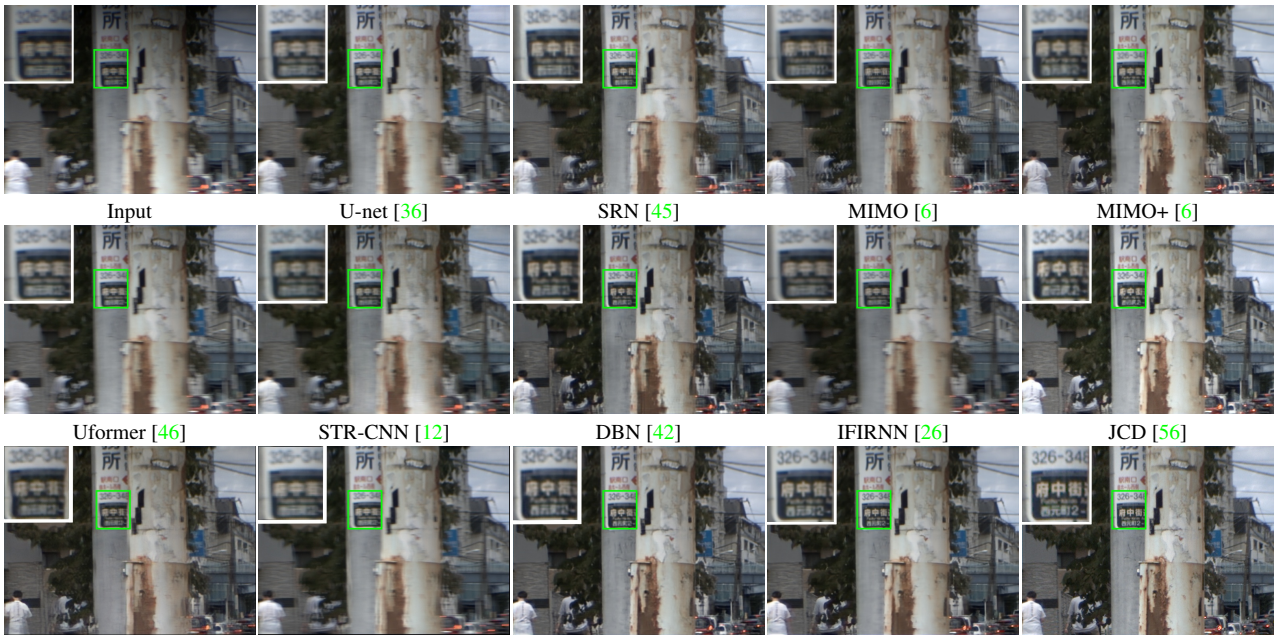

Figure 4: **Qualitative results on the GRR-real dataset.** We compare our method with state-of-the-art methods from RS correction, GS deblurring, and GRR video deblurring.

Table 1: **Quantitative comparison on GRR-real dataset**. We compare our method with state-of-the-art motion deblurring, RS correction, and GRR video deblurring approaches. According to 'exposure mode/input number', they could be divided as 'GS/Single', 'GS/Multi', 'RS/Multi', 'GRR/Multi', and 'GRR/single'. The performance is measured with mean PSNR/SSIM (↑). 'Full' denotes evaluation using full-size frames, while 'Top', 'Middle', and 'Bottom' represent using only a patch of images from the top, middle, and bottom areas.

Method	Mode / Input	Effectiveness				Efficiency		
		Top	Middle	Bottom	Full	Time (s)	Params (M)	FLOPs (T)
Input	–	15.12 / 0.67	21.83 / 0.78	20.08 / 0.76	17.61 / 0.74	–	–	–
U-net [36]	GS / Single	26.34 / 0.92	23.83 / 0.85	22.74 / 0.83	23.57 / 0.87	0.0062	9.50	0.072
SRN [45]		25.11 / 0.78	24.34 / 0.72	24.08 / 0.72	23.84 / 0.74	0.0140	10.25	0.509
MIMO [6]		27.16 /0.90	24.55 / 0.82	24.04 / 0.79	24.48 / 0.84	0.0123	6.81	0.315
MIMO+ [6]		28.46 / 0.91	26.31 / 0.85	25.78 / 0.83	26.12 / 0.86	0.0249	16.11	0.724
Uformer [46]		25.04 / 0.93	23.70 / 0.86	22.20 / 0.84	22.92 / 0.87	0.2147	20.63	0.257
STR-CNN [12]	GS / Multi	20.88 / 0.83	23.28 / 0.77	22.60 / 0.75	21.39 / 0.78	0.0194	0.93	0.367
DBN [42]		25.11 / 0.90	25.90 / 0.85	25.87 / 0.82	24.97 / 0.85	0.0229	15.31	1.046
IFIRNN [26]		27.19 / 0.90	23.85 / 0.78	23.21 / 0.77	23.97 / 0.81	0.0135	1.64	0.581
JCD [56]	RS / Multi	27.64 / 0.90	23.65 / 0.80	19.76 / 0.77	22.31 / 0.82	0.2625	8.67	0.326
DSUR [23]		24.81 / 0.87	23.87 / 0.78	23.39 / 0.76	23.35 / 0.80	0.3018	3.90	0.225
NGS$_8$ [47]	GRR / Multi	31.71 / 0.93	30.54 / 0.90	29.12 / 0.87	30.03 / 0.90	0.0233	8.67	1.266
NGS$_1$ [47]	GRR / Single	26.61 / 0.89	23.53 / 0.78	22.83 / 0.76	23.67 / 0.81	0.0187	4.56	0.083
RSS-T		30.90 / 0.93	28.60 / 0.88	27.86 / 0.86	28.64 / 0.90	0.1479	11.34	0.176

results in Figure 4, although the images obtained by the existing networks exhibit mitigated distortions compared to the input, local details and structures were not sufficiently corrected, whereas our method produces sharper images that even visually outperform NGS$_8$.

Thirdparty Evaluation To further prove the practical value of our method, we evaluate it on another GRR dataset from [47]. As shown in Table 2, our method still outperforms other algorithms and is just slightly inferior to NGS$_8$ with as many as eight input frames (26.62/0.86 *vs.* 27.29/0.85).

Table 2: **Comparison on another GRR dataset [47].**

Method	SET-II			
	Full	Top	Middle	Bottom
Input	17.82 / 0.73	23.64 / 0.77	21.45 / 0.77	15.54 / 0.66
SRN [45]	25.05 / 0.81	24.32 / 0.79	25.65 / 0.81	27.02 / 0.83
STRCNN [12]	22.59 / 0.81	22.99 / 0.79	23.46 / 0.81	23.66 / 0.83
DBN [42]	22.57 / 0.81	23.24 / 0.80	23.81 / 0.81	23.24 / 0.82
IFIRNN[26]	25.17 / 0.82	24.77 / 0.80	25.62 / 0.81	26.94 / 0.84
ESTRNN [54]	22.72 / 0.83	23.42 / 0.81	26.03 / 0.83	22.86 / 0.83
DSUR [23]	22.50 / 0.80	22.49 / 0.78	23.87 / 0.81	23.38 / 0.83
JCD [56]	25.33 / 0.80	24.77 / 0.78	25.71 / 0.80	27.43 / 0.83
NGS_8 [47]	27.29 / 0.85	26.96 / 0.84	27.57 / 0.85	28.35 / 0.86
NGS_1 [47]	25.19 / 0.79	24.48 / 0.77	25.53 / 0.79	27.27 / 0.82
RSS-T	26.62 / 0.86	25.90 / 0.84	26.82 / 0.86	28.43 / 0.87

Table 3: **Architecture ablation on our GRR-real dataset.**

	PE		FFN		Fusion		PSNR / SSIM
	WinPE	PixPE	LeFNN	Se-FNN	AFF	CFF-SE	
v1			✓		✓		25.69 / 0.85
v2	✓		✓		✓		26.50 / 0.86
v3		✓	✓		✓		26.10 / 0.86
v4	✓	✓	✓		✓		26.83 / 0.87
v5	✓	✓	✓			✓	27.24 / 0.87
RSS-T	✓	✓		✓		✓	28.64 / 0.90

4.2. Ablation Study

Model Architecture Ablation We conducted experiments to analyze the effectiveness of the core components of our RSS-T. The hierarchical position encoding consists of WinPE and PixPE. The feed-forward network is also implemented by LeFFN [46] to compare with our Se-FNN (v5 *vs.* RSS-T). As for the cross-scale fusion part, AFF [46] and our CFF-SE are used to explore the performance difference (v4 *vs.* v5). All the experimental results are listed in Table 3. From v1, v2, v3, and v4, when WinPE and PixPE are separately presented in baseline model v1, the performance goes up but the combination of them achieved the best metrics. Our Se-FNN and CFF-SE modules contributed to the further performance improvement of PSNR by $+1.81$dB compared with the LeFFN and AFF.

Window Size Ablation We introduced three types of window size (*i.e.,* horizontal, squared, and full-width window) and clarified the advantages of the horizontal window for GRR deblurring in Section 3.1. Here, we present experimental results to support the claim. As Figure 5 shows, all horizontal window settings are superior to the standard squared one from quantitative and qualitative perspectives. The extreme case $(2, \text{full})$ is severely trapped in low-rank bottleneck [2] as the pixel number of this window is far more than the head size.

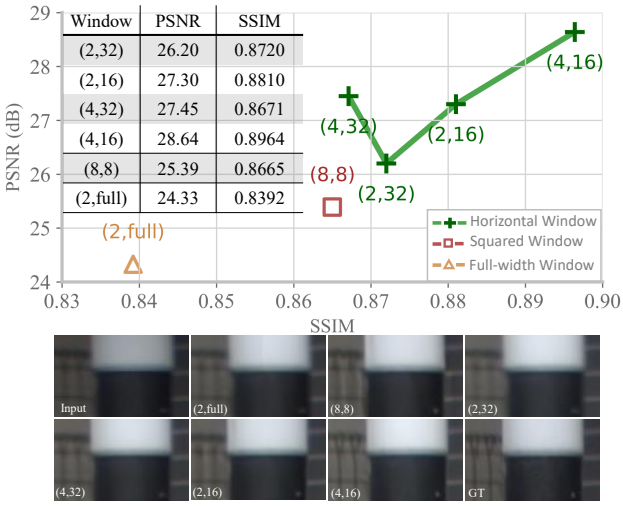

Figure 5: **Window size ablation.** Experimental results of all different window-size settings on our GRR-real dataset.

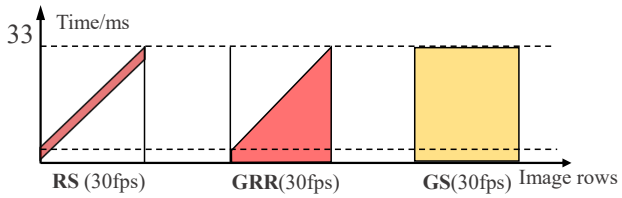

Figure 6: **Temporal alignment of three shutters.** When synthesizing datasets, we strictly align the exposure of three modes.

4.3. Comparison on Three Shutter Modes

Until now, a direct comparison of three shutter modes is still missing, although [47] has demonstrated that GRR has its superiority against RS in obtaining sharp images in the scenario of video restoration. Here, we evaluate them on single image Restoration. For a fair comparison, we synthesize strictly aligned images virtually captured by three modes (RS, GRR, and blur) for each scene by following the protocol of [28, 45, 6] (details in the supplemental material), and compare with SOTA methods for each mode.

When compared with RS mode, Table 4 shows that GRR combined with our RSS-T model has gained significant advantages in obtaining single sharp image, even though DSUR and JCD take multiple frames as input. These results are consistent with the discussion in [47] that the GRR mode enables us to convert the RS correction problem into a more tractable one. Therefore, given an RS camera that supports GRR mode, we recommend switching to GRR mode for better restoration from a single input, if losing some capture speed of the original RS is permissible. Also, the exposure should be adjusted to avoid saturation of bottom rows.

We additionally present comparisons with row-

Table 4: **Quantitative comparison on three modes**. The subscript of each method denotes the number of input frames. The reason we choose multi-frame RS correction is discussed in Section 4.1. MIMO+ and Uformer are trained on blur mode. DSUR and JCD are trained on the RS counterpart, while RSS-T is trained on GRR mode.

Method	Full	Top	Middle	Bottom
$DSUR_2$ [23]	24.51 / 0.81	28.42 / 0.89	23.84 / 0.78	24.21 / 0.77
JCD_3 [56]	23.52 / 0.75	25.85 / 0.79	23.87 / 0.75	23.18 /0.70
RSS-T_1	27.47 / 0.89	27.40 / 0.90	27.9 / 0.89	28.10 / 0.89
MIMO+$_1$[6]	25.16 / 0.86	26.75 / 0.87	25.54 / 0.85	25.58 / 0.85
Uformer$_1$[46]	24.28 / 0.84	25.54 / 0.86	24.26 / 0.83	24.57 / 0.83

independent deblurring issues under identical scenes. For fairness, we artificially elongate the exposure time of a GS camera to synthesize blurred images. As shown in Figure 6, the exposure time of GS is set to be the same as that of the full scanning period of RS and GRR. From Table 4, row-dependent blur from GRR is more tractable than blur generated by a GS camera. Because of the characteristics of GRR exposure, it implicitly encodes the temporal ordering of latent frames that mitigate the motion ambiguity [13] of blur images. And local details from the top sharper yet noisier rows further enhance the reconstruction performance. The practical implication on usage is that, given a GS camera in hand that supports GRR mode, we recommend switching to the GRR mode for single image deblurring with our proposed algorithm. Note that there is no additional cost to pay here since GRR mode will not slow down the original GS frame rate nor cause additional saturation since the longest exposure time of GRR is the same as the GS exposure time (Figure 6).

5. Conclusion

In this paper, we analyzed image degradations for conventional GS and RS cameras and further highlighted an in-between exposure mode, named GRR, arising from the global reset feature equipped with many RS sensors. Instead of RS distortion, GRR gives rise to a specialized blurring effect with row-dependent blur magnitude and brightness level. Based on the characteristics of GRR degradation, an original Transformer block with spatial compensation and spectral enhancement was devised, which has the advantage of being able to rectify the GRR degradation. The proposed hierarchical positional encoding well preserved the absolute and relative position context and the spectral branch of the feed-forward network enhanced the model's ability to capture long-range discrepancies. The experimental results also verified that our tailored model is effective.

Potential Application As shown in Figure 7, GRR is mainly intended for industrial usage in controlled dark environment,

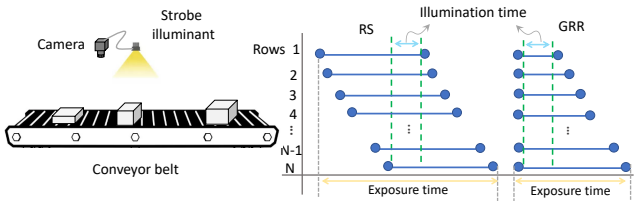

Figure 7: **Application scenario of GRR mode.** (a) Illustrated application scenario. (b) Synchronization between illumination and cameras.

such as parts inspection on a conveyor belt, together with strobe illuminant. In this scenario, GRR with synchronized illumination can capture distortion-free and blur-free images, with lower cost than using a GS sensor, and faster speed than using a RS sensor (Because, in this case, the exposure time of RS has to be elongated such that the first and last rows have sufficient overlap). Instead, without using strobe illumination, GRR boosted with our approach can still restore a sharp image. This will further reduce production and maintenance cost in practice.

Limitations When the illumination is bright enough, with appropriate exposure, GS can capture a sharp image without blur. The bottom parts of GRR can do the same, yet the top parts will be darker and thus noiser, or the top parts can be sharp, yet the the bottom parts will be blurry. In this scenario, GRR is clearly worse than GS, and our algorithm can help reduce the performance gap. On the other hand, RS will not have distortion and the image will be sharp and clean when dealing with completely static scenes. Yet GRR will have non-uniform brightness, although blur will not happen. Our algorithm can compensate this non-uniformity, yet GRR is worse than RS in this particular scenario.

Acknowledgement This research was supported in part by JSPS KAKENHI Grant Numbers 22H00529, 20H05951, 22H03620, 22H05015, JST AIP Acceleration Research JPMJCR22U4, the Value Exchange Engineering, a joint research project between Mercari, Inc. and RIISE.

References

[1] 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. 1, 3

[2] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International Conference on Machine Learning*, pages 864–873. PMLR, 2020. 5, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-

to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[4] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Advances in Neural Information Processing Systems*, 33:4479–4488, 2020. 3, 6

[5] Lu Chi, Guiyu Tian, Yadong Mu, Lingxi Xie, and Qi Tian. Fast non-local neural networks with spectral residual learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2142–2151, 2019. 3

[6] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, 2021. 3, 6, 7, 8, 9

[7] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[9] Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2021. 1, 3

[10] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3848–3856, 2019. 1

[11] 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

[12] 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. 3, 6, 7, 8

[13] 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. 9

[14] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004. 3

[15] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 3

[16] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 3

[17] Yizhen Lao and Omar Ait-Aider. A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4795–4803, 2018. 1, 3, 6

[18] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2034–2042, 2021. 3

[19] Anat Levin. Blind motion deblurring using image statistics. *Advances in Neural Information Processing Systems*, 19, 2006. 3

[20] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7721–7731, 2021. 3

[21] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. *arXiv preprint arXiv:2201.04676*, 2022. 5

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2, 3, 4

[23] 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. 1, 3, 6, 7, 8, 9

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 5

[25] 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. 1, 3

[26] Seungjun Nah, Sanghyun Son, and Kyoung Mu Lee. Recurrent neural networks with intra-frame iterations for video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8102–8111, 2019. 3, 6, 7, 8

[27] Mehdi Noroozi, Paramanand Chandramouli, and Paolo Favaro. Motion deblurring in the wild. In *German conference on pattern recognition*, pages 65–77. Springer, 2017. 3

[28] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pages 198–215. Springer, 2022. 8

[29] Chandramouli Paramanand and Ambasamudram N Rajagopalan. Non-uniform motion deblurring for bilayer scenes.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1122, 2013. 3

[30] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *European Conference on Computer Vision*, pages 327–343. Springer, 2020. 1

[31] Pulak Purkait and Christopher Zach. Minimal solvers for monocular rolling shutter compensation under ackermann motion. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 903–911. IEEE, 2018. 1, 3, 6

[32] Pulak Purkait, Christopher Zach, and Ales Leonardis. Rolling shutter correction in manhattan world. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–890, 2017. 1, 3, 6

[33] Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11882–11889, 2020. 1

[34] 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. 1, 3, 6

[35] 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. 1, 3, 6

[36] 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. 3, 6, 7

[37] Christian J Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, 2015. 3

[38] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 5

[39] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5114–5123, 2020. 1

[40] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2642–2650, 2021. 3

[41] Hyeongseok Son, Junyong Lee, Jonghyeop Lee, Sunghyun Cho, and Seungyong Lee. Recurrent video deblurring with blur-invariant motion estimation and pixel volumes. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021. 3

[42] 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. 6, 7, 8

[43] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 769–777, 2015. 3

[44] Yu-Wing Tai, Ping Tan, and Michael S Brown. Richardson-lucy deblurring for scenes under a projective motion path. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1603–1618, 2010. 3

[45] 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. 1, 3, 6, 7, 8

[46] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 1, 2, 3, 4, 6, 7, 8, 9

[47] Zhixiang Wang, Xiang Ji, Jia-Bin Huang, Shin'ichi Satoh, Xiao Zhou, and Yinqiang Zheng. Neural global shutter: Learn to restore video from a rolling shutter camera with global reset feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2022. 2, 6, 7, 8

[48] Patrick Wieschollek, Michael Hirsch, Bernhard Scholkopf, and Hendrik Lensch. Learning blind motion deblurring. In *Proceedings of the IEEE international conference on computer vision*, pages 231–240, 2017. 3

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[50] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 2, 3, 4

[51] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 3

[52] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 3

[53] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[54] Zhihang Zhong, Ye Gao, Yinqiang Zheng, and Bo Zheng. Efficient spatio-temporal recurrent neural network for video deblurring. In *European Conference on Computer Vision*, pages 191–207. Springer, 2020. 3, 8

[55] Zhisheng Zhong, Tiancheng Shen, Yibo Yang, Zhouchen Lin, and Chao Zhang. Joint sub-bands learning with clique structures for wavelet domain super-resolution. *Advances in neural information processing systems*, 31, 2018. 3

[56] 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. 6, 7, 8, 9

[57] 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. 1, 3

[58] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4551–4560, 2019. 1, 3, 6