Self-supervised Learning to Bring Dual Reversed Rolling Shutter Images Alive

Wei Shang[1,2], Dongwei Ren[1*], Chaoyu Feng, Xiaotao Wang, Lei Lei, Wangmeng Zuo[1,3]
[1]School of Computer Science and Technology, Harbin Institute of Technology
[2]City University of Hong Kong [3]Peng Cheng Laboratory, Shenzhen

Abstract

Modern consumer cameras usually employ the rolling shutter (RS) mechanism, where images are captured by scanning scenes row-by-row, yielding RS distortions for dynamic scenes. To correct RS distortions, existing methods adopt a fully supervised learning manner, where high framerate global shutter (GS) images should be collected as ground-truth supervision. In this paper, we propose a Self-supervised learning framework for Dual reversed RS distortions Correction (SelfDRSC), where a DRSC network can be learned to generate a high framerate GS video only based on dual RS images with reversed distortions. In particular, a bidirectional distortion warping module is proposed for reconstructing dual reversed RS images, and then a self-supervised loss can be deployed to train DRSC network by enhancing the cycle consistency between input and reconstructed dual reversed RS images. Besides start and end RS scanning time, GS images at arbitrary intermediate scanning time can also be supervised in SelfDRSC, thus enabling the learned DRSC network to generate a high framerate GS video. Moreover, a simple yet effective self-distillation strategy is introduced in self-supervised loss for mitigating boundary artifacts in generated GS images. On synthetic dataset, SelfDRSC achieves better or comparable quantitative metrics in comparison to state-of-the-art methods trained in the full supervision manner. On real-world RS cases, our SelfDRSC can produce high framerate GS videos with finer correction textures and better temporary consistency. The source code and trained models are made publicly available at https://github.com/shangwei5/SelfDRSC.

1. Introduction

Recent years have witnessed an increasing demand for imaging sensors, due to the widespread applications of digital cameras and smartphones. Although the Charge-Coupled Device (CCD) has been the dominant technology for imaging sensors, it is recently popular that modern consumer cameras choose the Complementary Metal-Oxide Semiconductor (CMOS) as an alternative due to its many merits, *e.g.*, easy integration with image processing pipeline and communication circuits, and low power consumption [14]. In CMOS sensors, rolling shutter (RS) scanning mechanism is generally deployed to capture images, *i.e.*, each row of CMOS array is exposed in the sequential time, which is different from CCD with global shutter (GS) scanning at one instant. Therefore, RS images suffer from distortions when capturing dynamic scenes, which not only affect human visual perception but also yield performance degradation or even failure in computer vision tasks [2, 13].

To correct RS distortions, pioneering works usually reconstruct GS images from a single RS image [20, 32] or multiple consecutive RS images [15, 6], where the latter ones usually have better performance. But consecutive RS images setting is ambiguous [30], *e.g.*, two RS cameras, moving horizontally at the same speed but with different readout time, can produce the same RS images. Most recently, a new RS acquisition setting, *i.e.*, dual RS images with reversed scanning directions (see Fig. 1), is proposed in [1, 30] to address this ambiguity. In [1], one GS image is reconstructed from dual RS images with reversed distortions, while in [30], Zhong *et al*. devote efforts to reconstruct high framerate GS videos. Nevertheless, both of them adopt a fully supervised learning manner, *i.e.*, ground-truth GS supervision is required to learn RS correction networks. Especially in [30], high framerate GS videos should be collected to serve as ground-truth. In the supervised learning manner, it is not easy to collect real-world training samples, while synthetic training samples would yield poor generalization ability when handling real-world RS cases.

In this paper, we aim ambitiously for the more challenging and practical task, *i.e.*, self-supervised learning to invert dual reversed RS images to a high framerate GS video, dubbed SelfDRSC, which to the best of our knowledge is studied for the first time. The primary design philosophy of our SelfDRSC is that latent GS images predicted by the DRSC network can be used to reconstruct dual RS images with reversed distortions, and then the DRSC network can

*Corresponding author: rendongweihit@gmail.com.

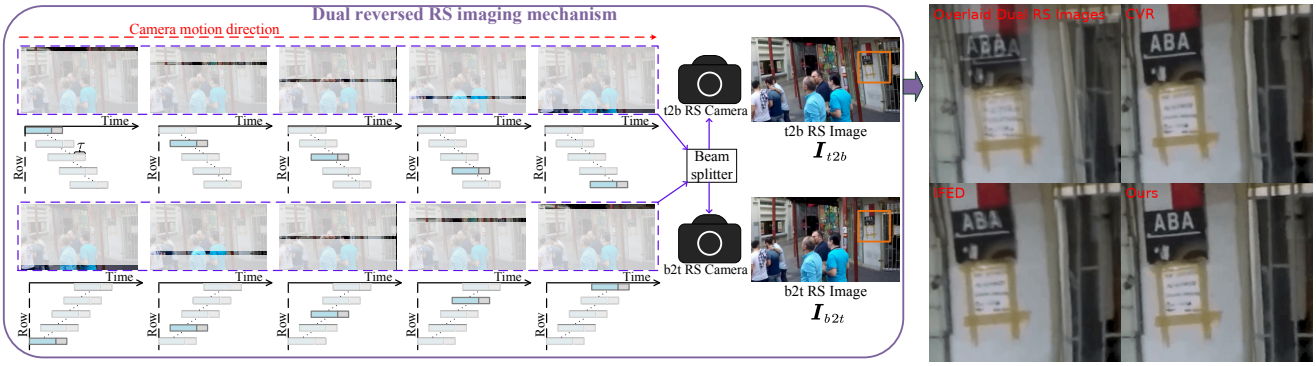


Figure 1. Illustration of capturing dual RS images with reversed scanning directions, *i.e.*, top-to-bottom ($\boldsymbol{I}_{t2b}$) and bottom-to-top ($\boldsymbol{I}_{b2t}$). In this work, we propose the first self-supervised learning method SelfDRSC to correct RS distortions. In comparison to state-of-the-art supervised RS correction methods CVR [7] and IFED [30], our SelfDRSC can generate high framerate GS videos with finer textures and better temporary consistency. The result can be displayed in an animated figure in the ***arXiv*** version.

be learned by enforcing the cycle consistency between input and reconstructed RS images. As shown in Fig. 2, a novel bidirectional warping (BDWarping) module is proposed, by which dual reversed RS images can be reconstructed, and then a self-supervised loss can be deployed to train the DRSC network. During training, an intermediate GS image at arbitrary RS scanning time is predicted, and in our BDWarping module, it can also be used to reconstruct another set of dual RS images, which serve as the extra self-supervision in SelfDRSC. In this way, the predicted GS images at intermediate scanning time can also be supervised, making the learned DRSC network be able to generate high framerate GS videos. Moreover, the DRSC network trained by individual self-supervised loss would yield undesirable boundary artifacts as shown in Fig. 3, and we introduce a self-distillation strategy into self-supervised loss to alleviate this issue, as shown in Fig. 4.

Extensive experiments on synthetic and real-world RS images have been conducted to evaluate our SelfDRSC. Although any ground-truth GS supervision is not exploited in our SelfDRSC, it still achieves comparable quantitative metrics on synthetic dataset, in comparison to state-of-the-art supervised RS correction methods. On real-world RS cases, our SelfDRSC can produce high framerate GS videos with finer textures and better temporary consistency.

2. Related Work

2.1. Rolling Shutter Distortion Correction

With the rising demand for RS cameras, RS distortion corrections have received widespread attention. Existing works on RS correction generally fall into two categories: single-image-based [20, 32] and multi-frame-based [15, 6, 30] methods. It is an ill-posed problem to correct RS distortion from a single image, and its performance is usually inferior. For multi-frame-based methods, it can be further divided into generating one specific image and generating a video sequence. For the former, Liu *et al*. [15] proposed an end-to-end model, which warped features of RS images to a GS image by a special forward warping block. For the latter, Fan *et al*. [6] designed a deep learning framework, which utilized the underlying spatio-temperal geometric relationships for generating a latent GS image sequence. Then Fan *et al*. [7] further proposed a context-aware model for solving complex occlusions and object-specific motion artifacts. Recently, Zhong *et al*. [30] proposed an end-to-end method IFED, and it can extract an undistorted GS sequence grounded on the symmetric and complementary nature of dual RS images with reversed distortion [1]. However, these methods all rely on supervised learning, which would yield poor generalization on real-world data. In this work, we aim to develop a self-supervised learning framework for inverting dual reversed RS images to high framerate GS videos with visually pleasing results.

2.2. Cycle Consistency-based Self-supervised Learning in Low-level Vision

The concept of cycle consistency has been utilized in several low-level vision tasks for self-supervised learning. Zhu *et al*. [31] proposed a cycle consistency loss for unpaired image-to-image translation. Chen *et al*. [5] enforced the results by fine-tuning existing methods in a self-supervised fashion, where they estimated the per-pixel blur kernel based on optical flows between restored frames, for reconstructing blurry inputs. Ren *et al*. [23] utilized two generative networks for respectively modeling the deep priors of clean image and blur kernel to reconstruct blurred image for enforcing cycle consistency. Liu *et al*. [16] claimed that motion cues obtained from consecutive images yield sufficient information for deblurring task and they re-rendered the blurred images with predicting optical flows for cycle consistency-based self-supervised learning. Bai *et al*. [3] presented a self-supervised video super-resolution

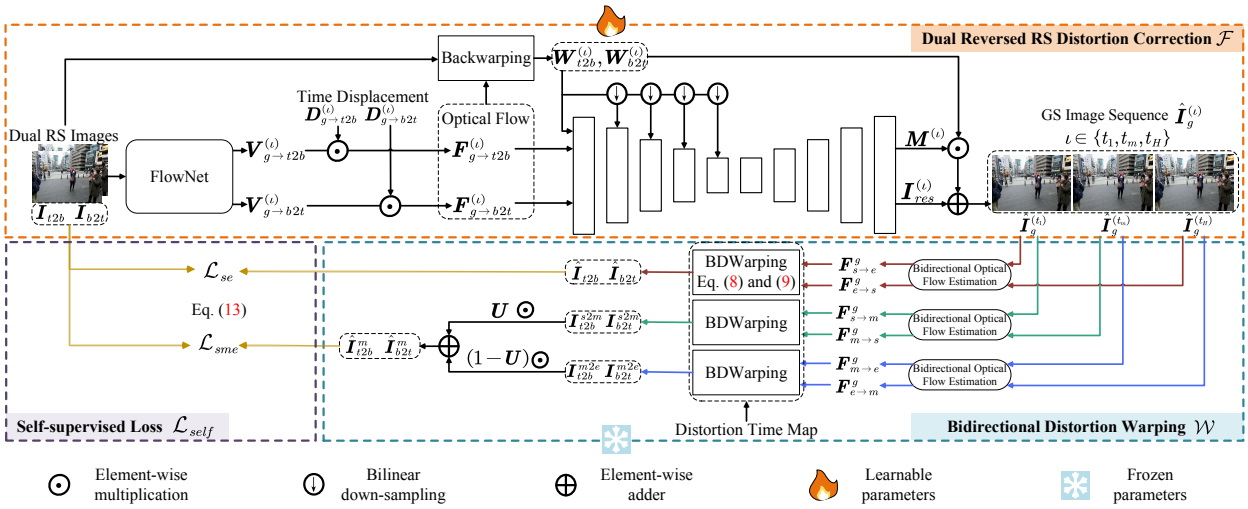


Figure 2. Training framework of our SelfDRSC, which consists of three modules, *i.e.*, DRSC network $\mathcal{F}$ for generating GS images $\{\hat{\boldsymbol{I}}_g^{(t_1)}, \hat{\boldsymbol{I}}_g^{(t_m)}, \hat{\boldsymbol{I}}_g^{(t_H)}\}$ from input dual RS images $\boldsymbol{I}_{t2b}$ and $\boldsymbol{I}_{b2t}$, a bidirectional distortion warping module $\mathcal{W}$ for reconstructing dual reversed RS images, and self-supervised loss $\mathcal{L}_{self}$ for enforcing the cycle consistency between input and reconstructed RS images. Moreover, a self-distillation loss $\mathcal{L}_{sd}$, referring to Fig. 4, is introduced into self-supervised loss for mitigating boundary artifacts in generated GS images. In inference phase, the learned DRSC model $\mathcal{F}$ is able to generate high framerate GS videos by giving multiple intermediate time t_m.

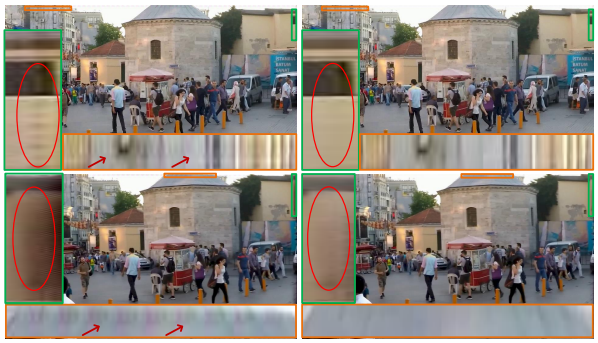
method, which can generate auxiliary paired data from the original low resolution input videos to constrain the network training. To sum up, in these methods, degraded images can be reconstructed based on the imaging mechanism, and then self-supervised loss can be employed to enforce the cycle consistency between reconstructed and original degraded images. In this work, we incorporate cycle consistency-based self-supervised loss with self-distillation for tackling DRSC problem.

3. Proposed Method

In this section, we first present the problem formulation of self-supervised learning for Dual reversed RS Correction (SelfDRSC). Then, the DRSC network architecture is briefly introduced, and more attention is paid on our proposed self-supervised learning framework including bidirectional distortion warping module and self-supervised loss with self-distillation.

3.1. Formulation of SelfDRSC

Recently, dual reversed RS imaging setting [1, 30] has been proposed to address the ambiguity issue in consecutive RS images, where two RS images are captured simultaneously by different scanning patterns, *i.e.*, top-to-bottom ($\boldsymbol{I}_{t2b}$) and bottom-to-top ($\boldsymbol{I}_{b2t}$), as shown in Fig. 1. We first give a formal imaging formation of dual RS images with H rows. Without loss of generality, we define the acquisition time t as the midpoint of the whole exposure period, *i.e.*, each RS image is captured from t_1 to t_H, having



w/o $\mathcal{L}_{sd}$ w/ $\mathcal{L}_{sd}$

Figure 3. Self-distillation $\mathcal{L}_{sd}$ for mitigating boundary artifacts.

$H-1$ readout instants τ, where $t_1 = t - \tau(H-1)/2$ and $t_H = t + \tau(H-1)/2$. Dual reversed RS images captured at time t can be defined as

$$\begin{aligned} \boldsymbol{I}_{t2b}^{(t)}[i] &= \boldsymbol{I}_g^{(t+\tau(i-(H+1)/2))}[i], \\ \boldsymbol{I}_{b2t}^{(t)}[i] &= \boldsymbol{I}_g^{(t+\tau(i-(H+1)/2))}[H-i+1], \end{aligned} \tag{1}$$

where $\boldsymbol{I}_g$ is the latent GS image. When scanning i-th rows for $\boldsymbol{I}_{t2b}^{(t)}$ and $\boldsymbol{I}_{b2t}^{(t)}$, the image contents are captured from $\boldsymbol{I}_g$ at the same instant time t_i but with reversed scanning directions. In the following, the superscripts in $\boldsymbol{I}_{t2b}^{(t)}$ and $\boldsymbol{I}_{b2t}^{(t)}$ are omitted, since $t = (t_1 + t_H)/2$.

Under the dual reversed RS imaging setting, RS distortions can be well distinguished, *i.e.*, dual reversed RS images $\boldsymbol{I}_{t2b}$ and $\boldsymbol{I}_{b2t}$ provide cues for reconstructing GS images between t_1 and t_H, even for the ambiguous case in

consecutive video setting. In the most recent state-of-the-art method IFED [30], fully supervised learning is employed for learning DRSC network, where high framerate GS video frames should be collected as ground-truth supervision. Albeit obtaining promising performance, high framerate GS frames are not trivial to collect. The common solution is to synthesize training datasets where video frame interpolation (VFI) is usually adopted to increase video framerate [30], thus restricting their performance on real-world cases.

In this work, we propose a novel SelfDRSC method for rolling shutter correction with dual reversed distortion, where only RS images are required for training DRSC network, without requiring ground-truth high framerate GS images. Formally, the optimization of SelfDRSC is defined as

$$\min_{\boldsymbol{\Theta}} \mathcal{L}\left(\{\boldsymbol{I}_{t2b}, \boldsymbol{I}_{b2t}\}, \mathcal{W}\left(\mathcal{F}\left(\boldsymbol{I}_{t2b}, \boldsymbol{I}_{b2t}; \boldsymbol{\Theta}\right)\right)\right), \quad (2)$$

which contains three key components: DRSC network $\mathcal{F}$ with parameters $\boldsymbol{\Theta}$, bidirectional distortion warping (BDWaring) module $\mathcal{W}$, and self-supervised learning objective $\mathcal{L}$. By taking dual RS images $\boldsymbol{I}_{t2b}$ and $\boldsymbol{I}_{b2t}$ as input, DRSC network $\mathcal{F}$ generates high framerate GS frames. To learn parameters $\boldsymbol{\Theta}$ of $\mathcal{F}$, BDWaring module $\mathcal{W}$ reconstructs dual RS images from generated GS frames, and self-supervised learning loss $\mathcal{L}_{self}$ is imposed to enforce the cycle consistency between input and reconstructed RS images. Moreover, self-distillation $\mathcal{L}_{sd}$ is introduced for mitigating boundary artifacts in generated GS images.

3.2. Network Architecture of DRSC $\mathcal{F}$

The architecture of $\mathcal{F}$ is similar with that in IFED [30], which consists of a RS correction module and a GS reconstruction module. In IFED [30], multiple GS frames are directly adopted as supervision for training DRSC network, and the output of IFED has fixed framerate. In contrast, our SelfDRSC does not require high framerate GS images as ground-truth supervision. Therefore, during training phase, our DRSC network $\mathcal{F}$ generates three images $\{\hat{\boldsymbol{I}}_g^{(t_1)}, \hat{\boldsymbol{I}}_g^{(t_m)}, \hat{\boldsymbol{I}}_g^{(t_H)}\}$ where t_m is an intermediate scanning time between start time t_1 and end time t_H. During inference phase, it allows DRSC network to generate videos with arbitrary framerate by giving different intermediate scanning time t_m. In the following, we take an example to show how intermediate GS image $\hat{\boldsymbol{I}}_g^{(t_m)}$ is predicted from dual reversed RS images $\boldsymbol{I}_{b2t}$ and $\boldsymbol{I}_{t2b}$. The start and end GS images $\hat{\boldsymbol{I}}_g^{(t_1)}$ and $\hat{\boldsymbol{I}}_g^{(t_H)}$ can be obtained in the same way.

RS Correction Module. For inverting RS images to GS images, a natural strategy is to warp RS images based on the optical flow $\boldsymbol{F}_{g\to t2b}^{(t_m)}$ and $\boldsymbol{F}_{g\to b2t}^{(t_m)}$ between input RS images and latent GS images. But direct estimation is a challenging problem due to the time displacement and relative motion [30]. Following IFED, we adopt a simple FlowNet [30] to estimate relative motion map $\boldsymbol{V}_{g\to t2b}^{(t_m)}$ and $\boldsymbol{V}_{g\to b2t}^{(t_m)}$ between latent GS images and the input RS images. As for the time displacement $\boldsymbol{D}^{(t_m)}$ between input RS images and latent GS images, they can be obtained based on the scanning mechanism of RS cameras. Formally, the values at i-th row in time displacement are given by

$$\begin{aligned} \boldsymbol{D}_{g\to t2b}^{(t_m)}[i] &= \frac{i-m}{H-1}, && i, m \in [1, \cdots, H], \\ \boldsymbol{D}_{g\to b2t}^{(t_m)}[i] &= \frac{(H-i)-(m-1)}{H-1}, && i, m \in [1, \cdots, H]. \end{aligned} \quad (3)$$

Then the optical flow between input RS images and latent GS images can be obtained by multiplying corresponding time displacement and relative motion map in the entry-by-entry manner, *i.e.*, $\boldsymbol{F}_{g\to t2b}^{(t_m)} = \boldsymbol{D}_{g\to t2b}^{(t_m)} \odot \boldsymbol{V}_{g\to t2b}^{(t_m)}$ and $\boldsymbol{F}_{g\to b2t}^{(t_m)} = \boldsymbol{D}_{g\to b2t}^{(t_m)} \odot \boldsymbol{V}_{g\to b2t}^{(t_m)}$. Finally, the corrected images $\boldsymbol{W}_{t2b}^{(t_m)} = \mathcal{B}(\boldsymbol{I}_{t2b}; \boldsymbol{F}_{g\to t2b}^{(t_m)})$ and $\boldsymbol{W}_{b2t}^{(t_m)} = \mathcal{B}(\boldsymbol{I}_{b2t}; \boldsymbol{F}_{g\to b2t}^{(t_m)})$ can be obtained from $\boldsymbol{I}_{t2b}$ and $\boldsymbol{I}_{b2t}$ by the backwarping operation $\mathcal{B}$ [30].

GS Reconstruction Module. The remaining issue is how to fuse warped images $\boldsymbol{W}_{t2b}^{(t_m)}$ and $\boldsymbol{W}_{b2t}^{(t_m)}$ for reconstructing latent GS image $\boldsymbol{I}_g^{(t_m)}$. In our work, encoder-decoder is adopted, where dual corrected images ($\boldsymbol{W}_{t2b}^{(t_m)}$ and $\boldsymbol{W}_{b2t}^{(t_m)}$) and optical flows ($\boldsymbol{F}_{g\to t2b}^{(t_m)}$ and $\boldsymbol{F}_{g\to b2t}^{(t_m)}$) are taken as input, and a fusing mask $\boldsymbol{M}^{(t_m)}$ and residual image $\boldsymbol{I}_{res}^{(t_m)}$ are generated as output. In particular, the reconstruction is performed in a multi-scale framework, where 5 levels are adopted. And finally, the GS image can be obtained by

$$\hat{\boldsymbol{I}}_g^{(t_m)} = \boldsymbol{I}_{res}^{(t_m)} + \boldsymbol{M}^{(t_m)} \odot \boldsymbol{W}_{t2b}^{(t_m)} + (\mathbf{1} - \boldsymbol{M}^{(t_m)}) \odot \boldsymbol{W}_{b2t}^{(t_m)}. \quad (4)$$

More details of $\mathcal{F}$ can be found in the arXiv version.

3.3. Self-supervised Learning for DRSC

To learn the parameters $\boldsymbol{\Theta}$ of DRSC network $\mathcal{F}$, we need to introduce supervision on $\{\hat{\boldsymbol{I}}_g^{(t_1)}, \hat{\boldsymbol{I}}_g^{(t_m)}, \hat{\boldsymbol{I}}_g^{(t_H)}\}$. Instead of collecting ground-truth GS images, we suggest that supervision can be exploited from the input RS images themselves. Generally, we introduce a bidirectional distortion warping module $\mathcal{W}$ to reconstruct dual reversed RS images, and self-supervised loss $\mathcal{L}$ can be adopted to learn the parameters $\boldsymbol{\Theta}$ without ground-truth GS images.

3.3.1 Reconstruction of Dual Reversed RS Images

Based on RS imaging mechanism, generated start and end GS images $\hat{\boldsymbol{I}}_g^{(t_1)}$ and $\hat{\boldsymbol{I}}_g^{(t_H)}$ can be accordingly exploited to reconstruct dual reversed RS images $\hat{\boldsymbol{I}}_{t2b}$ and $\hat{\boldsymbol{I}}_{b2t}$. Besides start and end GS images, we also provide a way to reconstruct RS images $\hat{\boldsymbol{I}}_{t2b}^m$ and $\hat{\boldsymbol{I}}_{b2t}^m$ from intermediate GS images $\hat{\boldsymbol{I}}_g^{(t_m)}$. In the following, we take top-to-bottom scanning pattern as an example to show the reconstruction of $\hat{\boldsymbol{I}}_{t2b}$.

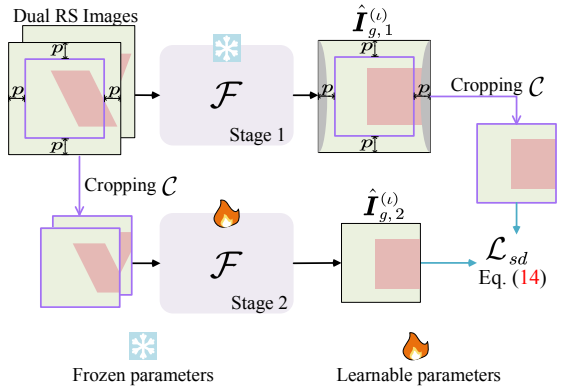


Figure 4. Self-distillation loss for mitigating boundary artifacts. The DRSC network $\mathcal{F}$ learned by individual $\mathcal{L}_{self}$ in Stage 1 generates GS images $\hat{\boldsymbol{I}}_{g,1}^{(\iota)}$ having high-quality center regions but suffering from boundary artifacts. For training $\mathcal{F}$ in Stage 2, the GS images $\hat{\boldsymbol{I}}_{g,1}^{(\iota)}$ are cropped by $\mathcal{C}$ with boundary cropping size p to serve as pseudo GS supervision for finetuning $\mathcal{F}$, whose boundary artifacts can be well mitigated.

Reconstructing RS Images from Start & End GS Frames $\hat{\boldsymbol{I}}_g^{(t_1)}$ and $\hat{\boldsymbol{I}}_g^{(t_H)}$. The purpose of our bidirectional distortion warping $\mathcal{W}$ is to reconstruct $\hat{\boldsymbol{I}}_{t2b}$ from $\hat{\boldsymbol{I}}_g^{(t_1)}$ and $\hat{\boldsymbol{I}}_g^{(t_H)}$, where the key issue is to obtain bidirectional distortion optical flows $\boldsymbol{F}_{t2b\to g}^{s2e}$ and $\boldsymbol{F}_{t2b\to g}^{e2s}$. Then $\hat{\boldsymbol{I}}_{t2b}$ can be reconstructed using a backwarping operation. First, given the corrected results $\hat{\boldsymbol{I}}_g^{(t_1)}$ and $\hat{\boldsymbol{I}}_g^{(t_H)}$, it is easy to compute bidirectional optical flows $\boldsymbol{F}_{s\to e}^{g}$ and $\boldsymbol{F}_{e\to s}^{g}$ between GS frames, by using a pre-trained optical flow estimation network (*e.g.*, PWC-Net [25] and GMFlow [27]). According to arbitrary time flow interpolation method in VFI [9], we can estimate optical flows between t_1 (or t_H) and arbitrary intermediate time, *e.g.*, linear approximation [10], flow reversal [28] and CFR [24]. We need to introduce distortion time map in RS imaging process to obtain optical flows $\boldsymbol{F}_{t2b\to g}^{s2e}$ and $\boldsymbol{F}_{t2b\to g}^{e2s}$ between GS images and reconstructed RS image $\hat{\boldsymbol{I}}_{t2b}$.

Hence, we design distortion time map for representing the interpolation time in each row. Distortion time map $\boldsymbol{T}_{s\to e}^{t2b}$ for estimating distortion optical flow from start time t_1 to end time t_H can be formulated as

$$\boldsymbol{T}_{s\to e}^{t2b}[i] = \frac{(i-1)\cdot\tau}{(H-1)\cdot\tau} = \frac{i-1}{H-1}, i\in[1,\cdots,H]. \quad (5)$$

And we can also get $\boldsymbol{T}_{e\to s}^{t2b} = \boldsymbol{1} - \boldsymbol{T}_{s\to e}^{t2b}$. According to CFR [24], we need to first obtain anchor flows $\boldsymbol{F}_1^{s2e}$ and $\boldsymbol{F}_1^{e2s}$, and complementary flows $\boldsymbol{F}_2^{e2s}$ and $\boldsymbol{F}_2^{s2e}$ for complementally filling the holes occurred in the reversed flows. Different from CFR, the time instance at each row is different. Anchor flows can be calculated as

$$\boldsymbol{F}_1^{s2e} = \boldsymbol{T}_{s\to e}^{t2b}\odot\boldsymbol{F}_{s\to e}^{g}, \text{ and } \boldsymbol{F}_1^{e2s} = \boldsymbol{T}_{e\to s}^{t2b}\odot\boldsymbol{F}_{e\to s}^{g}. \quad (6)$$

And the complementary flows are normalized as

$$\boldsymbol{F}_2^{e2s} = \boldsymbol{T}_{s\to e}^{t2b}\odot\boldsymbol{F}_{e\to s}^{g}, \text{ and } \boldsymbol{F}_2^{s2e} = \boldsymbol{T}_{e\to s}^{t2b}\odot\boldsymbol{F}_{s\to e}^{g}. \quad (7)$$

Then we can obtain distortion optical flow as follows

$$\boldsymbol{F}_{t2b\to g}^{s2e}(\mathbf{x}) = \frac{\boldsymbol{T}_{s\to e}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_2}w_2\boldsymbol{F}_2^{e2s}(\mathbf{y}_2) - \boldsymbol{T}_{e\to s}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_1}w_1\boldsymbol{F}_1^{s2e}(\mathbf{y}_1)}{\boldsymbol{T}_{e\to s}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_1}w_1 + \boldsymbol{T}_{s\to e}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_2}w_2},$$
$$\boldsymbol{F}_{t2b\to g}^{e2s}(\mathbf{x}) = \frac{\boldsymbol{T}_{e\to s}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_1}w_1\boldsymbol{F}_2^{s2e}(\mathbf{y}_1) - \boldsymbol{T}_{s\to e}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_2}w_2\boldsymbol{F}_1^{e2s}(\mathbf{y}_2)}{\boldsymbol{T}_{e\to s}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_1}w_1 + \boldsymbol{T}_{s\to e}^{t2b}(\mathbf{x})\cdot\sum_{\mathbb{N}_2}w_2}, \quad (8)$$

where $\mathbf{x}$ denotes a pixel coordinate, and $\mathbf{y}_1, \mathbf{y}_2$ are neighbors of $\mathbf{x}$. The neighbors are defined as $\mathbf{y}_1 \in \mathbb{N}_1$ with $\mathbb{N}_1 = \{\mathbf{y} | \text{round}\left(\mathbf{y} + \boldsymbol{F}_1^{s2e}(\mathbf{y})\right) = \mathbf{x}\}$, and $\mathbf{y}_2 \in \mathbb{N}_2$ with $\mathbb{N}_2 = \{\mathbf{y} | \text{round}\left(\mathbf{y} + \boldsymbol{F}_1^{e2s}(\mathbf{y})\right) = \mathbf{x}\}$. The round$(\cdot)$ is numerical rounding operator. The Gaussian weights $w_1 = \mathcal{G}(|\mathbf{x} - (\mathbf{y}_1 + \boldsymbol{F}_1^{s2e}(\mathbf{y}_1))|)$ and $w_2 = \mathcal{G}(|\mathbf{x} - (\mathbf{y}_2 + \boldsymbol{F}_1^{e2s}(\mathbf{y}_2))|)$ are depending on the distance between pixel coordinates. Finally, the RS image $\hat{\boldsymbol{I}}_{t2b}$ is reconstructed as

$$\hat{\boldsymbol{I}}_{t2b} = \boldsymbol{T}_{e\to s}^{t2b}\odot\mathcal{B}(\hat{\boldsymbol{I}}_g^{(t_1)};\boldsymbol{F}_{t2b\to g}^{s2e}) + \boldsymbol{T}_{s\to e}^{t2b}\odot\mathcal{B}(\hat{\boldsymbol{I}}_g^{(t_H)};\boldsymbol{F}_{t2b\to g}^{e2s}), \quad (9)$$

where $\mathcal{B}$ is the backwarping operation.

Reconstructing RS Images from Intermediate GS Frame $\hat{\boldsymbol{I}}_g^{(t_m)}$. Aiming to make SelfDRSC be able to generate GS frames at time of arbitrary scanline, we also need to constrain the generated intermediate GS image $\hat{\boldsymbol{I}}_g^{(t_m)}$ in the similar way. The only distinction is that bidirectional distortion warping is divided into two parts, *i.e.*, one is from t_1 to t_m and the other one is from t_m to t_H. According to Eq. (5), we can obtain distortion time map $\boldsymbol{T}_{s\to m}^{t2b}$ and $\boldsymbol{T}_{m\to e}^{t2b}$ as follows

$$\boldsymbol{T}_{s\to m}^{t2b}[i] = \begin{cases} \frac{i-1}{m-1}, & i\in[1,\cdots,m], \\ 1, & i\in[m+1,\cdots,H], \end{cases}$$
$$\boldsymbol{T}_{m\to e}^{t2b}[i] = \begin{cases} 0, & i\in[1,\cdots,m], \\ \frac{i-m-1}{H-m-1}, & i\in[m+1,\cdots,H]. \end{cases} \quad (10)$$

Then we can obtain $\boldsymbol{I}_{t2b}^{s2m}$ and $\boldsymbol{I}_{t2b}^{m2e}$ similar to Eq. (8) and Eq. (9). Finally, we can get the reconstructed RS frame $\hat{\boldsymbol{I}}_{t2b}^{m}$ with time mask $\boldsymbol{U}_{t2b}$

$$\hat{\boldsymbol{I}}_{t2b}^{m} = \boldsymbol{U}_{t2b}\odot\boldsymbol{I}_{t2b}^{s2m} + (\boldsymbol{1} - \boldsymbol{U}_{t2b})\odot\boldsymbol{I}_{t2b}^{m2e}, \quad (11)$$

where $\boldsymbol{U}_{t2b}[i] = \begin{cases} 0, & \text{if } i>m \\ 1, & \text{else} \end{cases}$.

3.3.2 Self-supervised and Self-distillation Losses

We have reconstructed two sets of dual reversed RS images, *i.e.*, $\hat{\boldsymbol{I}}_{t2b}$, $\hat{\boldsymbol{I}}_{b2t}$ and $\hat{\boldsymbol{I}}_{t2b}^{m}$, $\hat{\boldsymbol{I}}_{b2t}^{m}$, and thus a loss function ℓ can be imposed on their corresponding input RS images. The self-supervised loss function is defined as

$$\mathcal{L}_{self} = \mathcal{L}_{se} + \mathcal{L}_{sme}, \quad (12)$$

with

$$\begin{aligned}\mathcal{L}_{se} &= \ell(\hat{\boldsymbol{I}}_{t2b}, \boldsymbol{I}_{t2b}) + \ell(\hat{\boldsymbol{I}}_{b2t}, \boldsymbol{I}_{b2t}),\\ \mathcal{L}_{sme} &= \ell(\hat{\boldsymbol{I}}^{m}_{t2b}, \boldsymbol{I}_{t2b}) + \ell(\hat{\boldsymbol{I}}^{m}_{b2t}, \boldsymbol{I}_{b2t}),\end{aligned} \tag{13}$$

where ℓ is the combination of Charbonnier loss [12] and perceptual loss [11] in our experiments, and the hyper-parameters for balancing them are set as 1 and 0.1.

However, the DRSC model $\mathcal{F}$ trained only with $\mathcal{L}_{self}$ suffers from boundary artifacts (referring to the left part of Fig. 3). This is because boundary pixels are not reliable when reconstructing RS images, and the individual self-supervised loss $\mathcal{L}_{self}$ actually provides invalid supervision for boundary regions. Luckily, we find that the center part of corrected GS images is free from artifacts, and we further propose to introduce self-distillation into loss function for mitigating boundary artifacts. In particular, the training of SelfDRSC is performed in multiple stages, where $\mathcal{L}_{self}$ is first adopted to train $\mathcal{F}$ in Stage 1, and in the following stages, a self-distillation loss [8] is added to cooperate with $\mathcal{L}_{self}$. As shown in Fig. 4, by taking Stage 2 as an example, $\mathcal{F}$ from Stage 1 has been able to generate GS images having high-quality center regions. We assume generated GS images $\{\hat{\boldsymbol{I}}^{(t_1)}_{g,1}, \hat{\boldsymbol{I}}^{(t_m)}_{g,1}, \hat{\boldsymbol{I}}^{(t_H)}_{g,1}\}$ from Stage 1 have spatial size $H \times H$. When training $\mathcal{F}$ at Stage 2, we adopt a cropping operation to extract the center region of input RS images, and $\{\hat{\boldsymbol{I}}^{(t_1)}_{g,1}, \hat{\boldsymbol{I}}^{(t_m)}_{g,1}, \hat{\boldsymbol{I}}^{(t_H)}_{g,1}\}$ are also cropped in the same way to serve as pseudo GS images

$$\mathcal{L}_{sd} = \sum_{\iota \in \{t_1, t_m, t_H\}} \ell(\hat{\boldsymbol{I}}^{(\iota)}_{g,2}, \mathcal{C}(\hat{\boldsymbol{I}}^{(\iota)}_{g,1})), \tag{14}$$

where $\mathcal{C}$ is boundary cropping operation with size p, and the generated GS images $\{\hat{\boldsymbol{I}}^{(t_1)}_{g,2}, \hat{\boldsymbol{I}}^{(t_m)}_{g,2}, \hat{\boldsymbol{I}}^{(t_H)}_{g,2}\}$ in Stage 2 have spatial size $(H-2p) \times (H-2p)$. In SelfDRSC, the final loss function $\mathcal{L}$ is defined as

$$\mathcal{L} = \begin{cases} \mathcal{L}_{self}, & \text{when stage } n = 1 \\ \mathcal{L}_{self} + \mathcal{L}_{sd}, & \text{when stage } n = 2, ..., N \end{cases} \tag{15}$$

and we empirically find that $N = 2$, *i.e.*, self-distillation one time, is sufficient to mitigate boundary artifacts.

4. Experiments

In this section, our SelfDRSC is evaluated on synthetic and real-world RS images, and more video results can be found from the link ***Video Results***. We also provide an implementation using Mindspore at https://github.com/Hunter-Will/SelfDRSC-mindspore.

4.1. Datasets

4.1.1 Synthetic Dataset

The synthetic dataset RS-GOPRO from IFED [30] is used for quantitatively evaluating the competing methods. The training, validation and testing sets are randomly split to have 50, 13 and 13 sequences. For training and testing in IFED [30], 9 GS images are used as ground-truth. We note that ground-truth is not entirely captured by a GOPRO camera, but is partially synthesized by video interpolation methods, possibly yielding over-smoothed details as shown in Fig. 5. Thus full-reference image assessment (FR-IQA) metrics are not very trustworthy for this task, where the result by IFED has less textures than ours, but is better in FR-IQA metrics.

GT (PSNR↑/SSIM↑/LPIPS↓) IFED (22.23/0.711/0.105) Ours (20.01/0.656/0.131)

Figure 5. Interpolated GT frames yield over-smoothed details. The result by IFED has better FR-IQA metrics than Ours, but it has less texture details.

4.1.2 Real-world Testing Set

For testing in real-world, Zhong *et al.* [30] built a dual-RS image acquisition system, which consists of a beam-splitter and two RS cameras with dual reversed scanning patterns. Each sample includes two RS distorted images with reversed distortions but without ground-truth.

4.2. Implementation Details

Our SelfDRSC is trained in $N = 2$ stages, where self-distillation is introduced in Stage 2. The training is done in 205K iterations for Stage 1 and in 115K iterations for Stage 2. The optimization is implemented using AdamW [18] optimizer (β_1=0.9, β_2=0.999) and the initial learning rate is 1×10^{-4}, which is gradually reduced to 1×10^{-6} with the cosine annealing [17]. In Stage 1, patch size is 256×256, and batch size is 48. In Stage 2, patch size is increased to 320×320 considering that the boundary cropping size is $p = 32$, and batch size is 32. During training, the intermediate time t_m is randomly sampled between start time t_1 and end time t_H with sampling interval $(t_H - t_1)/8$.

4.3. Comparison with State-of-the-art Methods

To keep consistent quantitative evaluation with IFED [30], the output of competing methods should have 9 GS images. Thus, our SelfDRSC is compared with methods from two categories. (i) The first category contains cascade methods, where RS correction method DUN [15] for generating one GS image and a VFI model RIFE [9] is adopted for interpolating 8 GS images. Both of them are re-trained on RS-GOPRO dataset for a fair comparison. In Table 1, both cascade orders are considered, *i.e.*, DUN+RIFE and RIFE+DUN. (ii) The second category contains RS correction methods with multiple output images. There are only two works by adopting the dual reversed RS correction setting, *i.e.*, IFED [30] and Albl *et al.* [1]. Since the source

	Method	Inference time (s)	Parameters	NIQE↓ / NRQM↑ / PI↓	PSNR↑ / SSIM↑ / LPIPS↓	
Full-supervised	DUN+RIFE	0.638	14.62M	3.836 / 6.486 / 3.433	23.597 / 0.7653 / 0.1670	
	RIFE+DUN	4.078	14.62M	3.844 / 6.398 / 3.487	20.012 / 0.6520 / 0.1781	
	RSSR	0.976	26.03M	**3.293** / 6.826 / 2.963	22.729 / 0.7283 / 0.1026	
	CVR	2.088	42.70M	3.667 / 6.420 / 3.348	24.816 / 0.7804 / 0.0738	
	IFED	0.177	29.86M	3.657 / 6.405 / 3.358	**30.681 / 0.9121 / 0.0453**	(*Upper Bound)
Self-supervised	SelfDRSC (Ours)	0.182	28.75M	3.297 / **6.933 / 2.896**	28.704 / 0.8886 / 0.0546	
Ground-truth	—	—	—	3.506 / 6.639 / 3.159	—	

Table 1. Quantitative comparison on RS-GOPRO dataset [30], where the metrics are computed based on 9 GS images for a testing case. *Considering that our SelfDRSC has the similar DRSC network with IFED [30], these FR-IQA metrics PSNR, SSIM and LPIPS by IFED can be treated as the upper bound. Also, interpolated GS images by VFI method [9] may appear in 9 ground-truth GS images on RS-GOPRO dataset, *i.e.*, ground-truth GS images for computing FR-IQA metrics may not be truly captured by a GOPRO camera, making FR-IQA metrics not so reliable that we suggest further referring to NR-IQA metrics and visual comparison.

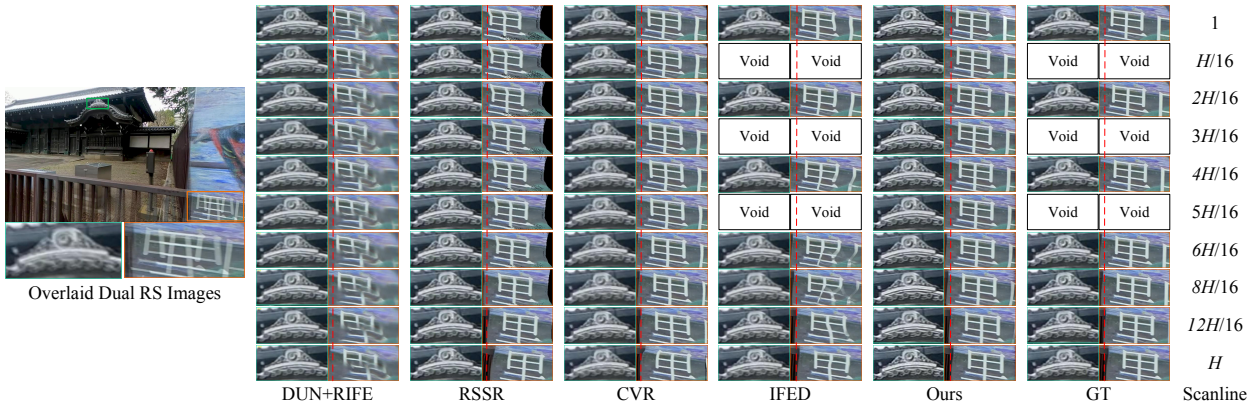


Figure 6. Visual results on RS-GOPRO. IFED [30] can only generate a GS video with 9 frames, since 9 ground-truth GS images from RS-GOPRO are used to train DRSC network in a supervised manner. Our SelfDRSC is able to generate GS videos with higher framerate. In this case, 17 GS frames are generated by SelfDRSC. It will be better viewed by zooming in.

code or experiment results of Albl *et al*. [1] are not publicly available, it is not included into comparison. Besides, we take RSSR [6] and CVR [7] into comparison, which are developed for correcting RS distortions from consecutive two RS images with only top-to-bottom scanning. For a fair comparison, they are re-trained based on the RS-GOPRO dataset. As for quantitative metrics, both FR-IQA (*i.e.*, PSNR, SSIM [26] and LPIPS [29]) and NR-IQA (*i.e.*, NIQE [21], NRQM [19] and PI [4]) metrics are employed to evaluate the competing methods.

4.3.1 Results on Synthetic Dataset

Table 1 reports quantitative comparison, where the IQA metrics are computed based on 9 GS images for each testing case. Our SelfDRSC, without any ground-truth high framerate GS images during training, achieves better FR-IQA metrics than the supervised methods except IFED. Considering that our SelfDRSC has similar DRSC network architecture with IFED [30], these FR-IQA metrics PSNR, SSIM and LPIPS by IFED can be treated as the upper bound of self-supervised DRSC methods on RS-GOPRO dataset. Besides, as mentioned above, interpolated frames by VFI method may appear in 9 ground-truth GS images for calculating FR-IQA values. And thus FR-IQA metrics may not be so reliable to indicate correction performance (please see Fig. 5). We suggest further referring to NR-IQA metrics and visual results for evaluating the competing methods.

Besides FR-IQA metrics, NR-IQA metrics are also employed as reference for quantitative evaluation, and more visual results especially videos are provided for comprehensive justification. We note that our SelfDRSC is better than ground-truth GS images in terms of NR-IQA values, since they may be interpolated using VFI method rather than original GS frames captured by GOPRO camera. As for the visual quality, reconstructed GS images by competing methods are presented in Fig. 6. IFED [30] can only generate a GS video with 9 frames. Our SelfDRSC is able to generate GS videos with higher framerate. In this case, 17 GS frames are generated by SelfDRSC, and are better corrected with finer textures. Also, our SelfDRSC ranks as top-2 efficient method in terms of inference time.

4.3.2 Results on Real-world Data

In Fig. 7, we provide an example on real-world dual reversed RS images. It is an animated figure to compare the video results by competing methods. RSSR and CVR cannot generalize well to real-world motion, and IFED can

Figure 7. Video results on real-world RS data. The results can be displayed in an animated figure in the ***arXiv*** version.

	NIQE↓ / NRQM↑ / PI↓	PSNR↑ / SSIM↑ / LPIPS↓
Optical flow estimation in $\mathcal{W}$:		
GMFlow-*kitti*	3.221 / 6.958 / 2.845	25.314 / 0.8080 / 0.0762
GMFlow-*sintel*	3.244 / 6.983 / 2.845	25.670 / 0.8221 / 0.0763
RIFE-Flow	3.570 / 6.735 / 3.144	25.539 / 0.8105 / 0.0653
PWC-Net	3.304 / 6.931 / 2.902	28.411 / 0.8848 / 0.0574
Warping strategy in $\mathcal{W}$:		
L.A. & S.	3.303 / 6.909 / 2.915	26.267 / 0.8466 / 0.0606
F. R. & B.	3.303 / 6.907 / 2.911	27.175 / 0.8665 / 0.0548
CFR & B.	3.304 / 6.931 / 2.902	28.411 / 0.8848 / 0.0574
Loss function in $\mathcal{L}$:		
$\mathcal{L}_{se}$	11.261 / 7.079 / 6.783	20.643 / 0.4575 / 0.1888
$\mathcal{L}_{sme}$	3.246 / 6.871 / 2.909	23.241 / 0.7514 / 0.0920
$\mathcal{L}_{self}$	3.304 / 6.931 / 2.902	28.411 / 0.8848 / 0.0574
$\mathcal{L}_{self} + \mathcal{L}_{sd}$	3.297 / 6.933 / 2.896	28.704 / 0.8886 / 0.0546

Table 2. Ablation results of SelfDRSC on RS-GOPRO dataset.

largely correct RS distortions but it is not as good as our SelfDRSC in terms of both textures and temporal consistency.

4.4. Ablation Study

To demonstrate the effectiveness of different optical flow estimation methods, warping ways, and loss functions. We implement ablation study on these elements. All the IQA metrics are computed based on 9 GS images. (i) For optical flow estimation method in $\mathcal{W}$, we use pre-trained RIFE-flow [9], PWC-Net [25], GMFlow-*kitti* and GMFlow-*sintel* [27]. We train all these variants with the same warping strategy and without $\mathcal{L}_{sd}$ for fair comparison on RS-GOPRO dataset. From Table 2, PWC-Net achieves the best performance on self-supervised dual reversed rolling shutter correction task. Hence, we use PWC-Net as our optical flow estimation method in the following. (ii) For the ways of distortion warping in $\mathcal{W}$, we use three combinations, *i.e.*, linear approximation [10] & splatting [22], flow reversal [28] & backwarping and CFR [24] & backwarping, which are abbreviated as L.A. & S., F.R. & B. and CFR & B., respectively. From Table 2, the third strategy gets the best performance since it combines the advantages of both linear approximation and flow reversal. (iii) We also train our method with different loss functions. We find that $\mathcal{L}_{sd}$ not only achieves about *+0.3dB* PSNR gains but also alleviates boundary artifacts in visual results in Fig. 3. Moreover, we verify that $\mathcal{L}_{se}$ and $\mathcal{L}_{sme}$ are both very important for self-supervised RS correction. For the update strategy of teacher model in self-distillation, we implement different momentum coefficients $c \in \{0.9, 0.99, 0.999, 1\}$, and we also set different boundary cropping sizes $p \in \{16, 32, 64\}$, and different training stage numbers N in the the arXiv version.

5. Conclusion

In this paper, we proposed a self-supervised learning framework for correcting dual reversed RS distortions, and high framerate GS videos can be generated by our SelfDRSC. In SelfDRSC, a novel bidirectional distortion warping module is proposed to obtain reconstructed dual reversed RS images that can be employed as cycle consistency-based supervision. The self-supervised learning loss with self-distillation is proposed for training DRSC network, where self-distillation is effective in mitigating boundary artifacts in generated GS images. Extensive experiments have been conducted to validate the effectiveness and generalization ability of our SelfDRSC on both synthetic and real-world data.

Acknowledgements

This work was supported in part by the National Key Research and Development Project (2022YFA1004100), the National Natural Science Foundation of China (62172127 and U22B2035), the Natural Science Foundation of Heilongjiang Province (YQ2022F004), the Hong Kong ITC Innovation and Technology Fund (9440288), and the CAAI-Huawei MindSpore Open Fund.

References

[1] Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, and Konrad Schindler. From two rolling shutters to one global shutter. In *CVPR*, pages 2505–2513, 2020. 1, 2, 3, 6, 7

[2] Cenek Albl, Zuzana Kukelova, and Tomas Pajdla. R6p-rolling shutter absolute camera pose. In *CVPR*, pages 2292–2300, 2015. 1

[3] Haoran Bai and Jinshan Pan. Self-supervised deep blind video super-resolution. *ArXiv preprint arXiv:2201.07422*, 2022. 2

[4] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *ECCV Workshops*, pages 0–0, 2018. 7

[5] Huaijin Chen, Jinwei Gu, Orazio Gallo, Ming-Yu Liu, Ashok Veeraraghavan, and Jan Kautz. Reblur2deblur: Deblurring videos via self-supervised learning. In *ICCP*, pages 1–9, 2018. 2

[6] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video. In *ICCV*, pages 4228–4237, 2021. 1, 2, 7

[7] Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *CVPR*, pages 17572–17582, 2022. 2, 7

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, 33:21271–21284, 2020. 6

[9] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *ECCV*, pages 1–16, 2022. 5, 6, 7, 8

[10] 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 *CVPR*, pages 9000–9008, 2018. 5, 8

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016. 6

[12] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE TPAMI*, 41(11):2599–2613, 2018. 6

[13] Yizhen Lao and Omar Ait-Aider. Rolling shutter homography and its applications. *IEEE TPAMI*, 43(8):2780–2793, 2020. 1

[14] Dave Litwiller. CCD vs. CMOS. *Photonics spectra*, 35(1):154–158, 2001. 1

[15] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *CVPR*, pages 5941–5949, 2020. 1, 2, 6

[16] Peidong Liu, Joel Janai, Marc Pollefeys, Torsten Sattler, and Andreas Geiger. Self-supervised linear motion deblurring. *IEEE Robotics and Automation Letters*, 5(2):2475–2482, 2020. 2

[17] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *ArXiv preprint arXiv:1608.03983*, 2016. 6

[18] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ArXiv preprint arXiv:1711.05101*, 2017. 6

[19] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *CVIU*, 158:1–16, 2017. 7

[20] Marci Meingast, Christopher Geyer, and Shankar Sastry. Geometric models of rolling-shutter cameras. *ArXiv preprint cs/0503076*, 2005. 1, 2

[21] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7

[22] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, pages 5437–5446, 2020. 8

[23] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *CVPR*, pages 3341–3350, 2020. 2

[24] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: Extreme video frame interpolation. In *ICCV*, pages 14489–14498, 2021. 5, 8

[25] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, pages 8934–8943, 2018. 5, 8

[26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[27] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. GMFlow: Learning optical flow via global matching. In *CVPR*, pages 8121–8130, 2022. 5, 8

[28] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *NeurIPS*, 32:1–10, 2019. 5, 8

[29] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[30] Zhihang Zhong, Mingdeng Cao, Xiao Sun, Zhirong Wu, Zhongyi Zhou, Yinqiang Zheng, Stephen Lin, and Imari Sato. Bringing rolling shutter images alive with dual reversed distortion. In *ECCV*, pages 233–249, 2022. 1, 2, 3, 4, 6, 7

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 2

[32] Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *CVPR*, pages 4551–4560, 2019. 1, 2