SUNet: Symmetric Undistortion Network for Rolling Shutter Correction –Supplementary Material–

Bin Fan Yuchao Dai* Mingyi He
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
binfan@mail.nwpu.edu.cn, daiyuchao@nwpu.edu.cn, myhe@nwpu.edu.cn

Abstract

In this supplementary material, we first provide additional RS correction results. Furthermore, the effectiveness of our approach in aggregating the contextual cues of two consecutive RS images is demonstrated. More ablation studies of the training loss are reported. We also include a video to verify the advantages of our pipeline against the state-of-the-art method [3] in RS sequence correction. Finally, we summarize the details of our network architecture.

1. More Experimental Results

We provide more qualitative comparisons in Fig. 1. Compared with the off-the-shelf algorithms (*e.g.* [5, 6, 3]) for RS correction, our pipeline can produce more reliable results. Specifically, the higher-quality GS images with richer details are restored, even for foreground objects with more severe RS distortion.

We also give quantitative results using the Learned Perceptual Image Patch Similarity (LPIPS) [4] metric. The smaller the LPIPS score, the more similar the predicted and the ground truth GS images. As shown in Table 1, one can see our method achieves superior performance compared with the state-of-the-art method: DSUN [3].

Table 1. Quantitative comparisons of the performance between our approach and the state-of-the-art method in terms of the LPIPS metric.

	LPIPS↓	
	Carla-RS	Fastec-RS
DSUN [3]	0.0703	0.1222
SUNet (Ours)	**0.0658**	**0.1205**

2. Necessity and Effectiveness of Contextual Aggregation

In Fig. 1(e)&(f) of the main manuscript, we have shown the forward and backward GS images $\boldsymbol{I}_{1\to g}$ and $\boldsymbol{I}_{2\to g}$ synchronously generated by our network. To better prove that our pipeline can effectively aggregate the contextual cues of two consecutive RS images, we further report the intermediate results of our network in Fig. 2, *i.e.*, forward/backward undistortion flows and GS images. The biggest motivation that utilizing contextual aggregation for RS correction is inspired by this obvious observation, *i.e.*, the first RS image $\boldsymbol{I}_1$ and the second RS image $\boldsymbol{I}_2$ have different contributions to the upper and lower regions of the corresponding ground truth time-centered GS image $\boldsymbol{I}_{GT}$ respectively, which is exemplified in Fig. 2(e)&(f). Through symmetric network design, our improved PWC-based architecture can obtain a plausible RS-aware undistortion flow (*i.e.*, the closer the pixel to the intermediate time τ, the smaller the value of the undistortion flow is generally) for subsequent accurate RS correction.

As manifested in Fig. 4, the limited information of the second RS image is insufficient to restore the rich texture details of the target GS image at intermediate time τ. Intuitively, the rear of the car in the backward GS image marked by the red circle in Fig. 2 (f) is similar to the content restored by DSUN [3] in Fig. 4(d) of the main manuscript. This is because the extra information of the first frame is not used for detail extraction and fusion, which indicates the necessity of **contextual aggregation**. Furthermore, our context-aware cost volume together with the symmetric consistency constraint is proven to be beneficial in effectively aggregating the contextual cues of two consecutive RS images, thereby resulting in high-quality time-centered GS images (at time τ) with more complete visual content.

3. An additional SfM Example

Here, we provide an example of 3D reconstruction to evaluate the performance of the original RS image and our corrected GS image in SfM. As shown in Fig. 3, direct use of the original RS image leads to erroneous and distorted 3D geometry, while our method can effectively remove RS artifacts and reconstruct an accurate and consistent 3D scene structure. For instance, the column becomes vertical and in the correct 3D position after being corrected by our method.

4. Ablation Study on the Training Loss

We report the impact of different combinations of the loss terms in training our model, as shown in Table 2. Note that removing the perceptual loss $\mathcal{L}_p$ will cause the corrected image to appear blurred effect, and the reconstruction loss $\mathcal{L}_r$ is particularly important. Our total loss function yields the best model, which facilitates better removal of RS artifacts to produce high-quality results.

Table 2. Effectiveness of different combinations of training losses.

	PSNR↑			SSIM↑	
	CRM	CR	FR	CR	FR
w/o $\mathcal{L}_r$	28.00	27.90	27.29	0.83	0.81
w/o $\mathcal{L}_p$	29.08	28.95	28.20	**0.85**	**0.84**
w/o $\mathcal{L}_c$	29.05	28.94	27.89	0.84	0.82
w/o $\mathcal{L}_s$	29.19	28.07	28.15	**0.85**	0.83
full loss	**29.28**	**29.18**	**28.34**	**0.85**	**0.84**

5. Demo Video

To evaluate the performance of our proposed pipeline on video sequences, we further utilize the RS image dataset from [2], where each sequence consists of 12 consecutive RS frames with significant image distortions. Additionally, we simulate RS image sequences in the autonomous driving environment by using the *Carla* simulator [1]. Based on these unseen scenarios, we compare our approach with DSUN [3] to evaluate the generalization ability of the model. The RS correction video results are included in "Demo_Video.mp4". We refer the readers to the video for the full results, and two screenshots are shown in Fig. 5. It is obvious that, in comparison with our proposed SUNet, DSUN [3] always fails to recover the texture on the ground. Also, combining with Fig. 6, one can further see that our approach obtains more satisfactory RS correction results with its excellent generalization ability.

6. Network Details

Fig. 7 displays the architecture of the 6-level feature pyramid extractor network. Note that the bottom level indicates the original input RS images. Fig. 8 illustrates the undistortion flow estimator network of $\boldsymbol{I}_1$ at the 4-th pyramid level. The optical flow estimator networks at other levels have similar structures but different feature channels. Note also that the top level did not adopt the upsampled undistortion flows and calculated the cost volume using the pyramid features of the first and second RS images directly.

References

[1] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: an open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2

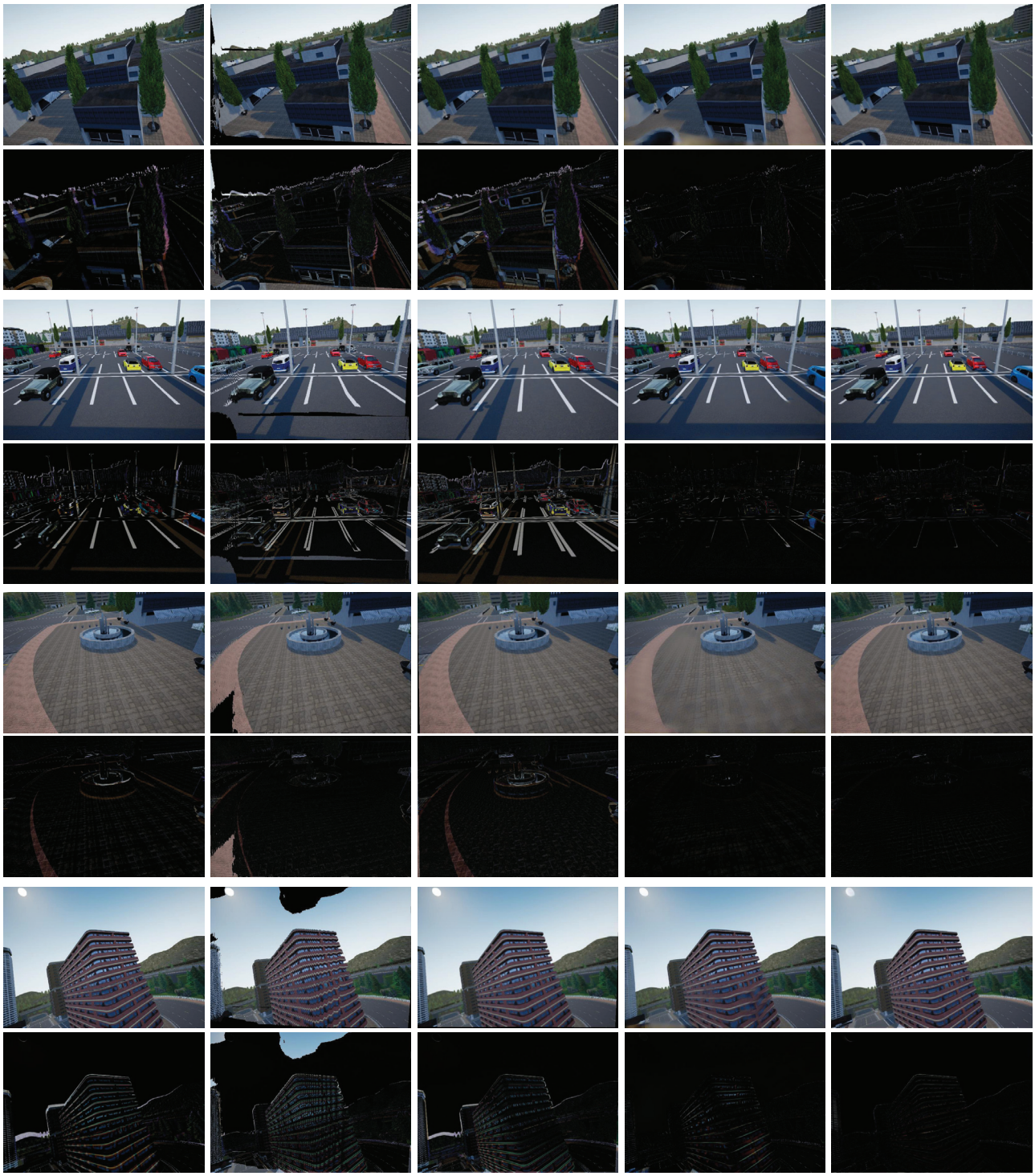
[2] Per-Erik Forssén and Erik Ringaby. Rectifying rolling shutter video from hand-held devices. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 507–514. IEEE, 2010. 2, 6, 7

[3] Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5941–5949. IEEE, 2020. 1, 2, 3, 5, 6

[4] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 586–595. IEEE, 2018. 1

[5] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of the International Conference on Computer Vision*, pages 948–956. IEEE, 2017. 1, 3

[6] 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 Conference on Computer Vision and Pattern Recognition*, pages 4551–4560. IEEE, 2019. 1, 3



(a) Original RS image 2 (b) Zhuang *et al.* [5] (c) Zhuang *et al.* [6] (d) Liu *et al.* [3] (e) Ours

Figure 1. Qualitative results against baseline methods. Even rows: absolute difference between the corresponding image and the ground truth GS image. (b-e) GS images predicted by Zhuang *et al.* [5], Zhuang *et al.* [6], Liu *et al.* [3], and our approach, respectively.

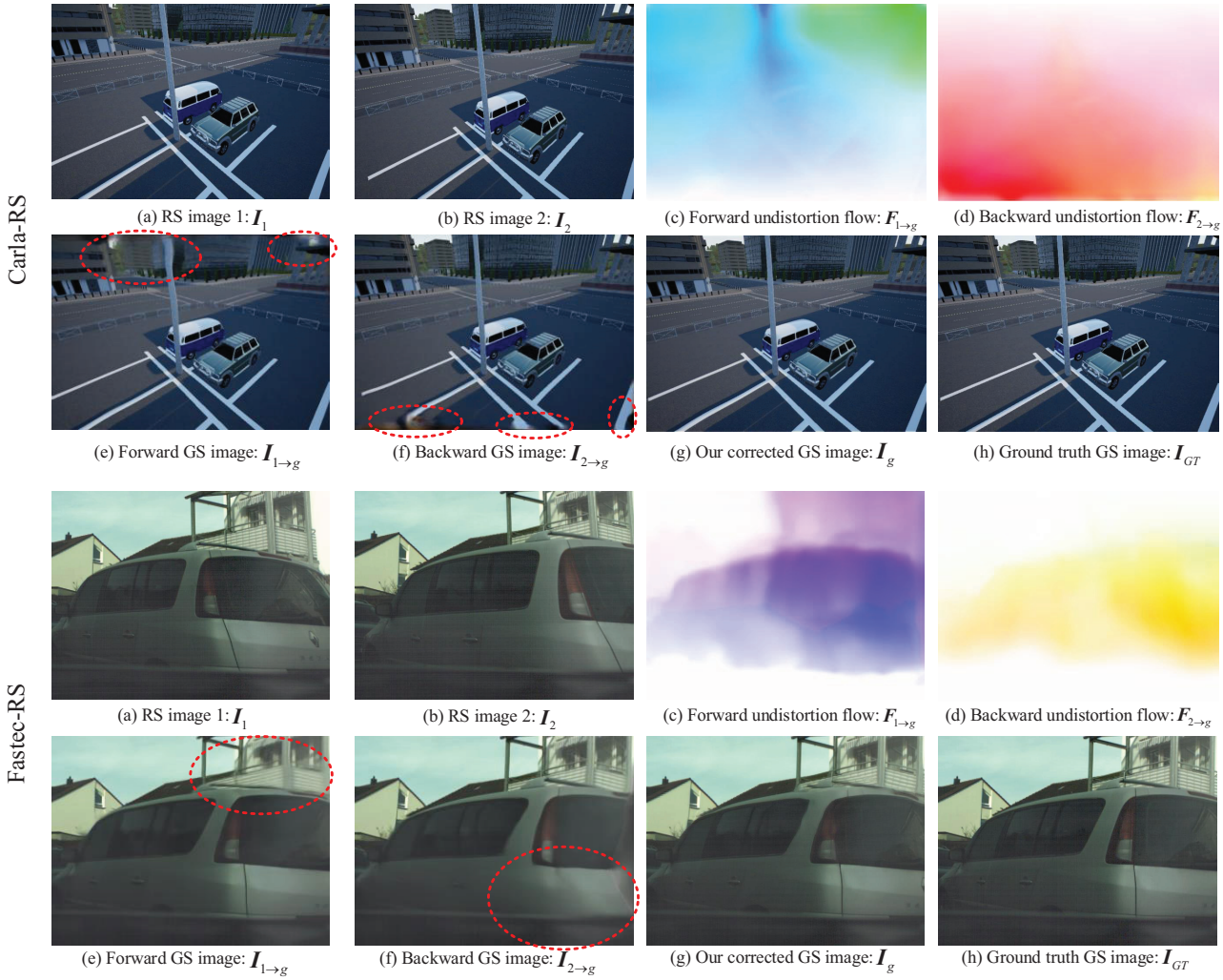


Figure 2. Intermediate examples of our network, including forward/backward undistortion flows and GS images. The upper two rows and the lower two rows show examples of Carla-RS and Fastec-RS datasets, respectively. Inputting two consecutive RS images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, our approach estimates the forward and backward undistortion flows $\boldsymbol{F}_{1\to g}$ and $\boldsymbol{F}_{2\to g}$ to predict the forward and backward GS images $\boldsymbol{I}_{1\to g}$ and $\boldsymbol{I}_{2\to g}$, which then are aggregated to produce a time-centered corrected GS image $\boldsymbol{I}_g$ as the ground truth GS image $\boldsymbol{I}_{GT}$. Note that the undistortion flow of a pixel closer to the intermediate time of two frames appears as a lighter color (i.e., smaller values), such as the last rows of (c) and the beginning rows of (d), which accounts for the basic scanline-dependent characteristics of the undistortion flows.

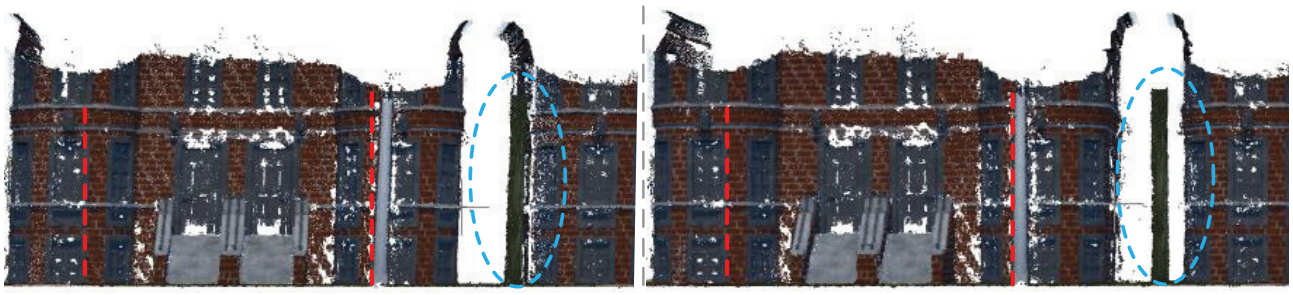


Figure 3. A 3D reconstruction example using the original RS images and our corrected GS images, respectively.

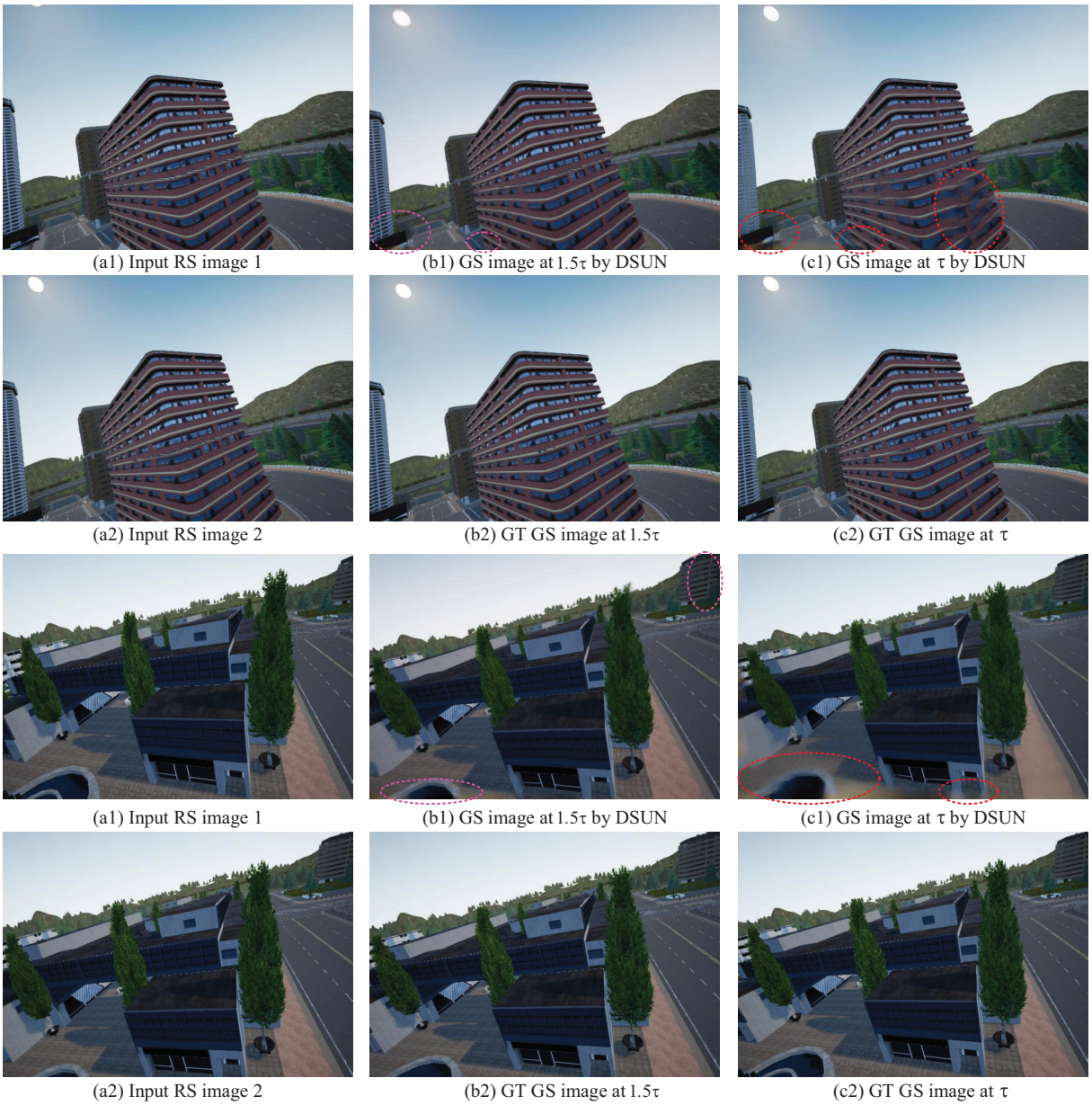


Figure 4. Two examples of vulnerability when generalizing DSUN [3] to estimate the time-centered GS image at time τ. One can see that there are obvious differences between the latent GS images at $\frac{3\tau}{2}$ and τ in (b2) and (c2), and the GS image at $\frac{3\tau}{2}$ is relatively similar to the second RS image. DSUN cannot well recover the details of the GS image (c1) corresponding to time τ (see red circles), although it can obtain a plausible GS image (b1) corresponding to time $\frac{3\tau}{2}$ (In fact it also contains missing content that is not particularly striking, see pink circles). This is because the recovery of the GS image at time τ is more challenging than that at time $\frac{3\tau}{2}$. At this more challenging task, only the limited information of the second frame image is insufficient, and the context information must be fully utilized. Note that due to the effective context aggregation based on the imaging characteristics of consecutive RS images, our method can reconstruct the rich details of the latent GS image at time τ, which is also the main contribution of our method.

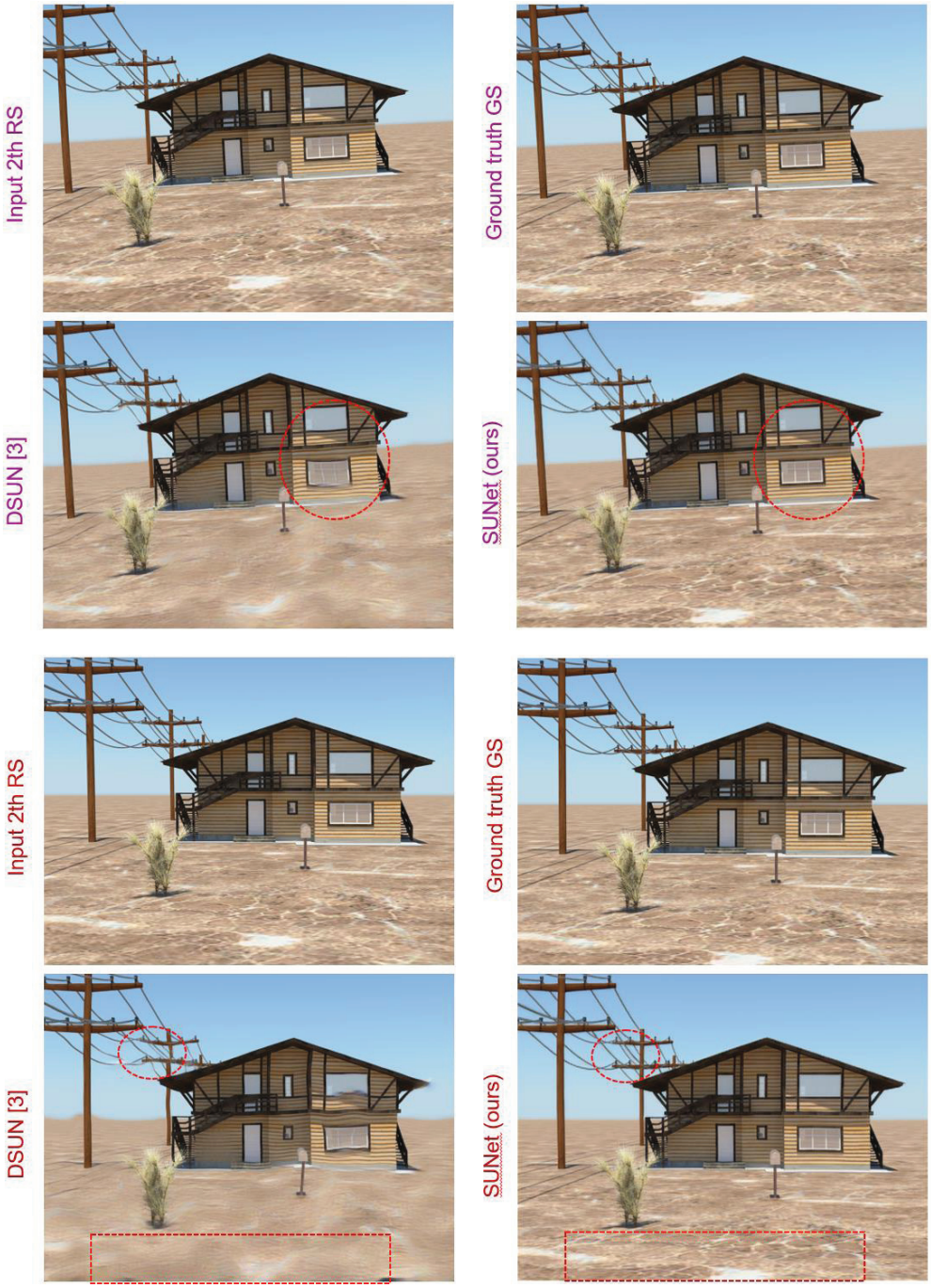


Figure 5. Two examples extracted from the demo video. DSUN [3] performs worse when dealing with unseen scenes provided by [2], while our proposed SUNet has excellent generalization performance to produce a coherent video with more detailed textures and fewer ghosting artifacts.

Figure 6. An example result from an RS sequence [2] captured by a fast-moving iPhone 3GS camera in the real world.

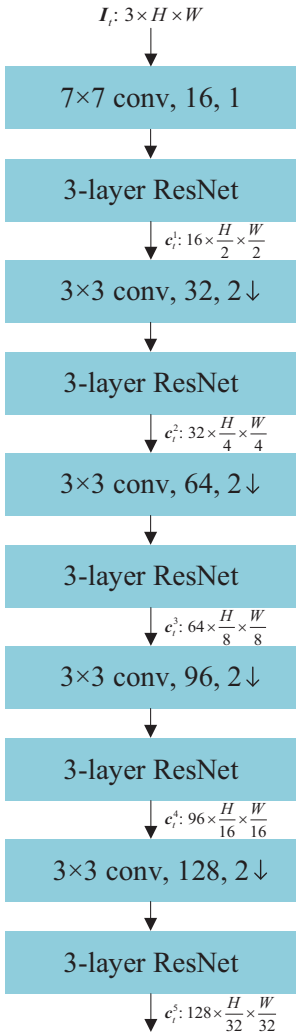


Figure 7. The feature pyramid extractor network. The first RS image $\boldsymbol{I}_1$ and the second RS image $\boldsymbol{I}_2$ are encoded using the same network. The convolutional layer and the $\times 2$ downsampling layer at each level is implemented using a single convolutional layer with a stride of 2, followed by a ReLU. Each ResNet layer contains two sequential blocks consisting of: a 2D convolution with a 3×3 kernel, a ReLU and a 2D convolution. $\boldsymbol{c}_t^l$ denotes extracted features of RS image t at level l.

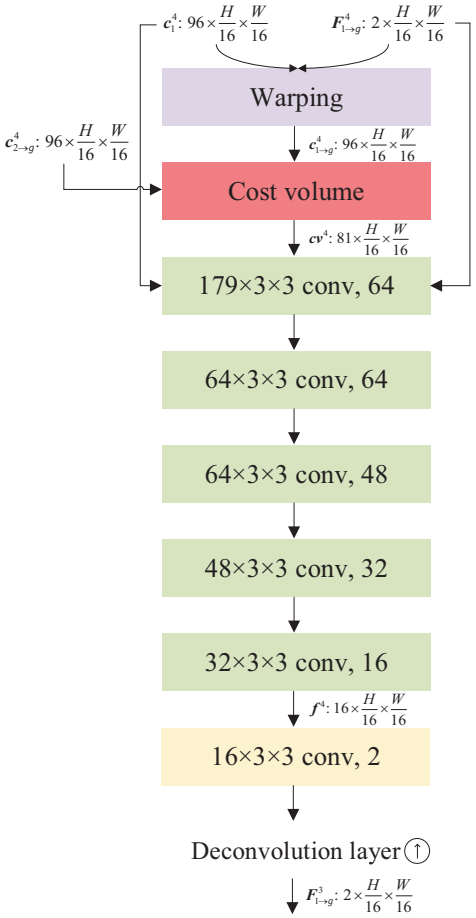


Figure 8. The undistortion flow estimator network of the first RS image $\boldsymbol{I}_1$ at pyramid level 4. Inspired by the DenseNet connections, each convolutional layer is followed by a ReLU except the last (yellow) one that outputs the undistortion flow. Deconvolution is then performed to return an upsampled undistortion flow $\boldsymbol{F}_{1\to g}^3$ for subsequent processing.