Zero-Shot Dual-Lens Super-Resolution
——Supplementary Material——

Ruikang Xu Mingde Yao Zhiwei Xiong
University of Science and Technology of China
{xurk, mdyao}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn

The supplementary material is organized as follows:

- Section 1 provides additional analysis on the alignment process.
- Section 2 provides more details of synthesized data generation.
- Section 3 provides more implementation details.
- Section 4 provides more details and additional comparison results on the real-world datasets.

1. Additional Analysis on Alignment

1.1. How the frequency information of the LR image changes during warping?

To generate the aligned HR-LR image pair, we warp the LR view toward the HR view by downsampling the HR view as a bridge (*e.g.*, using the bilinear operator). This downsampled HR view will contain different frequency information from the original LR view, since the predefined degradation cannot match the realistic one. Generally, the information gap between HR and downsampled HR will be smaller than that between HR and original LR, since the realistic degradation tends to be more severe than the predefined one. Therefore, only applying the photometric consistency (L_2) loss for alignment will inevitably introduce additional information in the warped LR image, as shown in Figure 1. This results in misguided SR model training since the realistic degradation will be pushed towards the predefined degradation (bilinear in this case). To keep the realistic degradation during warping, we propose to constrain the alignment process in spatial, frequency, and feature domains simultaneously.

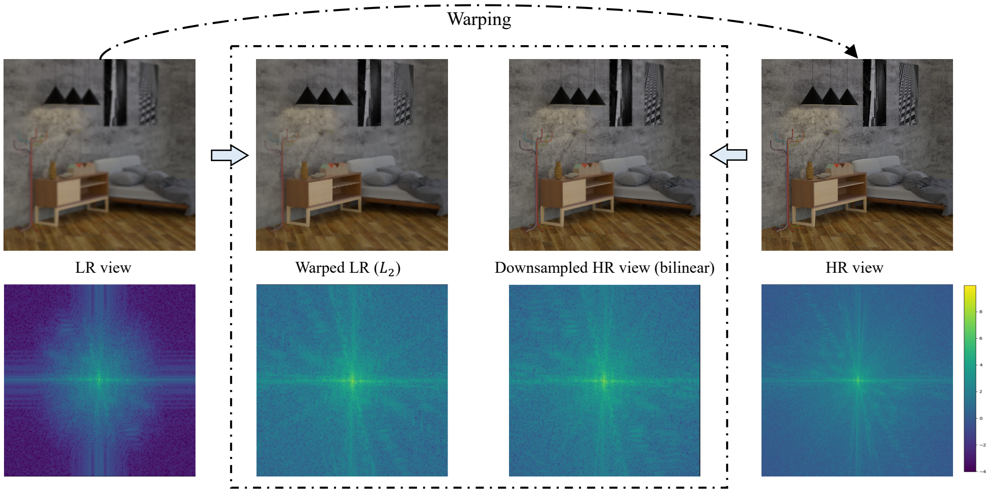


Figure 1. Analysis on the frequency domain during the alignment process.

1.2. How about warping HR toward LR?

As the alternative approach to generate the aligned HR-LR image pair, one can also warp the HR view to the LR view by upsampling the LR view as a bridge (*e.g.*, using the bilinear operation), while constraining the alignment process in spatial, frequency, and feature domains simultaneously. Compared with LR-to-HR warping, HR-to-LR warping may result in more severe frequency change in the warped HR image, and thus limited performance of super-resolution. We perform the ablation study on the HCI_new dataset with isotropic Gaussian downsampling for 2× and 4× super-resolution, where only the input of the alignment network is replaced. Experiment results are shown in Table 1, which validates the above argument.

Table 1. Ablation on the alignment direction.

Warping	2×		4×	
	PSNR	SSIM	PSNR	SSIM
HR-to-LR	30.45	0.8327	28.23	0.7335
LR-to-HR	31.01	0.8529	28.87	0.7536

2. Synthesized Data Generation

2.1. Dual-lens Pair Generation

- **Middlebury2021.** The Middlebury2021 stereo dataset [9] is captured by a mobile device (Apple iPod touch 6G) mounted on a robot arm. It contains 24 samples with a resolution of 1920 × 1080. We randomly select 6 samples for testing and the remaining samples for training (for methods using external training data). We downsample the left view with image-specific degradation as the wide-angle image (the original image can be then used as HR groundtruth for evaluation), while the center area of the right view is cropped to simulate the telephoto image. An example is shown in Figure 2.

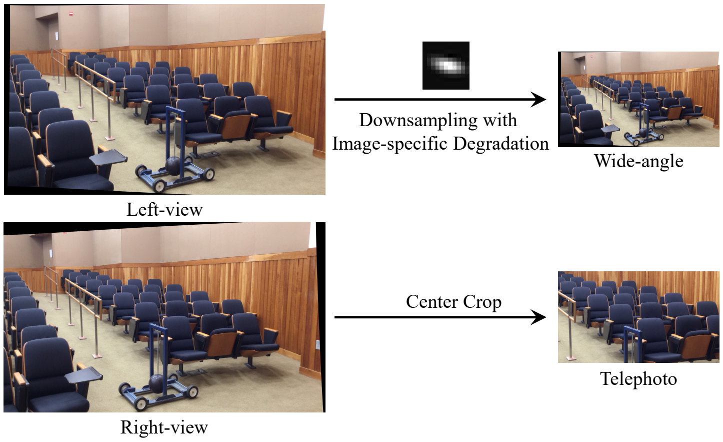


Figure 2. Pipeline of the simulated dual-lens data generation.

- **HCI_new.** The HCI_new light-field dataset [2] is a synthetic dataset including 24 samples with a resolution of 512 × 512. We randomly select 8 samples as the testing set and the remaining samples as the training set (for methods using external training data). Each sample consists of 9 × 9 views, where we downsample the view (3, 3) as the wide-angle image and crop the view (8, 8) as the telephoto image.

2.2. Image-Specific Degradation Generation

We use three groups of image-specific degradation kernels to generate the LR wide-angle image, *i.e.*, isotropic and anisotropic Gaussian downsampling, and isotropic Gaussian downsampling with slight JPEG compression. For isotropic Gaussian kernels, following [7, 8, 14], we set the kernel size as 21 for 2× and 4× SR with the length of axes is distributed in $(0.2, 2.0)$ and $(0.2, 4.0)$. For anisotropic Gaussian kernels, following [1, 10], we generate the kernels by a covariance matrix

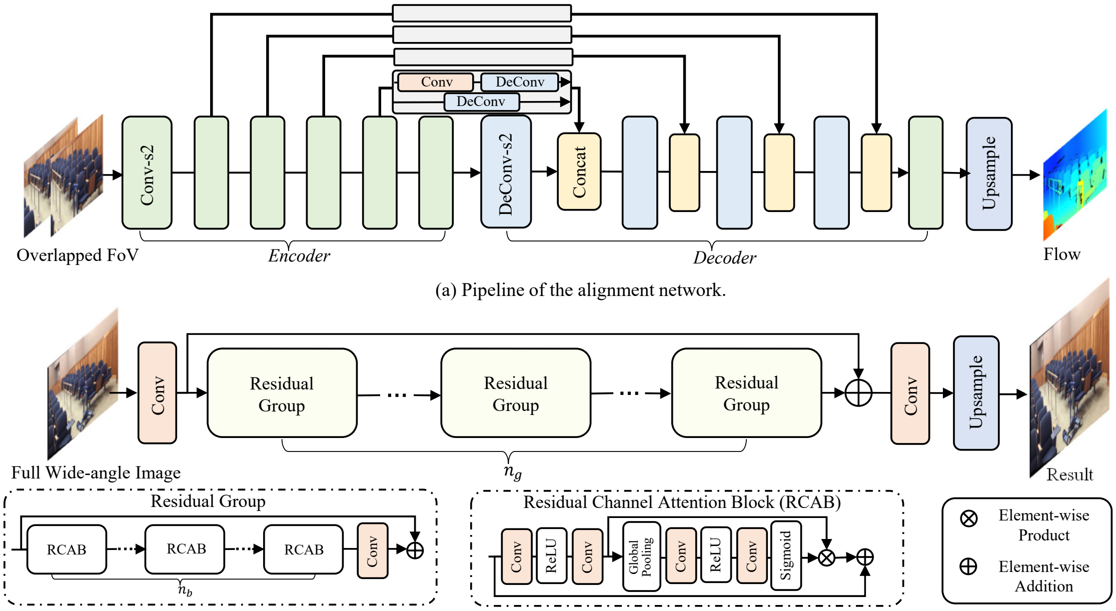
$$\Sigma = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{1}$$

where the length of both axes λ_1, λ_2 are uniformly distributed in $(0.6, 5.0)$, and the random angle θ for rotation are uniformly distributed in $[-\pi, \pi]$. We set the kernel size as 11 and 31 for 2× and 4× SR. For JPEG compression, we uniformly set the compression quality parameter $r = 75$.

3. Implementation Details

3.1. Network Architecture

For the alignment network, we modify FlowNet-S [3] by decreasing the convolution layers to suit the limited training data, and the network structure is shown in Figure 3 (a). We use the RCAN [15] backbone as the image-specific SR network, and the network structure is shown in Figure 3 (b), where we set the number of residual groups $n_g = 10$ and the number of residual channel attention blocks $n_b = 20$. Note that the embodiments of the alignment and SR networks both can be replaced with other networks.



(a) Pipeline of the alignment network.

(b) Pipeline of the image-specific SR network.

Figure 3. Network architectures of the alignment and image-specific SR.

3.2. Training Details

For training the alignment network, we use the Adam optimizer [5] ($\beta1 = 0.9, \beta2 = 0.99$) to train the model with an initial learning rate 1×10^{-4}. We set $\lambda_1 = 0.005, \lambda_2 = 0.001, \lambda_3 = 0.01$ (the weighting factors in the loss function) and set $m = 10, n = 10$ (the number of used layers in feature contrastive loss). For training the image-specific SR network, we set the batch sizes to 8 and the crop LR patches with size 48×48. We also use the Adam optimizer [5] ($\beta1 = 0.9, \beta2 = 0.99$) with initial learning rate 1×10^{-4}.

3.3. Training Time

The training time of ZeDuSR consists of two parts (i.e., alignment and SR), and mainly depends on the model complexity of the SR backbone. Here we take RCAN [15] (a heavy model) and VDSR [4] (a light one) as two examples. The average training time is shown in Table 2 below using NVIDIA 3090Ti GPU, which is comparable to the representative zero-shot method KernelGAN when both use VSDR. Note that KernelGAN also consists of two stages (i.e., kernel estimation and SR).

Table 2. Training Time on HCI_new for 2× SR.

Method	IG	AG	Training Time
KernelGAN (VDSR)	-	29.78/0.8042	29(s) + 87(s)
ZeDuSR (RCAN)	31.01/0.8529	30.02/0.8146	31(s) + 605(s)
ZeDuSR* (RCAN)	31.17/0.8594	30.23/0.8183	31(s) + 242(s)
ZeDuSR (VDSR)	29.97/0.8522	30.01/0.8122	31(s) + 87(s)
ZeDuSR* (VDSR)	31.11/0.8588	30.15/0.8158	31(s) + 35(s)

4. Additional Comparison Results on Real-world Datasets

4.1. Dataset Details

- **CameraFusion.** CameraFusion [11] is a dual-lens image dataset collected by iPhone11 Pro Max, which has a wide-angle view (from 26mm lens) and a telephoto view (from 52mm lens) and supports 2× SR. The resolution of both two views is 4032 × 3024. This dataset contains 146 samples. We randomly select 26 samples for testing set and the remaining samples as the training set (for methods using external training data).

- **RealMCVSR.** RealMCVSR [6] is a triple-lens video dataset collected by iPhone12 Pro Max, where the ultra-wide view (from 30mm lens) and the telephoto view (from 147mm lens) support 4× SR, while the ultra-wide view and the wide-angle view (from 59mm lens) support 2× SR. The resolution of all three views is 1080 × 1920. We use the first frame of each video clip. This dataset contains 161 samples. We randomly select 16 samples as the testing set and the remaining samples as the training set (for methods using external training data).

4.2. Non-refernece Evaluations

Since we do not have ground-truth in real-world scenarios, we conducted a user study for statistical comparison in our paper. We supplement two state-of-the-art non-reference metrics in Table 3 below, which quantitatively demonstrate the superiority of our method over the main competitors SelfDZSR and DCSR-SRA (ranked 2nd and 3rd in the user study). Note that MANIQA is the winner of the NTIRE@CVPR 2022 Perceptual Image Quality Assessment Challenge.

Table 3. Non-reference MANIQA [12] and PaQ-2-PiQ [13] metrics (both the higher, the better) on the real-world datasets.

Method	iPhone11 (2×)	iPhone12 (2×)	iPhone12 (4×)
DCSR-SRA	0.4725/65.79	0.4816/66.30	0.2900/43.08
SelfDZSR	0.4851/66.51	0.4849/67.09	0.2911/43.26
ZeDuSR	**0.4901/67.03**	**0.4903/67.89**	**0.2927/43.65**

4.3. More Visual Results

More visual results on real-world datasets can be found below.

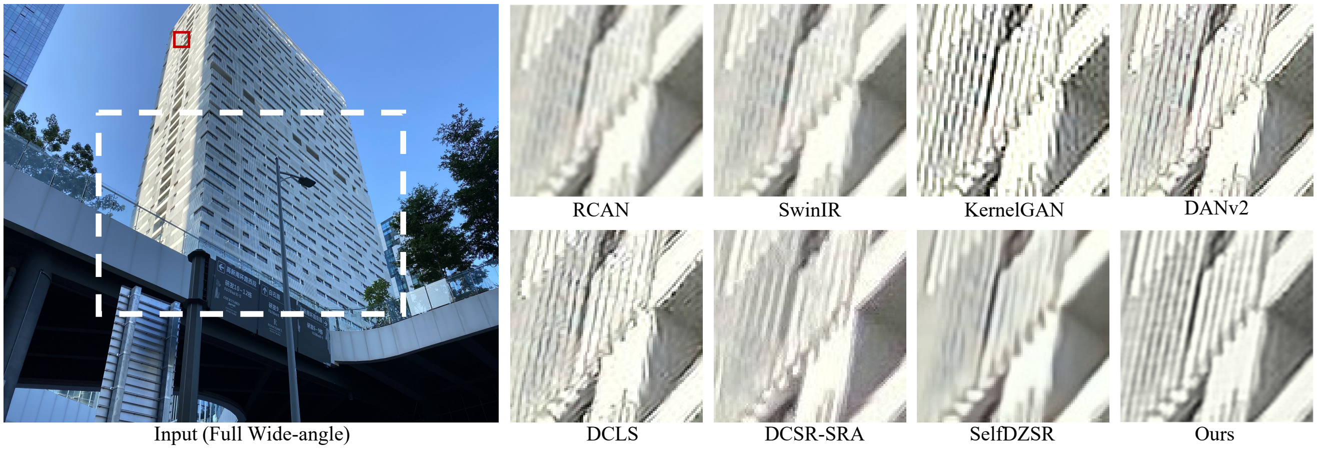


Figure 4. Visual comparisons for 2× SR on real-world data captured by iPhone11. The white dotted box indicates the overlapped FoV.

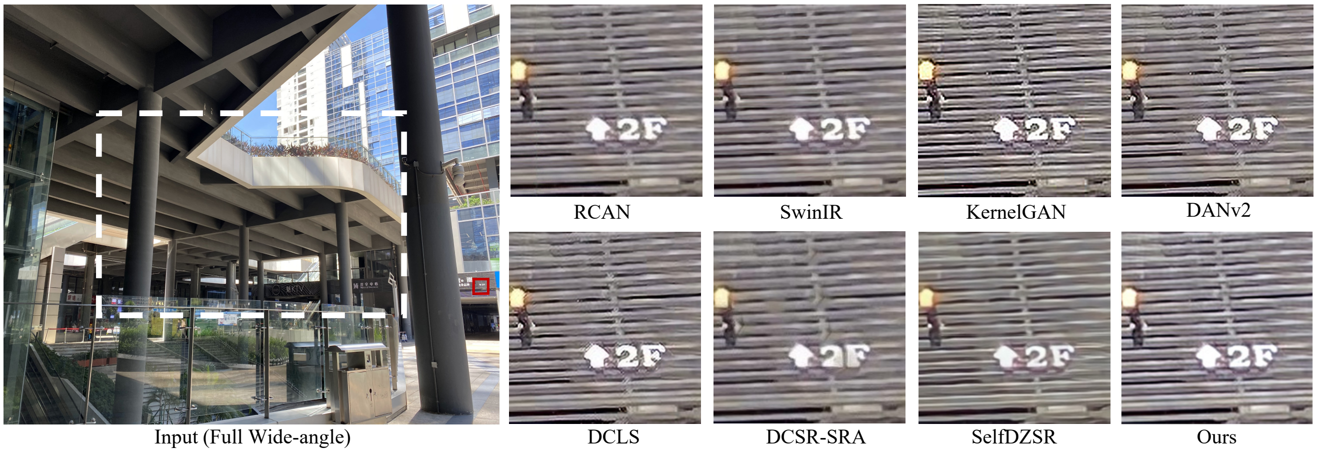


Figure 5. Visual comparisons for 2× SR on real-world data captured by iPhone11. The white dotted box indicates the overlapped FoV.

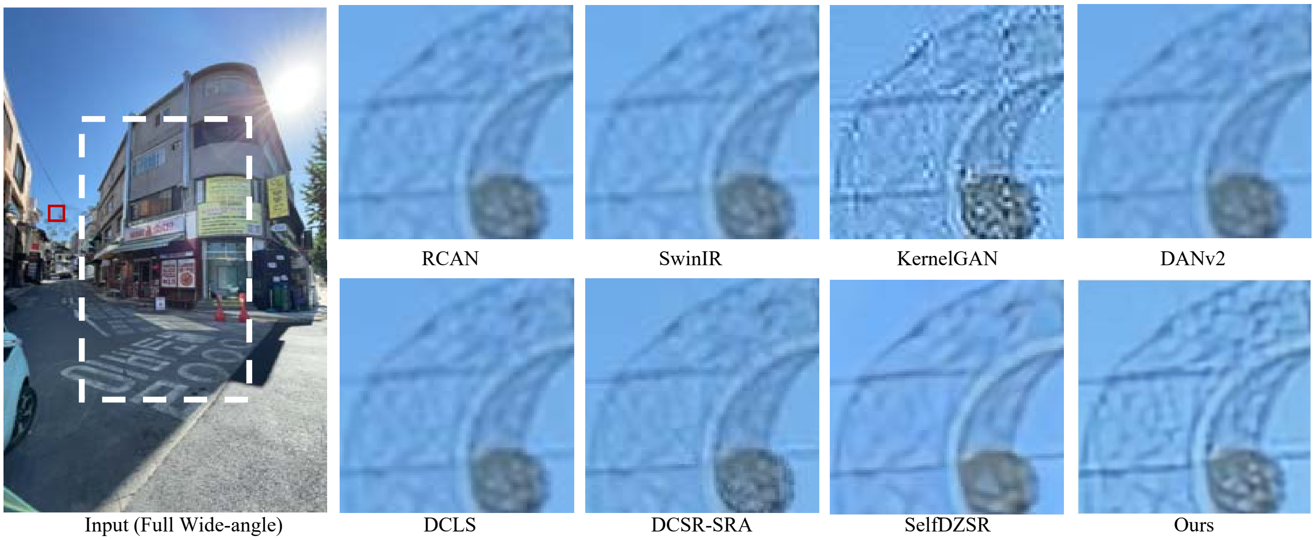


Figure 6. Visual comparisons for 2× SR on real-world data captured by iPhone12. The white dotted box indicates the overlapped FoV.

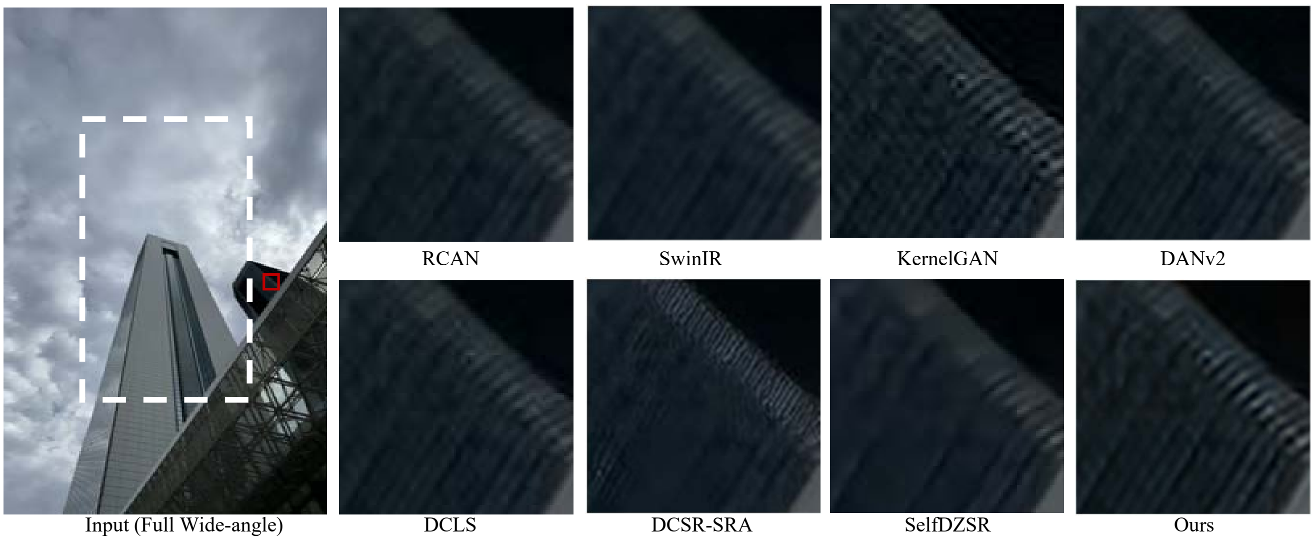


Figure 7. Visual comparisons for 2× SR on real-world data captured by iPhone12. The white dotted box indicates the overlapped FoV.

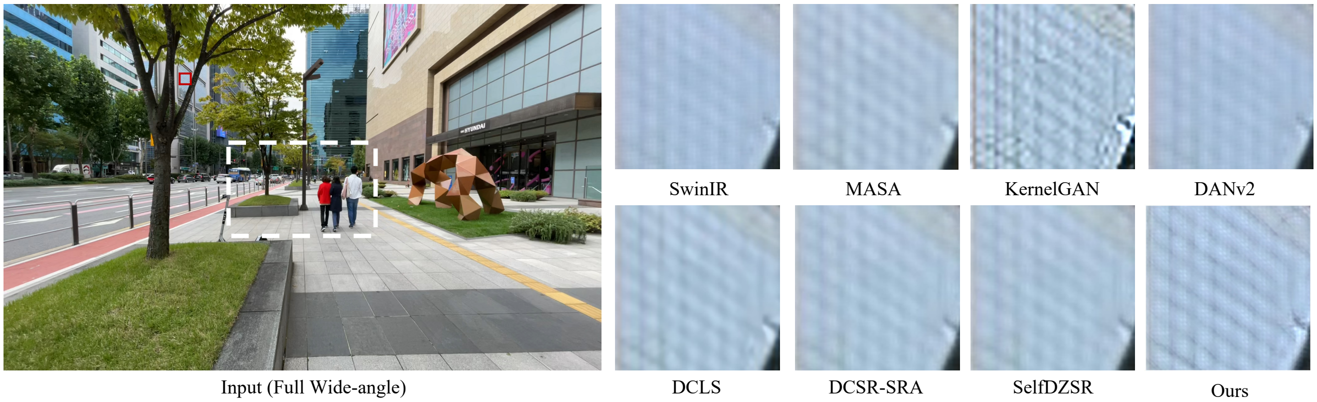


Figure 8. Visual comparisons for 4× SR on real-world data captured by iPhone12. The white dotted box indicates the overlapped FoV.

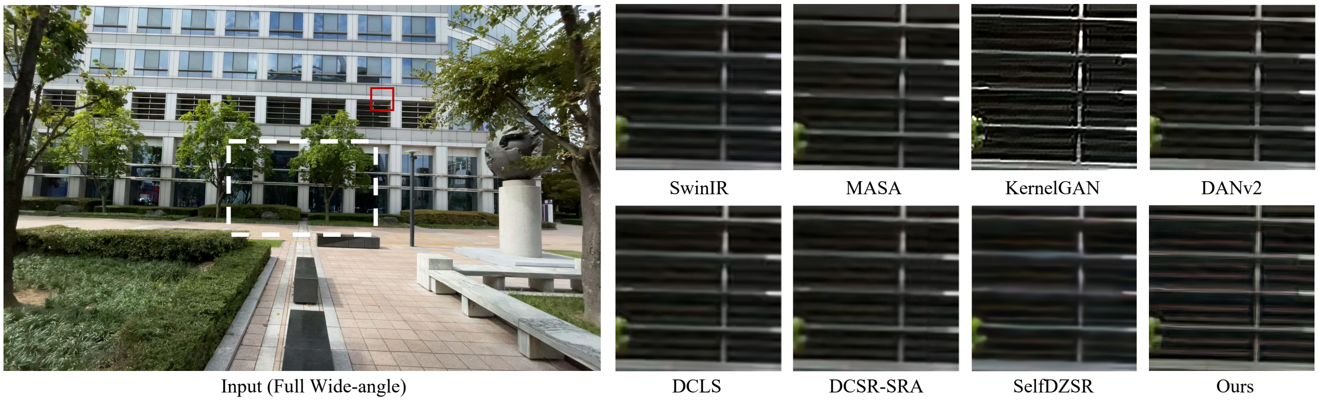


Figure 9. Visual comparisons for 4× SR on real-world data captured by iPhone12. The white dotted box indicates the overlapped FoV.

References

[1] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. Blind super-resolution kernel estimation using an internal-gan. In *NeurIPS*, 2019. 3

[2] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, 2016. 2

[3] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[4] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, 2016. 4

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[6] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022. 4

[7] Ziwei Luo, Haibin Huang, Lei Yu, Youwei Li, Haoqiang Fan, and Shuaicheng Liu. Deep constrained least squares for blind image super-resolution. In *CVPR*, 2022. 3

[8] Zhengxiong Luo, Yan Huang, Shang Li, Liang Wang, and Tieniu Tan. Unfolding the alternating optimization for blind super resolution. In *NeurIPS*, 2020. 3

[9] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014. 2

[10] Jae Woong Soh, Sunwoo Cho, and Nam Ik Cho. Meta-transfer learning for zero-shot super-resolution. In *CVPR*, 2020. 3

[11] Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, and Qifeng Chen. Dual-camera super-resolution with aligned attention modules. In *ICCV*, 2021. 4

[12] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPRW*, 2022. 4

[13] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *CVPR*, 2020. 4

[14] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *CVPR*, 2020. 3

[15] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 3, 4