Learning to Zoom Inside Camera Imaging Pipeline —Supplementary Materials—

Chengzhou Tang[1*] Yuqiang Yang[2*] Bing Zeng[2] Ping Tan[1] Shuaicheng Liu[2†]
[1]Simon Fraser University
[2]University of Electronic Science and Technology of China

A. Kernel Estimation

Sec. 3.2 gives the reason why we deal with single image super-resolution in the RAW-to-RAW domain. It is because the kernel assumption is violated by the image processing units that operate in a local neighborhood, such as the demosaicing and the denoising.

In this section, we further explain this motivation intuitively. We compare the final results and the intermediate kernels generated by our default method on RAW and the variation on RGB images. As shown in Fig. 1, the estimated kernels on the last column are more centralized and symmetric for our default method, while the kernels diverge and deviate from ideal distributions for the RGB variation. This observation is consistent with the one in Sec. 3.2 and verify the necessity of the RAW-to-RAW domain.

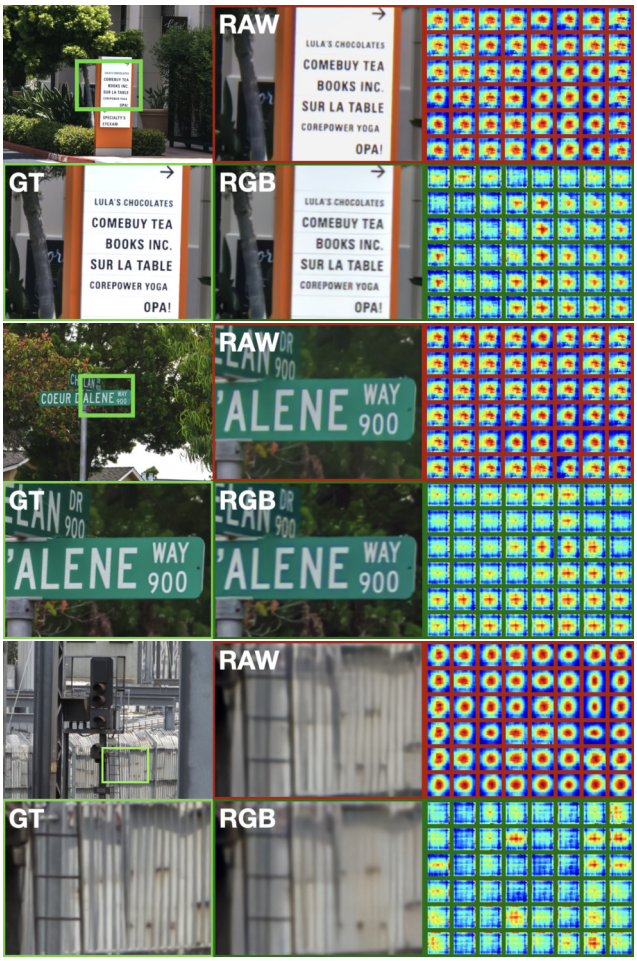


Figure 1. Comparisons of final results and intermediate kernels between our default method on RAW and the variation on RGB.

B. RAW Data Alignment

In this section, we further describe the details of our RAW-to-RAW image alignment in Sec. 4.1.

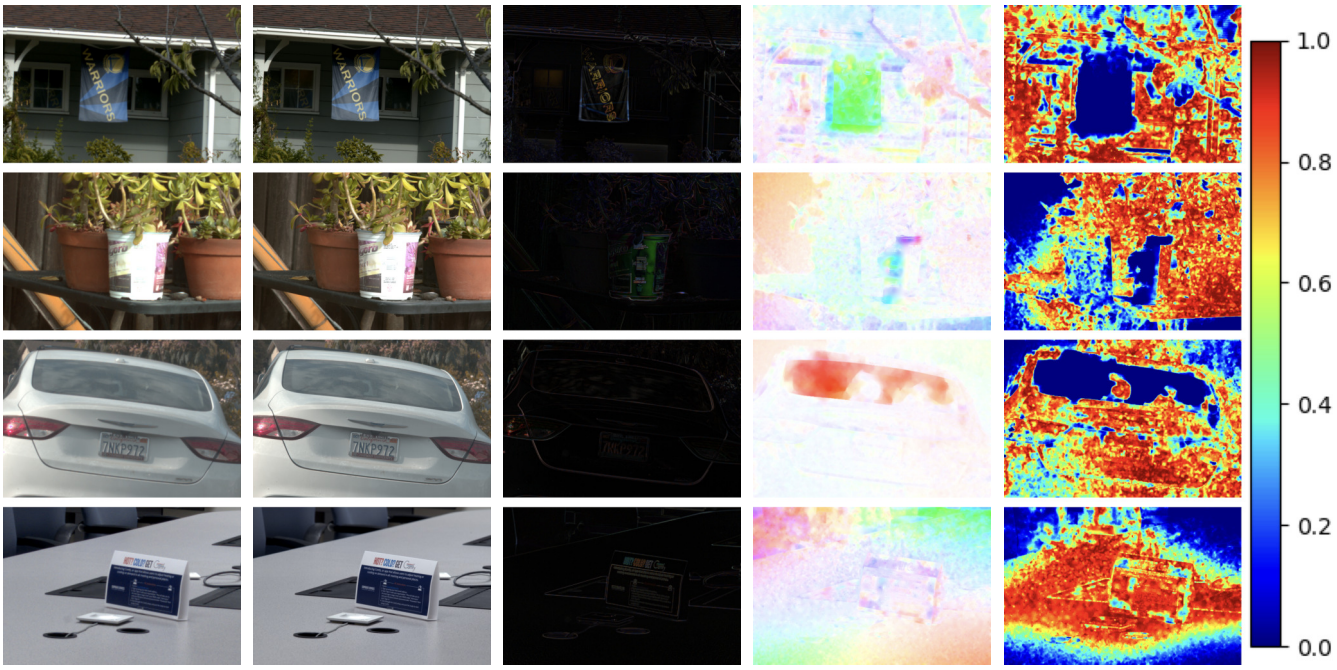
- Given a RAW image pair $\{\boldsymbol{x}, \boldsymbol{y}\}$, where $\boldsymbol{x}$ is the HR image and $\boldsymbol{y}$ has the same resolution but is the Bicubic up-sampled from the LR image, we apply black level correction, white balancing and demosaicing first, then compute and compensate lens distortion. We denote the processed image pair as $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\}$, the undistortion mapping from $\boldsymbol{y}$ to $\hat{\boldsymbol{y}}$ as $\mathcal{U}_{\boldsymbol{y}\to\hat{\boldsymbol{y}}}$, and the distortion mapping from $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$ as $\mathcal{D}_{\hat{\boldsymbol{x}}\to\boldsymbol{x}}$.

- Then, we extract the regions of interest (ROI) from images. With a slight abuse of notation, we reuse $\{\boldsymbol{x}, \boldsymbol{y}\}$ and $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\}$ after ROI cropping.

  For RealSR [1], since the ROI is not always centered, we detect the ROI by matching feature points [4] between $\hat{\boldsymbol{x}}$ and the original RGB image $\boldsymbol{x}_o$ from the dataset, estimate the transformation from $\boldsymbol{x}_o$ to $\hat{\boldsymbol{x}}$, and then crop the minimum exterior rectangle of the transformed $\boldsymbol{x}_o$ in $\hat{\boldsymbol{x}}$. It is the same for the ROI extraction of $\hat{\boldsymbol{y}}$.

  For SR-RAW [9], the ROI is always centered, so we keep the full image of $\hat{\boldsymbol{x}}$ and crop the ROI in $\hat{\boldsymbol{y}}$ based

*Equal contribution
†Corresponding author



(a) (Bicubic up-sampled) LR image $\boldsymbol{y}$. (b) (Warped) HR image $\boldsymbol{x}_{\mathcal{M}}$ (c) Color difference $\boldsymbol{c}$ (d) Optical flow $\boldsymbol{u}$ (e) Confidence map $\boldsymbol{w}$

Figure 2. Examples of (c) the color difference $\boldsymbol{c}$, (d) the optical flow $\boldsymbol{u}$, and (e) the final confidence map $\boldsymbol{w}$ computed from (a) $\boldsymbol{y}$ and (b) $\boldsymbol{x}_{\mathcal{M}}$.

on the ratio of focal lengths $s = f_{\boldsymbol{y}}/f_{\boldsymbol{x}}$, where $f_{\boldsymbol{x}}$ and $f_{\boldsymbol{y}}$ are the focal length of $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ respectively. Then, if the size of $\hat{\boldsymbol{x}}$ is $[W, H]$, we crop the center region with size $[sW, sH]$ in $\hat{\boldsymbol{y}}$ as its ROI.

- Then, we estimate the global transformation model $\mathcal{T}_{\hat{\boldsymbol{y}}\to\hat{\boldsymbol{x}}}$ from $\hat{\boldsymbol{y}}$ to $\hat{\boldsymbol{x}}$. $\mathcal{T}_{\hat{\boldsymbol{y}}\to\hat{\boldsymbol{x}}}$ is initialized as identity transformation and then refined by minimizing the photometric error:

$$\min_{\mathcal{T}} \sum_{\boldsymbol{p}} \|\hat{\boldsymbol{x}}_{\mathcal{T}(\boldsymbol{p})} - (\alpha\hat{\boldsymbol{y}}_{\boldsymbol{p}} + \beta)\|, \tag{1}$$

where $\mathcal{T}(\boldsymbol{p})$ is the transformed coordinate of $\boldsymbol{p}$ by $\mathcal{T}$, α and β are the luminance compensation parameters. The Eq. (1) is the same as in [1] except that it is operated on the demosaicked RAW pair $\{\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}\}$ instead of RGB.

- Finally, we compose the mapping from the $\boldsymbol{y}$ to $\boldsymbol{x}$ as $\mathcal{M}_{\boldsymbol{y}\to\boldsymbol{x}} = \mathcal{D}_{\hat{\boldsymbol{x}}\to\boldsymbol{x}} \odot \mathcal{T}_{\hat{\boldsymbol{y}}\to\hat{\boldsymbol{x}}} \odot \mathcal{U}_{\boldsymbol{y}\to\hat{\boldsymbol{y}}}$. After the mapping has been produced for each RAW pair $\{\boldsymbol{x}, \boldsymbol{y}\}$, we utilize the mappings for training, where $\boldsymbol{y}$ is *only* cropped without any pixel interpolation, while the high-resolution ground-truth $\boldsymbol{x}^*$ is warped by $\mathcal{M}_{\boldsymbol{y}\to\boldsymbol{x}}$ for loss computation.

However, the mapping $\mathcal{M}_{\boldsymbol{y}\to\boldsymbol{x}}$ is usually imperfect and contains misalignment due to parallax, illumination changes, or even moving objects. Therefore, we compute a confidence map $\boldsymbol{w}$ to measure the pixel-wise alignment quality between $\{\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{y}\}$.

- First, we compute the pixel-wise color difference between $\{\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{y}\}$ as $\boldsymbol{c}_{\boldsymbol{p}} = \boldsymbol{x}_{\mathcal{M}(\boldsymbol{p})} - \boldsymbol{y}_{\boldsymbol{p}}$.
- Second, we compute the optical flow map $\boldsymbol{u}$ between $\{\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{y}\}$ by the NL-Classic method in [6]. Note that the optical flow vector here does not have to be accurate. Instead, it only needs to have larger norm $\|\boldsymbol{u}_{\boldsymbol{p}}\|$ at misaligned pixels.
- Finally, we compute the per-pixel confidence $\boldsymbol{w}_{\boldsymbol{p}}$ as:

$$\boldsymbol{w}_{\boldsymbol{p}} = \frac{\boldsymbol{m}_{\boldsymbol{p}}}{z} \exp\left(-\frac{\|\boldsymbol{c}_{\boldsymbol{p}}\|^2}{\sigma_c^2} - \frac{\|\boldsymbol{u}_{\boldsymbol{p}}\|^2}{\sigma_u^2}\right), \tag{2}$$

where σ_c and σ_u are the standard deviations that control the color sensitivity and motion sensitivity respectively, z is the maximum weight over all pixels, which normalizes all the weights into $(0, 1]$. Additionally, $\boldsymbol{m}$ is the binary mask that excludes pixels with large color difference or flow vector norm, *i.e.*, $\boldsymbol{m}_{\boldsymbol{p}} = 0$ if $\|\boldsymbol{c}_{\boldsymbol{p}}\| > 0.5$ or $\|\boldsymbol{u}_{\boldsymbol{p}}\| > 1.5$ pixels, otherwise $\boldsymbol{m}_{\boldsymbol{p}} = 1$.

If $\{\boldsymbol{x}_{\mathcal{M}}, \boldsymbol{y}\}$ are perfectly aligned, both the color difference and the flow vector norm should be close to zero. Otherwise, as shown in Fig. 2, the misalignments are caused by non-static objects, parallax, lighting changes and reflections, which lead to large color difference or/and large flow vector norm, and finally give low confidence.

C. Additional Results

In this section, we give additional results besides the ones in the main paper. We still made comparisons with the *officially* released models from the methods that are designed for real data including LP-KPN [1], MIRNet [7], RAW-to-sRGB [11], Zoom-learn-Zoom [9]; the ones that are trained on synthetic data including LapSRN [2], EDSR [3], RCAN [10]; and the ones require external kernels as inputs for inference including SRMD [8] and ZSSR [5].

Methods	RealSR			SR-RAW		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Bicubic	27.24	0.821	0.335	27.09	0.796	0.314
EDSR [3]	27.34	0.828	0.331	27.24	0.805	0.310
RCAN [10]	27.36	0.828	0.330	27.25	0.806	0.309
SRMD [8]	27.54	0.830	0.326	27.42	0.813	0.301
MIRNet [7]	28.47	0.858	0.301	-	-	-
Ours	**33.36**	**0.937**	**0.196**	**34.26**	**0.944**	**0.173**

Table 1. Quantitative comparisons with methods designed on synthetic data and real data. '-' indicates there is no released model for such a test.

Comparisons of $3\times$: We first make comparisons under $3\times$ scale ratio. Both the quantitative comparisons in Tab. 1 and qualitative comparisons in Fig. 3 and Fig. 6 indicate our method performs better than others by a large margin.

Qualitative Comparisons of $4\times$: We also show additional qualitative results in Fig. 5 and Fig. 6 for RealSR [1] and SR-RAW [9] respectively.

Qualitative Comparisons of $2\times$: Since the $2\times$ scale ratio case is less challenging, the qualitative differences between different methods is less significant than the $3\times$ and the $4\times$ cases. Even though, our method still performs visually better on some images, especially the ones contains letters as shown in Fig. 7.

References

[1] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proc. CVPR*, pages 3086–3095, 2019. 1, 2, 3, 4, 6, 8

[2] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proc. CVPR*, pages 624–632, 2017. 3

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. CVPRW*, pages 136–144, 2017. 3

[4] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proc. CVPR*, 2020. 1

[5] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proc. CVPR*, pages 3118–3126, 2018. 3

[6] Deqing Sun, Stefan Roth, and Michael J. Black. Secrets of optical flow estimation and their principles. In *Proc. CVPR*, pages 2432–2439, 2010. 2

[7] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Proc. ECCV*, 2020. 3

[8] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *Proc. CVPR*, pages 3262–3271, 2018. 3

[9] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proc. CVPR*, pages 3762–3770, 2019. 1, 3, 5, 7, 8

[10] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proc. ECCV*, pages 286–301, 2018. 3

[11] Zhilu Zhang, Haolin Wang, Ming Liu, Ruohao Wang, Jiawei Zhang, and Wangmeng Zuo. Learning raw-to-srgb mappings with inaccurately aligned supervision. In *Proc. ICCV*, pages 4348–4358, 2021. 3

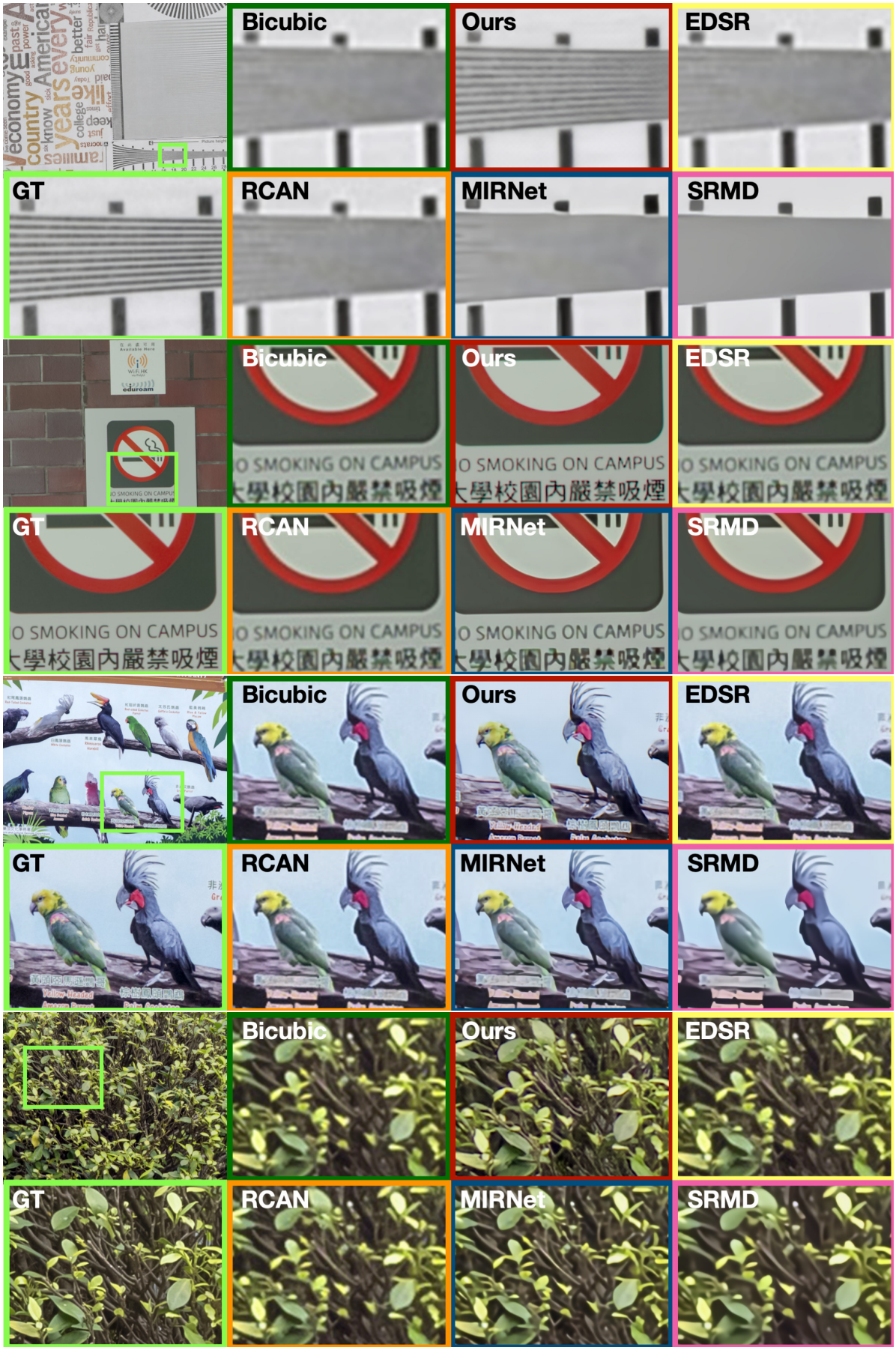


Figure 3. Qualitative comparisons on RealSR [1] under 3× scale ratio.

Figure 4. Qualitative comparisons on SR-RAW [9] under 3× scale ratio.

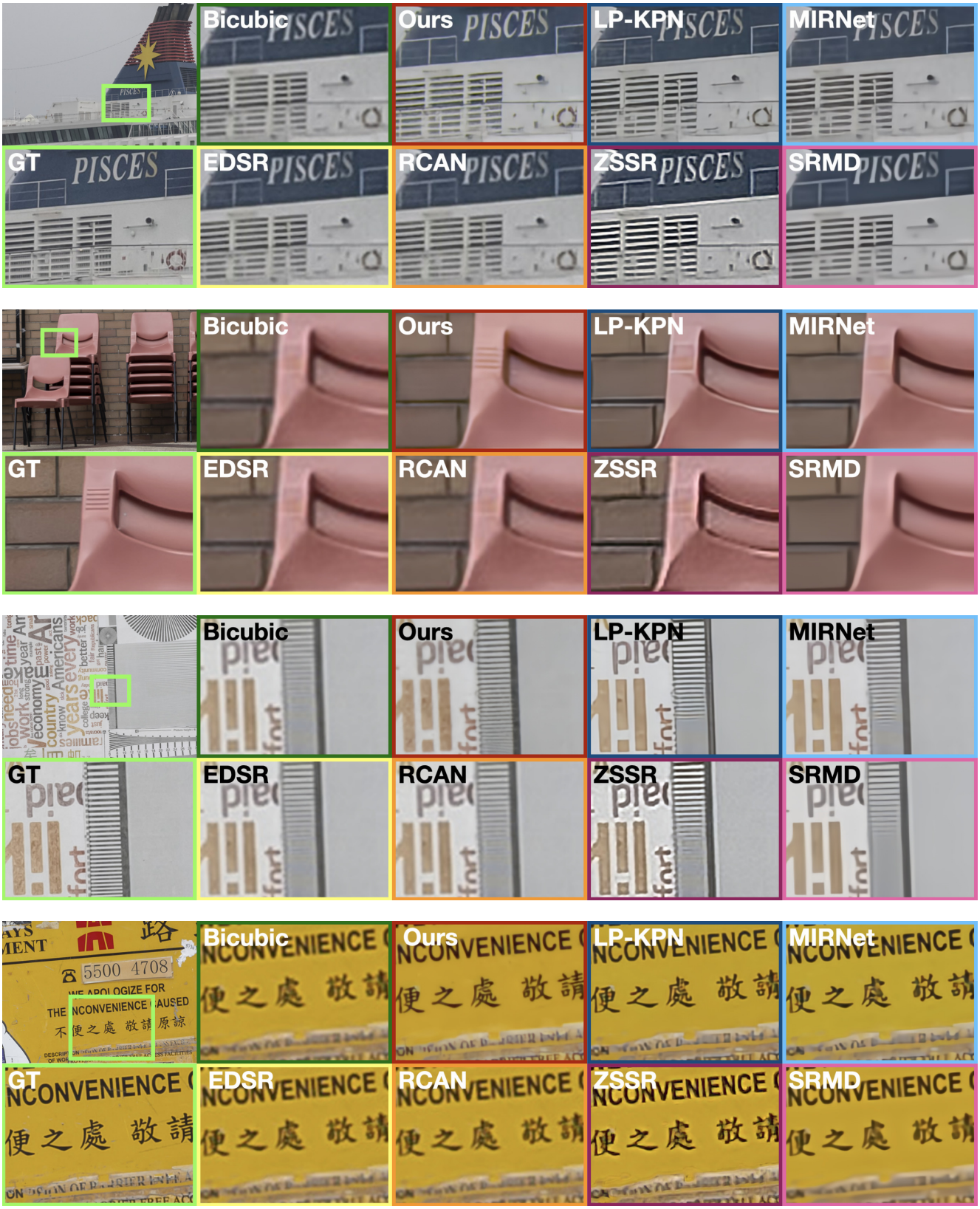
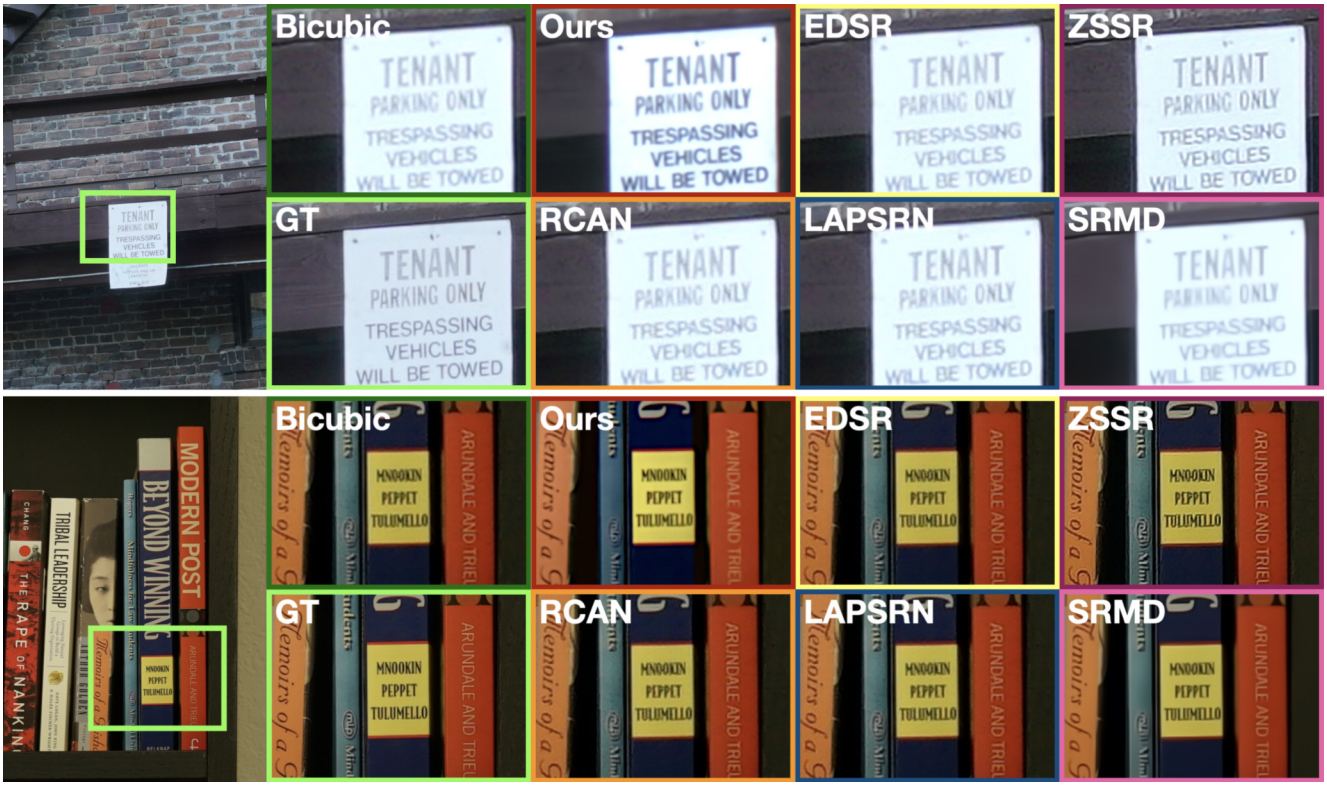


Figure 5. Qualitative comparisons on RealSR [1] under $4\times$ scale ratio.

Figure 6. Qualitative comparisons on SR-RAW [9] under $4\times$ scale ratio.

(a)



(b)

Figure 7. Qualitative comparisons of 2× scale ratio on (a) RealSR [1] and (b) SR-RAW [9]