Diffusion Transformer meets Multi-level Wavelet Spectrum for Single Image Super-Resolution

Supplementary Material

This supplementary material includes more details and analyses on the proposed method. First of all, we introduce additional implementation details about the proposed DTWSR. Then we provide more qualitative comparison to demonstrate the superiority of our method on both fidelity and image quality. Next we give additional analyses on DTWSR, and finally we present additional visualization results to show the effectiveness of our model.

1. Additional implementation details

1.1. Model architecture

Table S1 presents the detailed architecture configurations of the proposed DTWSR. Specially, we use slightly different hyper-parameters for general SISR and face SISR. The final parameter size is comparable to IDM (94M) [5] and DiWa (92M) [8], but much less than SR3 (550M) [10],

Table S1. Architecture hyper-parameters for the proposed DTWSR. Layers of the dual-decoder is the numbers of TransBlock in LEDec/HDDec.

Task	General 4×	Face 8×	Face 16×
Layers of dual-decoder	8/14	6/16	6/16
Hidden size	512	512	512
Heads	8	8	8
Minimal patch size	4	2	2
Parameters	107.8M	108.5M	108.5M

1.2. Implementation

We train the model using AdamW optimizer (with $\beta_1 = 0.5$ and $\beta_2 = 0.9$), with a fixed learning rate of 3×10^{-4} for the generator and 3×10^{-5} for the discriminator. The overall diffusion timesteps is set to 4. For the hyper-parameters of α, β, γ in Equation 14, α is initialized as 0.15 and gradually reduced to 0.02 at the final iteration; β and γ are set as 0.5.

For face SISR, images are super-resolved from 16^2 to 128^2 for $8\times$ and from 16^2 to 256^2 for $16\times$. All ground-truth (GT) and LR images are obtained from the original dataset by bicubic downsampling. Horizontal flips are used randomly for data augmentation in model training. For general SISR, we crop the original images into patches of 160^2 pixels as HR images. They are then downsampled to 40^2 by the bicubic kernel as the corresponding LR images. Random rotation and horizontal flips are performed for data augmentation in model training.

The detailed training and sampling process are outlined in Algorithm 1 and Algorithm 2 respectively.

Algorithm 1: Training process

```
Require: $I_{lr}$:LR image, $I_0$: HR image, $\mathrm{G}_\phi$: generator,
    $\mathrm{D}_\theta$: discriminator, $T$: max timestep, $q(\cdot)$: diffusion
    and denoising process (as refer to [12]).
repeat
    $t \sim Uniform(\{1, ..., T\})$
    $I_t \sim q(I_t|I_0)$
    $I_{t-1} \sim q(I_{t-1}|I_t, I_0)$
    $\tilde{I}_{t-1} \sim q(I_{t-1}|I_t, \mathrm{G}_\phi(I_{lr}, I_t, t))$
    Take a gradient descent step on
        $\nabla_\theta \left|\log(\mathrm{D}_\theta(\tilde{I}_{t-1}, I_t, t)) - \log(\mathrm{D}_\theta(I_{t-1}, I_t, t))\right|$
    $\tilde{I}_0 \sim \mathrm{G}_\phi(I_{lr}, I_t, t)$
    $\tilde{I}_{t-1} \sim q(I_{t-1}|I_t, \tilde{I}_0)$
    Take a gradient descent step on
        $\nabla_\phi \left(\left\|\tilde{I}_0 - I_0\right\| - \left|\log(\mathrm{D}_\theta(\tilde{I}_{t-1}, I_t, t))\right|\right)$
until $\mathrm{G}_\phi$ converged
```

Algorithm 2: Sampling process

```
Require: $I_{lr}$: LR image, $\mathrm{G}_\phi$: generator, $T$: max timestep,
    $q(\cdot)$: diffusion and denoising process (as refer to [12]).
$I_T \sim \mathcal{N}(0, \mathbf{I})$
for all $t = T, ..., 1$ do
    $\tilde{I}_0 = \mathrm{G}_\phi(I_{lr}, I_t, t)$
    $I_{t-1} \sim q(I_{t-1}|I_t, \tilde{I}_0)$
end for
return $I_0$
```

2. More comparisons and visualizations

2.1. Inference speed

Firstly, We test the inference speed of our model and compare with other diffusion model based methods[1] on 1 NVIDIA Tesla A100. As shown in Table S2, our method is more efficient than others.

2.2. Additional qualitative visualization

In this section, we provide more visual comparisons with state-of-the-art methods on both face and general SISR. For

[1] All methods are evaluated based on the official codes and hyper parameters.

Table S2. Comparison on the inference speed of super-resolving a single image.

Task	Face 8×				General 4×		
Method	SR3 [10]	DiWa [8]	IDM [5]	Ours	SRDiff [7]	DiWa	Ours
Times(s)	36.2	5.44	40.1	0.11	22.3	62.5	14.9

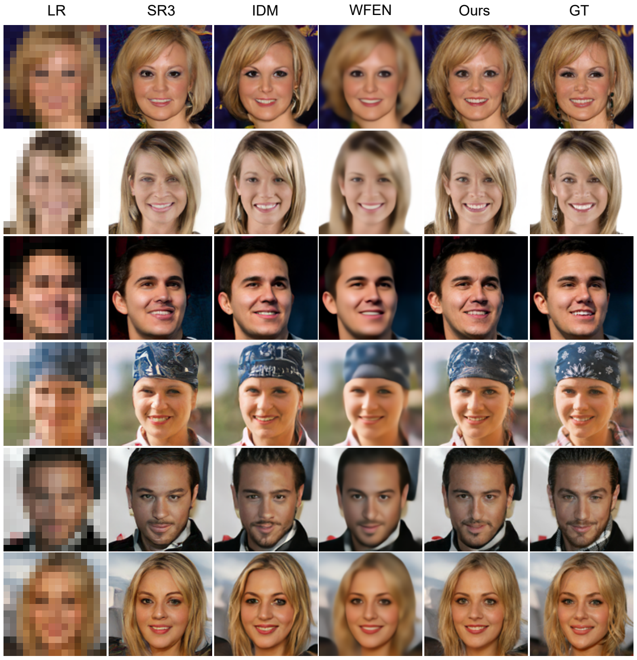


Figure S1. Qualitative comparison on CelebA [6] 8× SISR (16 × 16 to 128 × 128). Zoom in for best view.

face SISR, Figure S1, Figure S2 and Figure S3 show SISR results of 8×, 16×, 12× and 15×. It can be observed that our method consistently achieves higher fidelity and image quality (see the eyes, mouth, hair and skin texture). Qualitative comparison on general SISR is presented in Figure S4.

2.3. General SISR

To verify the generalizability of DTWSR, we perform the comparison with diffusion model-based SOTA methods on Manga109 (comic image dataset) [4], Set5 [2] and Set14 [13] SISR. Figure S5 and Figure S6 represent the comparisons on Mange109 and Set14 respectively, where the results generated by DTWSR are better than others in structural integrity and authenticity of detail.

3. Further analysis

In this section, further analyses are conducted to illustrate the effectiveness of the proposed model.

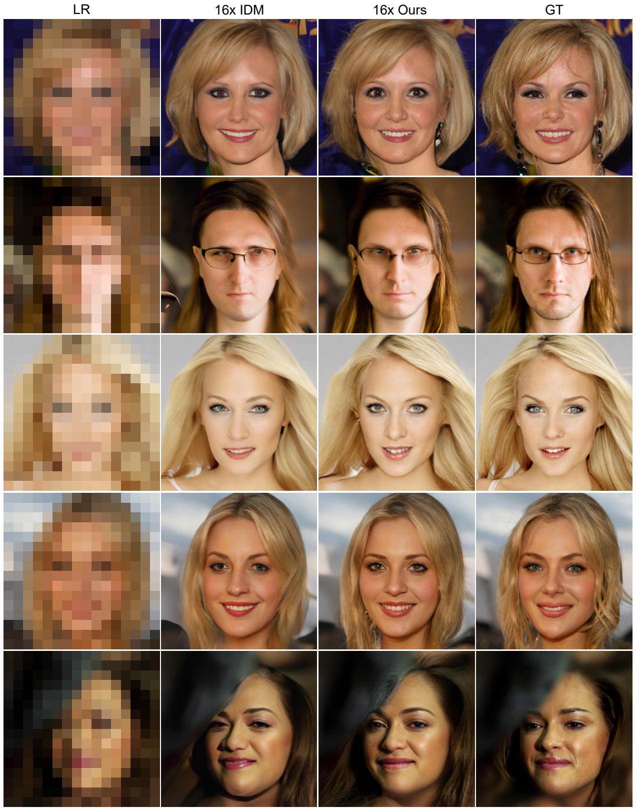


Figure S2. Qualitative comparison on CelebA [6] 16× SISR (16 × 16 to 256 × 256). Zoom in for best view.

3.1. LR conditioning effected by dual-decoder

We use the relative log amplitudes of Fourier transformed feature map [9, 11] to analyze the LR conditioning provided in each decoder, as shown in Figure S7b. Specifically, a higher value indicates that the model captures more information from LR input in the corresponding frequency band.

Dual-decoder design. As shown in Figure S7a, compared to freq-DiT, the dual-decoder (DTWSR (a)) enables the model to capture more information from LR condition, owning to the LR conditional features captured by each decoder can be more targeted.

LF residual. As shown in Figure S7b, the LR amplitudes of DTWSR (b) are smaller than those of DTWSR (a) in LEDec, but the opposite results are shown in HDDec, which indicates that part of conditional LF denoising in LEDec is transferred to HDDec. This suggests that LR conditioning in HDDec also is used for LF sub-band denoising, hindering its focus on guiding HF denoising.

Tailored attention mask. The tailored attention masks are proposed to avoid the unnecessary interaction among token. In this way, the LR condition focuses on conditioning denoising of LF components in the LEDec and conditioning denoising of HF components in the HDDec. As

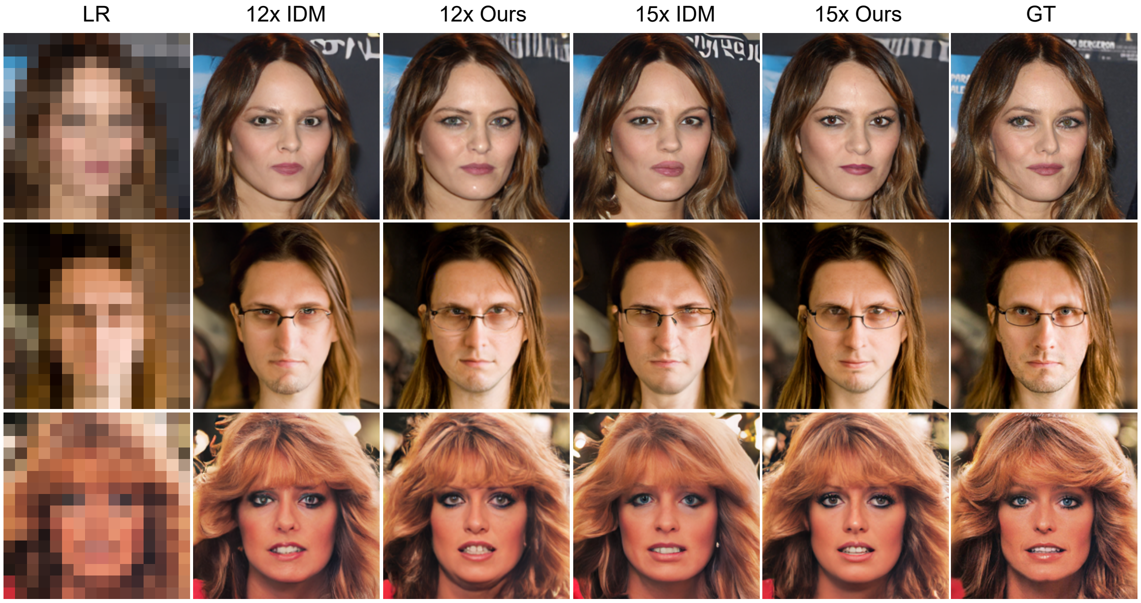


Figure S3. Qualitative comparison on CelebA [6] in 12× and 15× SISR (16 × 16 to 192 × 192 and 240 × 240). Zoom in for best view.

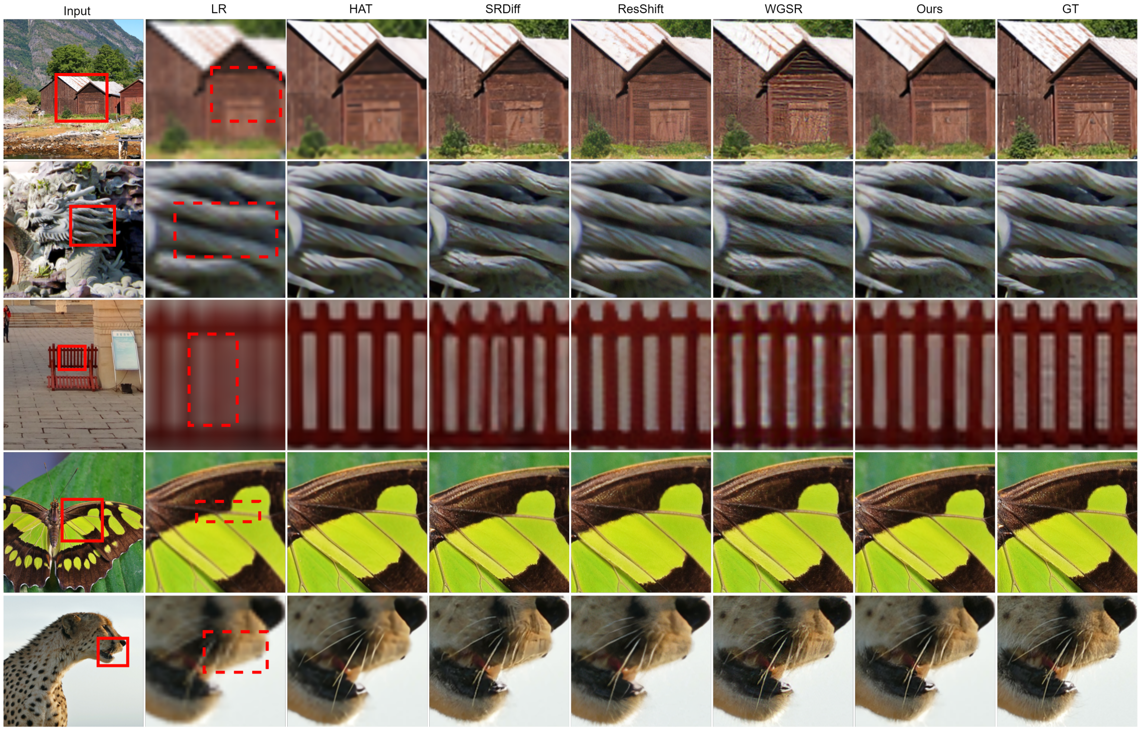


Figure S4. Qualitative comparison on DIV2K [1] in 4× SISR. Zoom in for best view.

shown in Figure S7a, DTWSR (ours) captures more information across all frequency bands, demonstrating that the proposed components facilitates the LR conditioning for better fidelity.

3.2. Relationships in multi-wavelet spectrum

We visualize the attention maps in LEDec and HDDec separately to show the relationships among sub-bands in multi-wavelet spectrums. The attention maps are obtained by averaging the attention scores across multiple

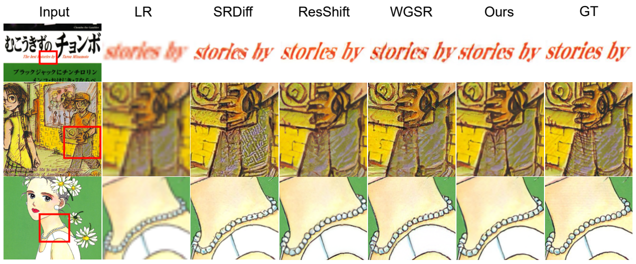


Figure S5. Qualitative comparison on Manga109 4× SISR. Zoom in for best view.

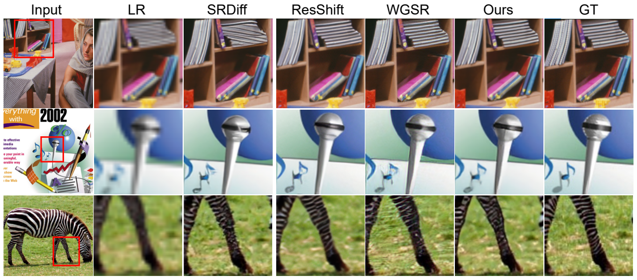
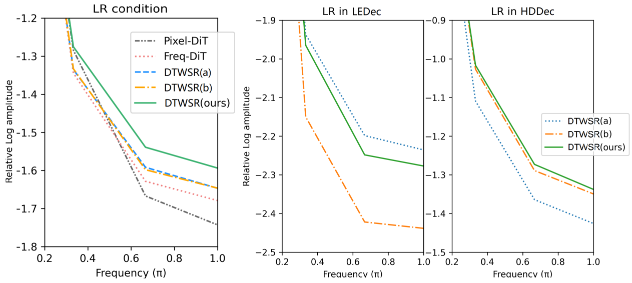


Figure S6. Qualitative comparison on Set14 4× SISR. Zoom in for best view.



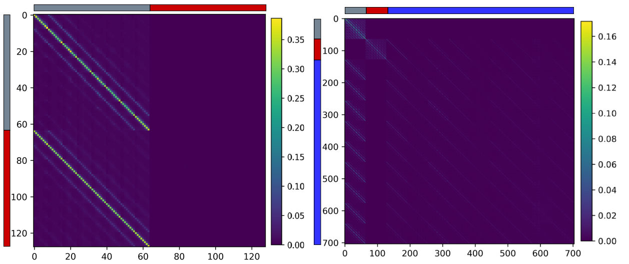
(a) (b)

Figure S7. The relative log amplitudes of Fourier transformed feature map. LR condition for LF and HF denoted as LR in LEDec and HDDec. The higher amplitude value indicates more dense information, and amplitude value is zero at 0π.

layers. As shown in Figure S8a, low-frequency sub-band focuses mostly on the LR input for generation. In Figure S8b, it can be observed that (1) besides the LR condition, low-frequency sub-band also pays attention to high-frequency sub-bands for information supplementing; (2) high-frequency sub-bands not only attend to the LR condition, but also interact with other sub-bands on features that have the same semantic information.

3.3. Visualization of the ablation experiments.

Figure S9 shows that the results obtained by DTWSR (ours) are more consistent with the ground truth in identities (*e.g.*, eyes and mouth in the first row) and details (*e.g.*, hair and skin in the second row), which reveals the effectiveness of the proposed method. As shown in Figure S10, the absence of the pyramid tokenization (DWTSR (c)) leads to degradation in both texture details (such as hair) and facial features (such as eyes and mouths). When the model not to consider the relationships between different levels of frequency bands, noticeable artifacts appear in the image, such as in the hair.



(a) Attention map in LEDec (b) Attention map in HDDec

Figure S8. Example of attention maps with masks on 8× face SISR. For 16 × 16 to 128 × 128 SISR with the minimum patch size of 2, the lengths of tokens are 128 and 704 in LEDec and HDDec respectively. (a) 0-63 are LR condition tokens (indicated by Gray lines), 64-127 are noisy low-frequency sub-band tokens (indicated by Red lines). (b) 0-63 are LR condition tokens, 64-127 are low-frequency sub-band tokens from LEDec, 128-703 are noisy high-frequency sub-bands tokens (indicated by Blue lines). Zoom in for best view.

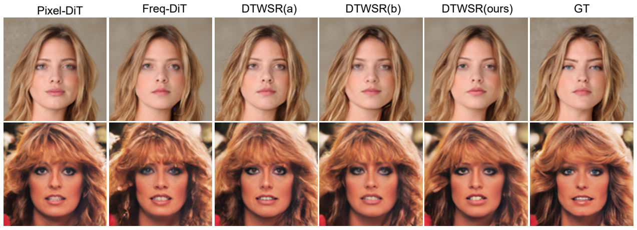


Figure S9. Comparison of our ablations for 8× face SISR. In the first row, the result generated by DTWSR (ours) is more similar to ground truth, especially in terms of face characteristics, *e.g.*, eyes and mouth. The second row indicates that our result is more authentic in details than others, such the orientation of hair in front of the forehead, *etc.*. Zoom in for best view.

3.4. Non-2^n magnification

For SISR with non-2^n magnification, the size of LR image cannot match that of low-frequency sub-band. Although the requested information for decoding can be obtained via self-attention operation, the inconsistency of receptive field between LR feature and each wavelet sub-bands will affect model performance. In this work, we upsample the LR image to have the same size as the low-frequency sub-band, so that the resulted features after pyramid patchify have the same receptive field. As demonstrated in Figure S11, with-

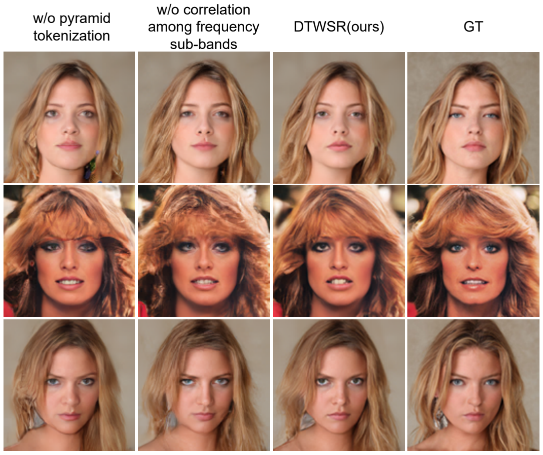


Figure S10. Comparison of our ablations for 8× face SISR. The ”w/o pyramid tokenization” refers to replacing the pyramid tokenization in DTWSR with a vanilla tokenization with a patch size of 4. The ”w/o correlation among frequency sub-bands” refers to the model does not consider the correlation among the multiple scale frequency sub-bands by applying a designed attention mask to the attention modules.

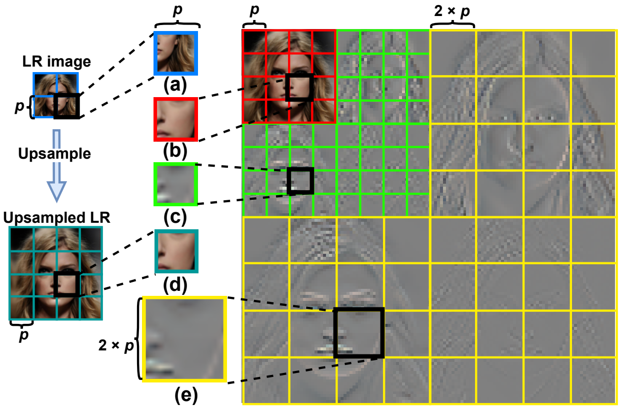


Figure S11. Illustration of the receptive field between LR and each wavelet sub-bands in Non-2^n magnification SISR. Here we take 2-level wavelet spectrum as an example. The minimum patch size is p. Each grid represents the feature receptive field after patchify. The receptive field from the original LR image (a) is inconsistent with those from the wavelet spectrum (b)(c)(e). The deviation can be alleviated after upsampling (comparing (d) to (b)(c)(e)).

out LR upsampling, the receptive field of the resulted feature is not aligned with those from the wavelet spectrum (comparing the contents in (a) and (b)(c)(e) in Figure S11). By upsampling LR to the size of low-frequency sub-band, the resulted feature after patchify will have the same receptive field as those from each wavelet sub-band (comparing the contents in (d) and (b)(c)(e) in Figure S11).

4. Additional SISR results

Figure S12 presents more high-resolution face images of DTWSR sampled on CelebA [6] in 16 × 16 to 128 × 128 and 256 × 256. More high-resolution general images of DTWSR sampled on DIV2K [1] and Manga109 [4] are shown in Figure S13, S14, S15, S16, S17 and S18. More comparisons with SOTA method on real-world SR are shown in Figure S19.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 126–135, 2017. 3, 5, 7, 8, 9, 10

[2] Marco Bevilacqua, Aline Roumy, Christine M. Guillemot, and Marie-Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Brit. Mach. Vis. Conf.*, 2012. 2

[3] 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 *Int. Conf. Comput. Vis.*, pages 3086–3095, 2019. 13

[4] Azuma Fujimoto, Toru Ogawa, Kazuyoshi Yamamoto, Yusuke Matsui, T. Yamasaki, and Kiyoharu Aizawa. Manga109 dataset and creation of metadata. *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, 2016. 2, 5, 11, 12

[5] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10021–10030, 2023. 1, 2

[6] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Int. Conf. Learn. Represent.*, 2018. 2, 3, 5

[7] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[8] Brian B Moser, Stanislav Frolov, Federico Raue, Sebastian Palacio, and Andreas Dengel. Waving goodbye to low-res: A diffusion-wavelet approach for image super-resolution. In *2024 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2024. 1, 2

[9] Namuk Park and Songkuk Kim. How do vision transformers work? In *Int. Conf. Learn. Represent.*, 2022. 2

[10] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4713–4726, 2022. 1, 2

[11] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng Yan. Inception transformer. *Adv. Neural Inform. Process. Syst.*, 35:23495–23509, 2022. 2

[12] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. In *Int. Conf. Learn. Represent.*, 2022. 1

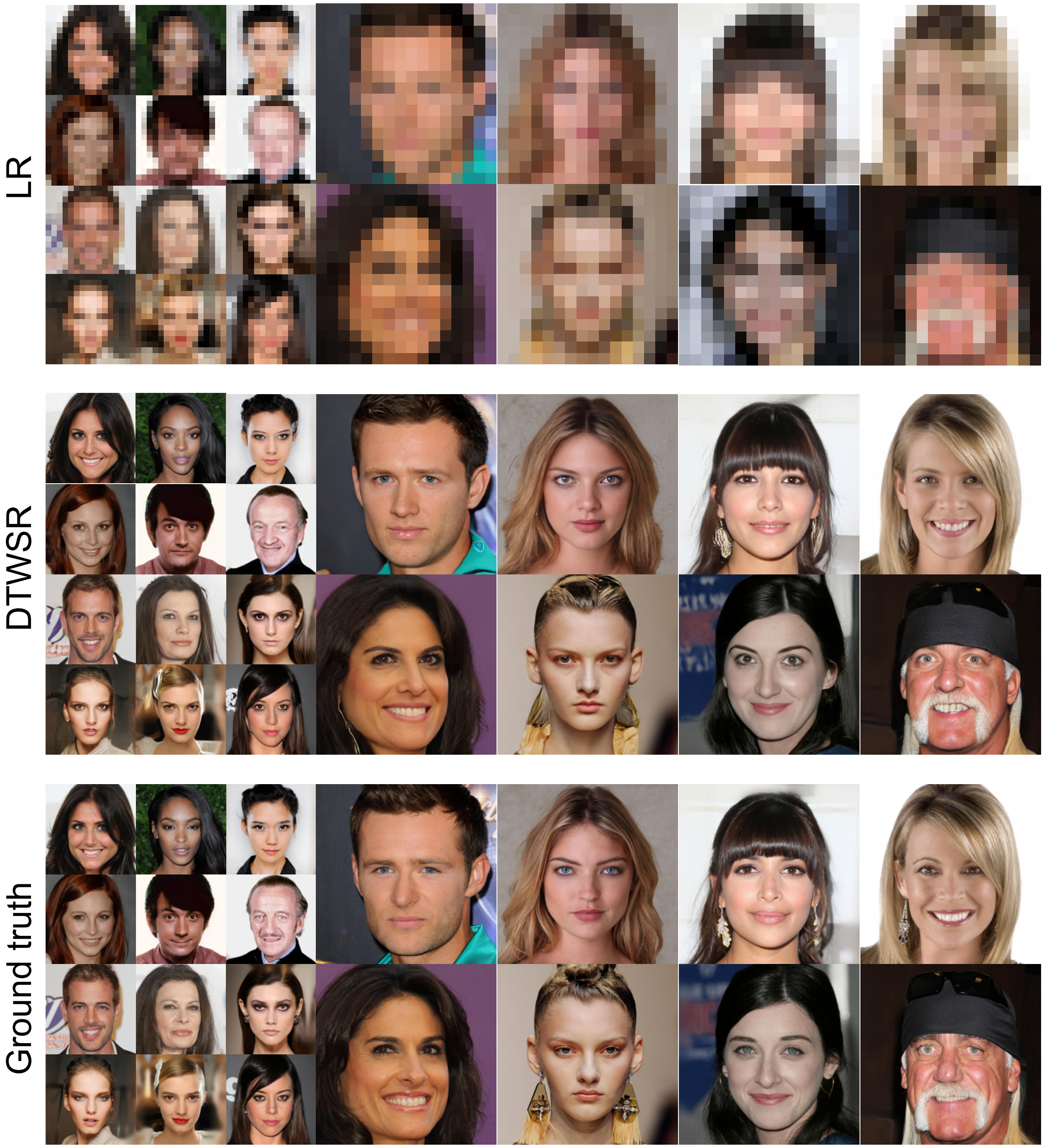


Figure S12. Additional 8× and 16× face SISR results of DTWSR. Zoom in for best view.

[13] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, 2010. 2

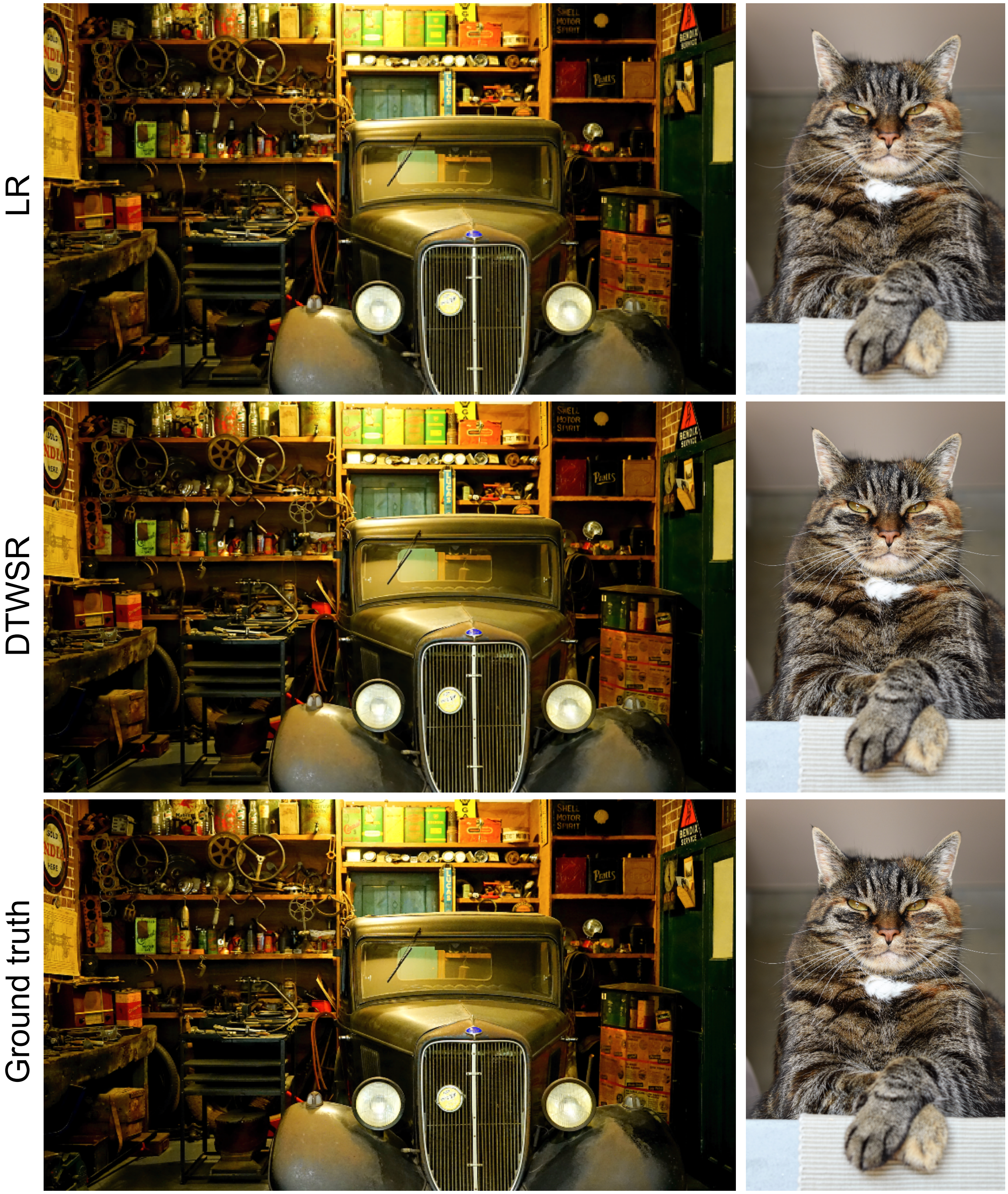


Figure S13. Additional 4× general (DIV2K[1]) SISR results of DTWSR. Zoom in for best view.

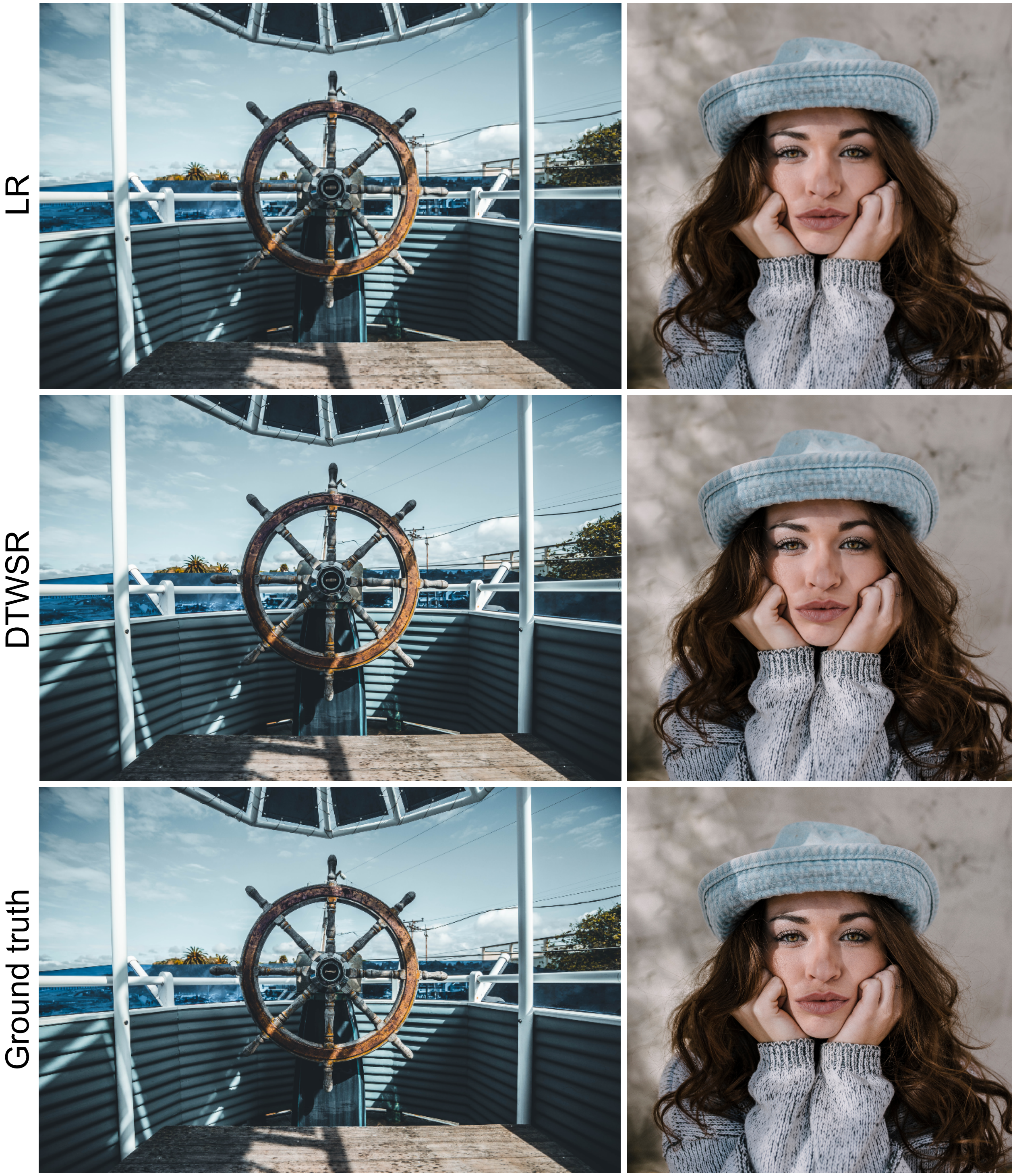


Figure S14. Additional 4× general (DIV2K[1]) SISR results of DTWSR. Zoom in for best view.

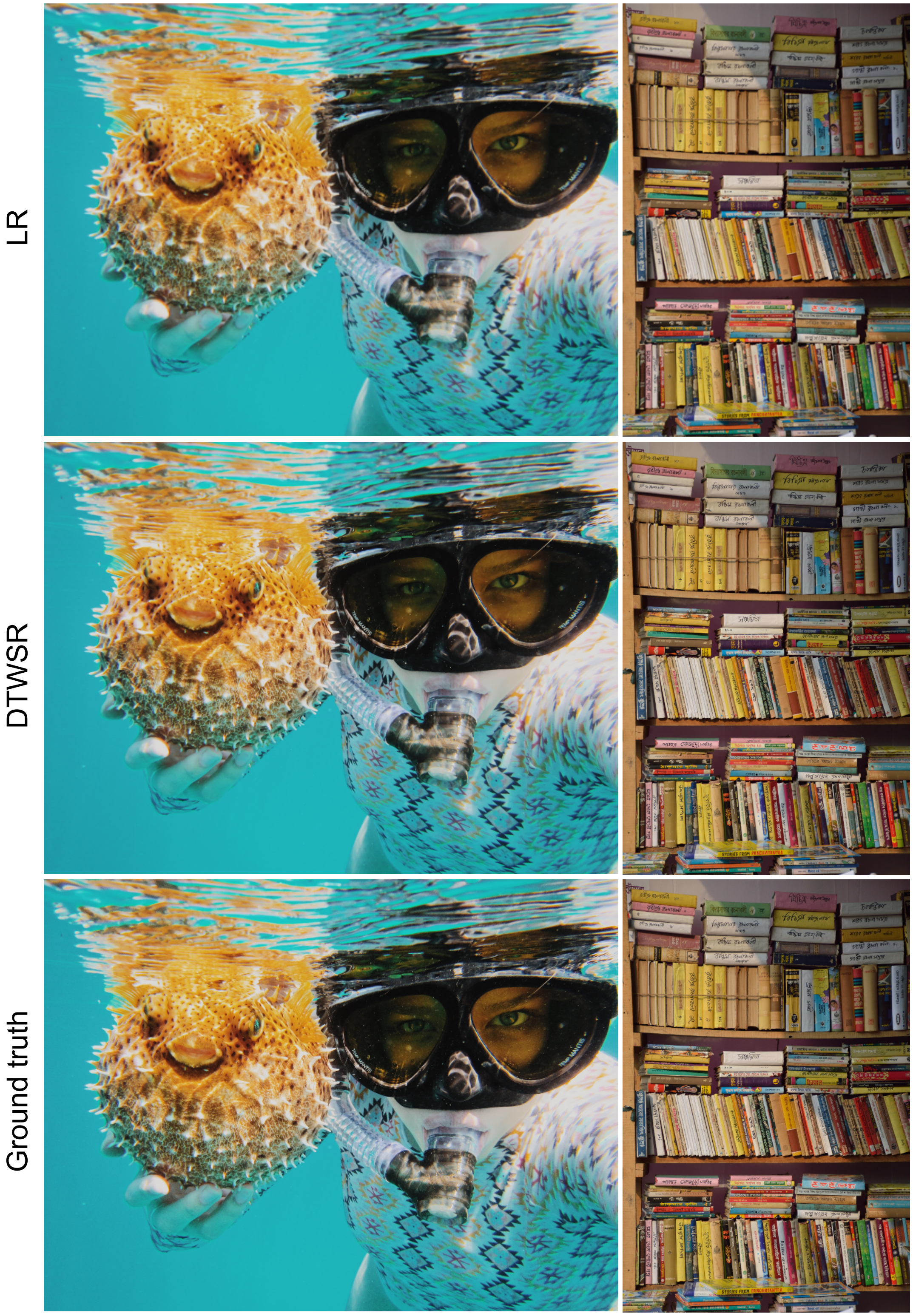


Figure S15. Additional 4× general (DIV2K[1]) SISR results of DTWSR. Zoom in for best view.

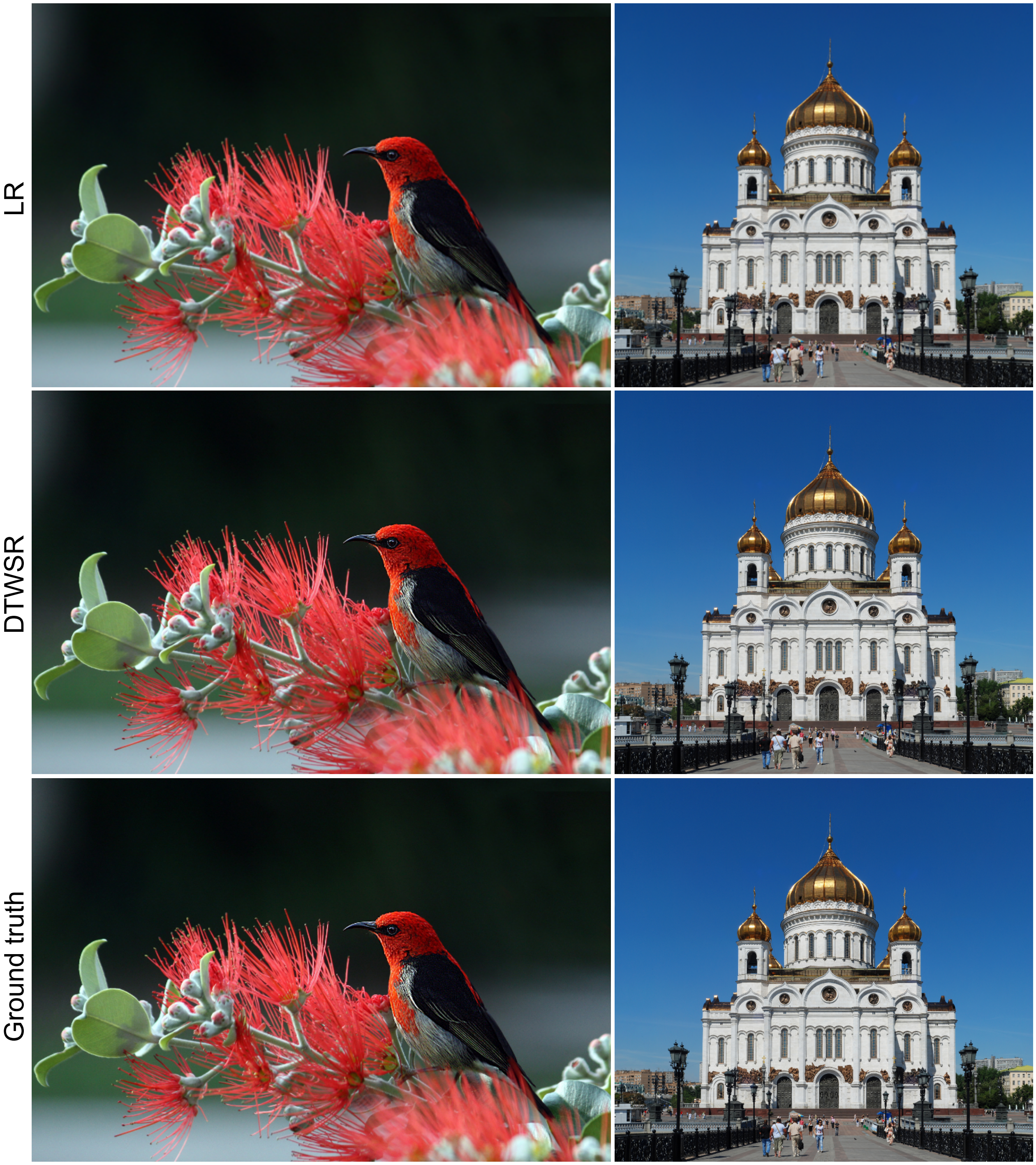


Figure S16. Additional 4× general (DIV2K[1]) SISR results of DTWSR. Zoom in for best view.

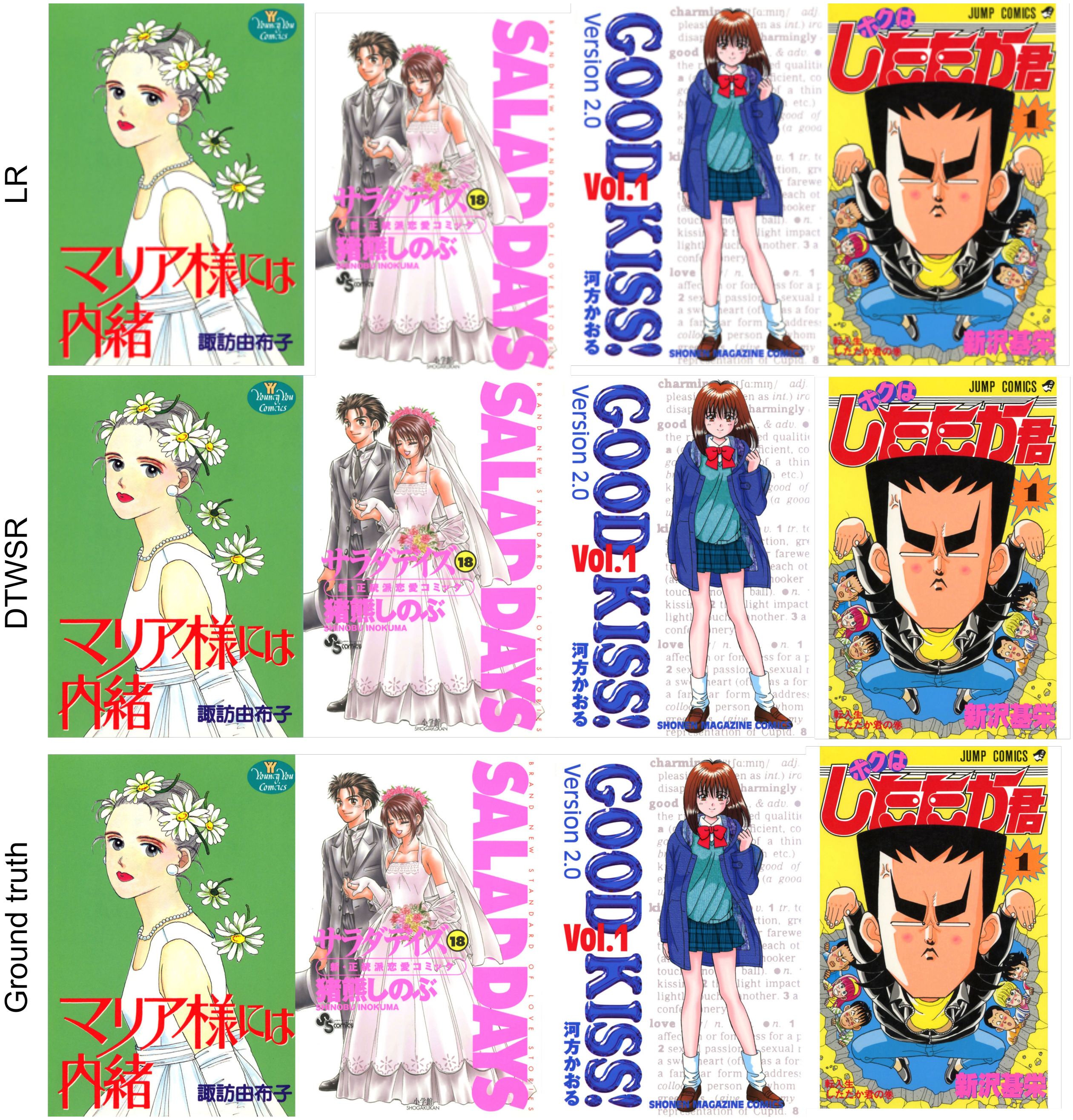


Figure S17. Additional 4× general (Manga109[4]) SISR results of DTWSR. Zoom in for best view.

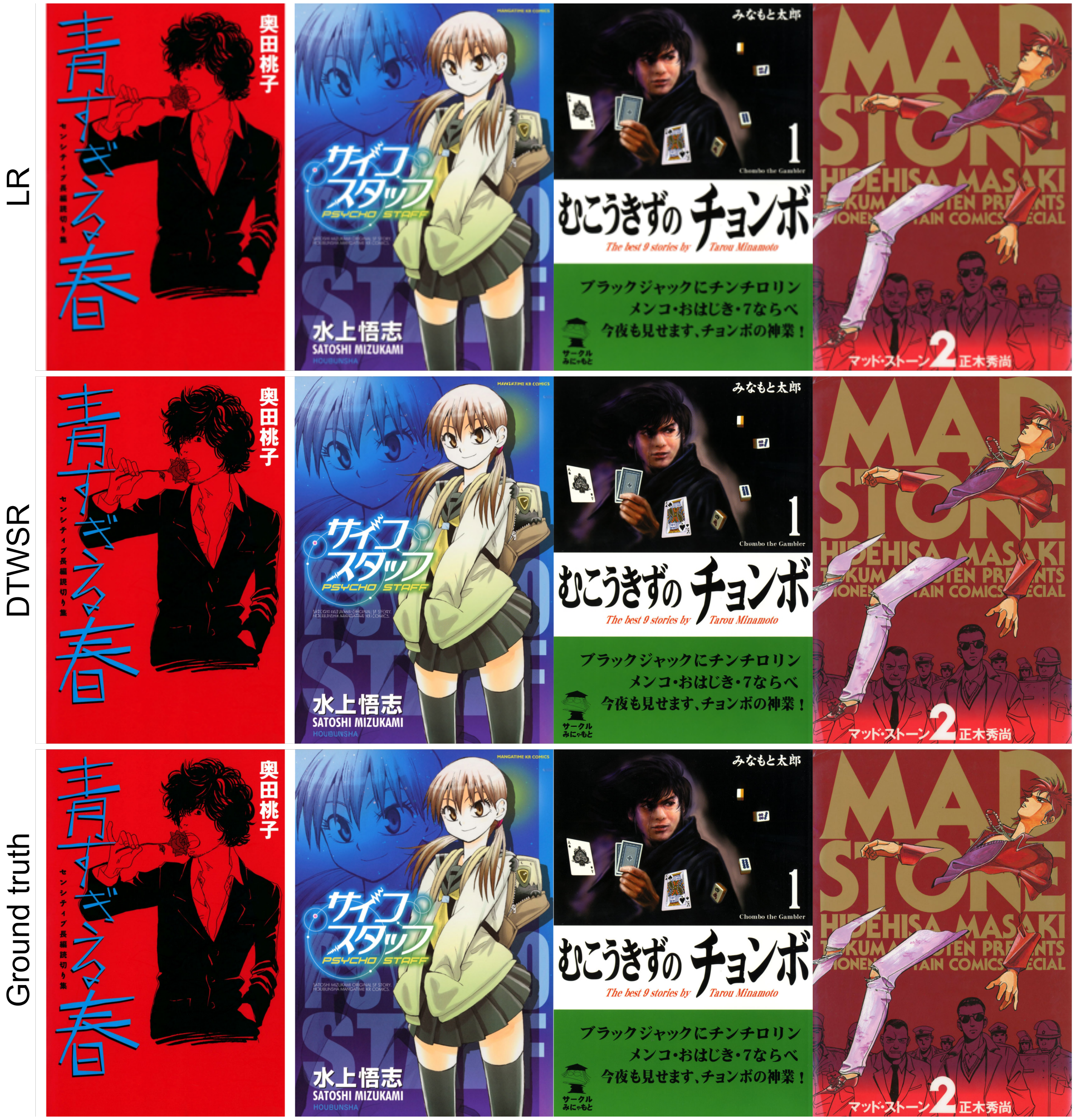


Figure S18. Additional 4× general (Manga109[4]) SISR results of DTWSR. Zoom in for best view.

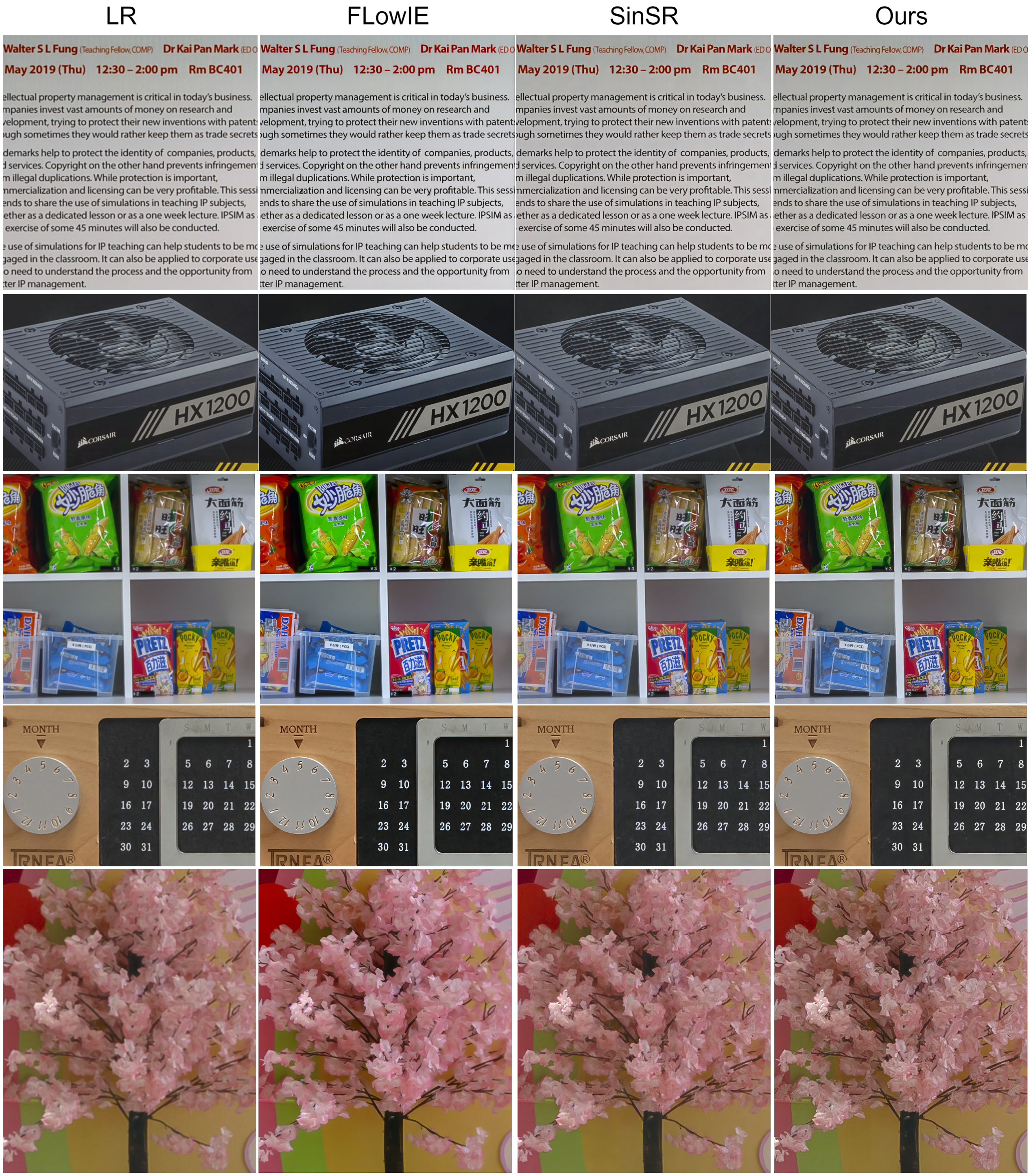


Figure S19. Comparison on real-world SR [3]. Zoom in for best view.