Supplementary Material to "Fine-structure Preserved Real-world Image Super-resolution via Transfer VAE Training"

The following materials are provided in this supplementary file:
1. More visual comparisons in Real-ISR task and STISR task (referring to Sec. 4.2 in the main paper);
2. Comparisons with GAN-based methods.

1. More Visual Comparisons

We provide more visual comparisons of the SD-based Real-ISR methods in Figs. 1 and 2 to demonstrate that our TVT method can not only preserve image fine-structures but also exhibit good generative capability. As shown in Fig. 1, TVT successfully generates clear sunglasses, while other methods struggle to reproduce complete sunglasses. As shown in Fig. 2, TVT effectively restores the windows in the building, while all the other methods fail. Fig. 3 presents visual comparisons of STISTR methods. It is evident that TVT restores the texts much more accurately than its competitors.

2. Comparisons with GAN-based Methods

We compare TVT with three representative GAN-based SR methods: RealESRGAN [9], BSRGAN [12] and LDL [6]. The quantitative results are presented in Table 1. One can see that our proposed TVT method achieves the best performance on no-reference metrics (CLIPIQA [8], MUSIQ [5], Q-Align [11], TOPIQ [3], HyerIQA [7], and AFINE-NR [4]) on all the three test datasets (DIV2K-val [1], RealSR [2] and DRealSR [10]). For reference-based metrics, TVT also demonstrates competitive results (*e.g.*, LPIPS, DISTS and AFINE-FR).

The visual comparisons are illustrated in Fig. 4. It can be clearly found that the proposed TVT method can generate more realistic details than those GAN-based methods. For example, in the first image, TVT restores clear windows, while the GAN-Based methods struggle to restore windows with complete structures.

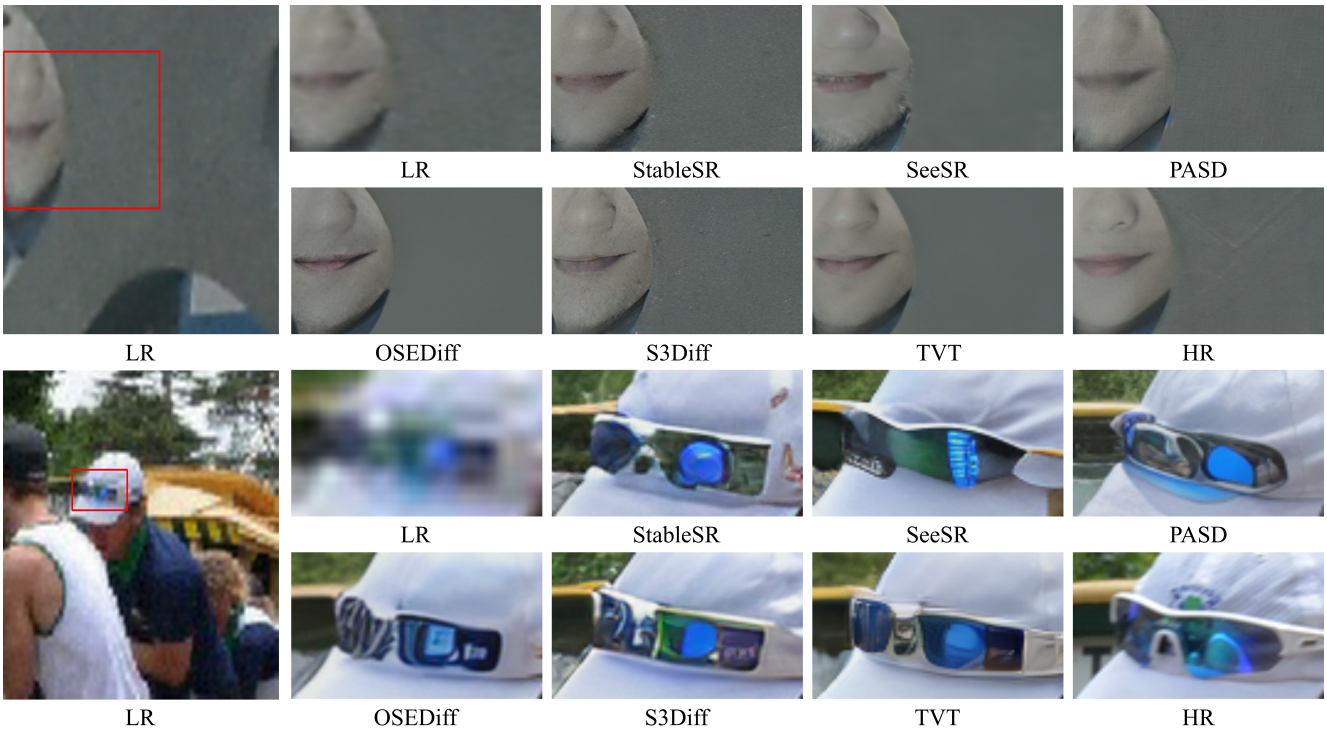

Figure 1. Visual comparison with SD-based Real-ISR methods.

Figure 2. Visual comparison with SD-based Real-ISR methods.

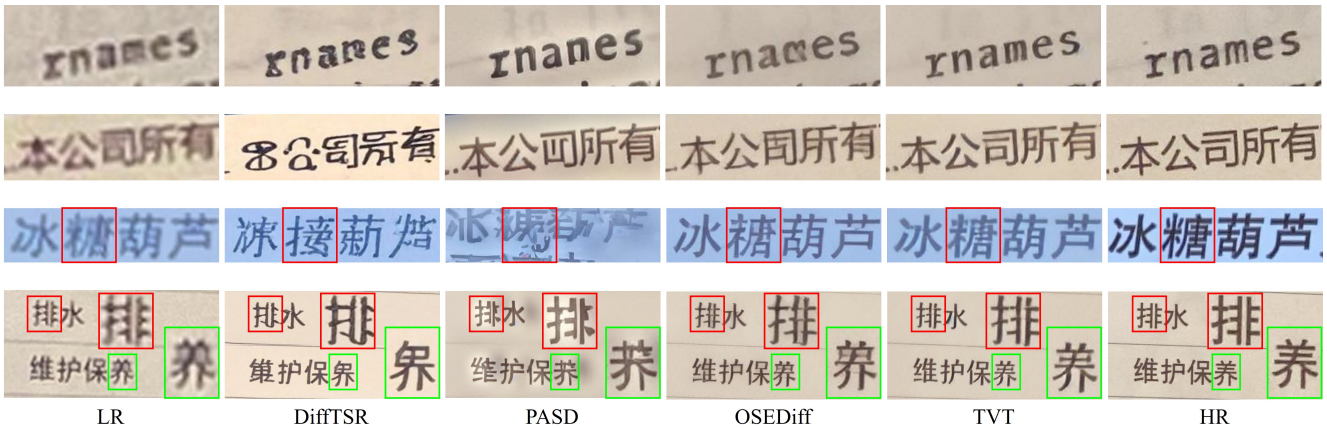

Figure 3. Visual comparison with STISR methods.

Table 1. Quantitative comparison between TVT and the state-of-the-art GAN-based SR methods on synthetic (DIV2K) and real-world (RealSR, DrealSR) test datasets. The best and second best results are highlighted in **red** and **blue**, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	AFINE-FR↓	DISTS↓	CLIPIQA↑	MUSIQ↑	Q-Align↑	TOPIQ↑	HyperIQA↑	AFINE-NR↓
DIV2K	RealESRGAN	**24.29**	**0.6371**	**0.3112**	**-0.5529**	**0.2141**	**0.5277**	61.05	**3.064**	0.5297	0.5664	**-0.8269**
	BSRGAN	**24.58**	0.6269	0.3351	-0.1088	0.2275	0.5247	**61.19**	2.855	**0.5460**	**0.5729**	-0.7550
	LDL	23.83	**0.6344**	0.3256	-0.3922	0.2227	0.5179	60.04	2.986	0.5144	0.5549	-0.8090
	TVT	24.23	0.6292	**0.2773**	**-0.9132**	**0.1860**	**0.6986**	**68.67**	**3.920**	**0.6791**	**0.6794**	**-0.8966**
RealSR	RealESRGAN	25.68	0.7614	0.2710	-0.6811	**0.2060**	0.4490	60.36	3.107	0.5148	0.5216	-0.9132
	BSRGAN	**26.37**	**0.7651**	**0.2656**	**-0.7995**	0.2124	**0.5116**	**63.28**	**3.181**	**0.5626**	**0.5505**	-0.8728
	LDL	25.28	0.7565	0.2750	-0.6386	0.2120	0.4558	60.93	3.085	0.5120	0.5288	**-0.9269**
	TVT	**25.81**	**0.7596**	**0.2587**	**-0.8787**	**0.2061**	**0.6882**	**69.89**	**3.770**	**0.6829**	**0.6761**	**-1.0237**
DrealSR	RealESRGAN	**28.61**	**0.8051**	**0.2819**	-0.7286	**0.2089**	0.4516	54.27	2.869	0.4624	0.4952	-0.7874
	BSRGAN	**28.70**	0.8028	0.2858	**-0.9056**	0.2144	0.5093	**57.16**	**2.957**	**0.5061**	**0.5315**	-0.7628
	LDL	28.19	**0.8124**	**0.2792**	-0.7307	**0.2127**	**0.4476**	53.94	2.850	0.4518	0.4888	**-0.7937**
	TVT	28.27	0.7899	0.2900	**-0.8057**	0.2205	**0.7220**	**65.56**	**3.641**	**0.6591**	**0.6655**	**-0.9073**

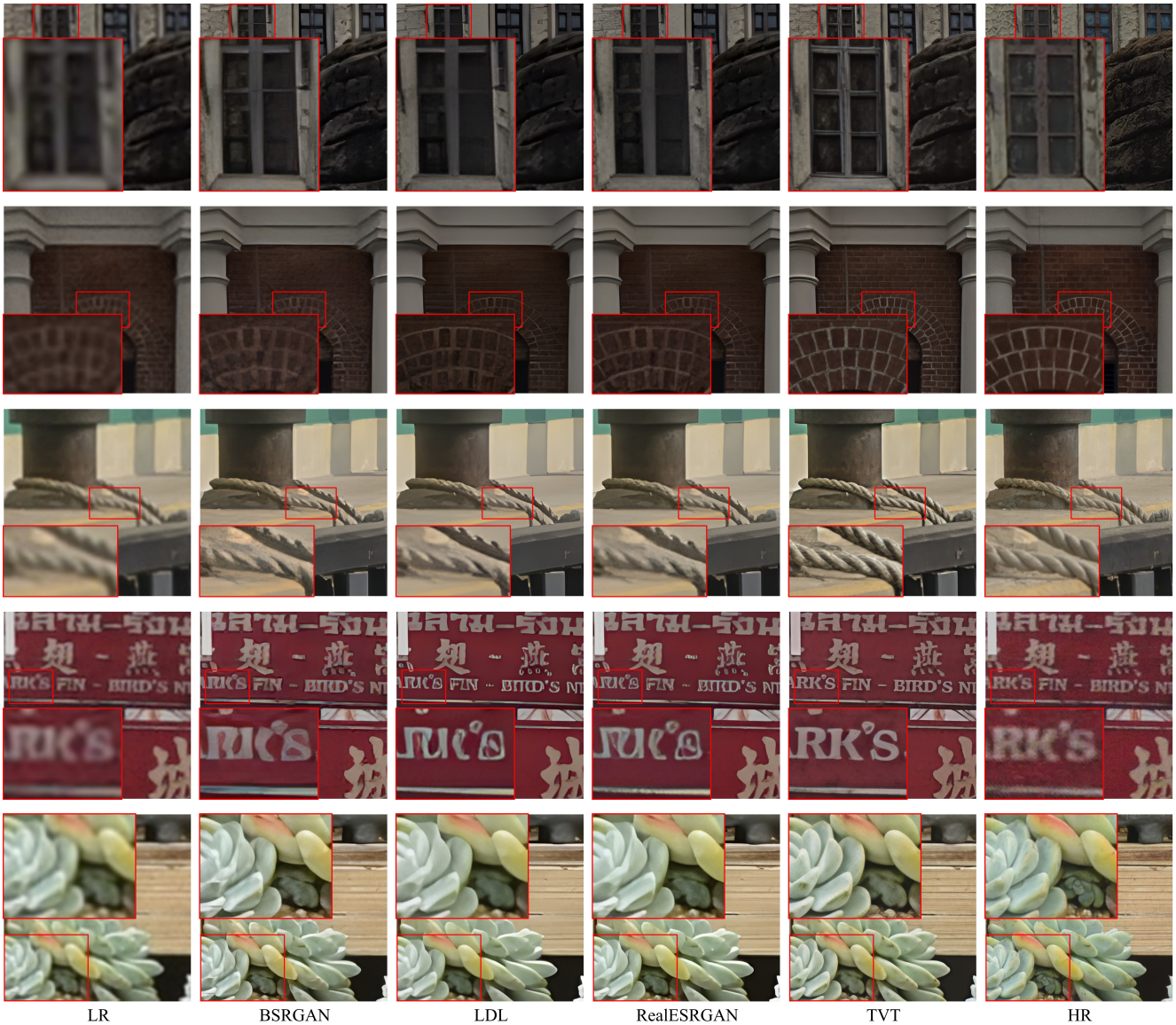

Figure 4. Visual comparison with GAN-based Real-ISR methods.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 1

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2019. 1

[3] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. 1

[4] Du Chen, Tianhe Wu, Kede Ma, and Lei Zhang. Toward generalized image quality assessment: Relaxing the perfect reference quality assumption. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12742–12752, 2025. 1

[5] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021. 1

[6] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5657–5666, 2022. 1

[7] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020. 1

[8] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 1

[9] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1905–1914, 2021. 1

[10] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 101–117. Springer, 2020. 1

[11] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 1

[12] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 1