Supplementary Material

A. Implementation Details

A.1. Data Processing.

Following [4, 7, 19, 20], the super-resolution process is conducted with a scale factor of 4, upsampling images from 128×128 to 512×512. To create the degraded dataset, Ground Truth (GT) images are first randomly cropped from their original sources. Subsequently, these GT images are synthesized into 128x128 degraded data via the well-established Real-ESRGAN [5] degradation pipeline, involving various corruptions such as noise, blurring, and compression. This data processing method is widely adopted and well-established. [15, 17, 21, 24]. Moreover, to minimize memory overhead and accelerate training, we pre-encode the low-quality, high-quality, and teacher-generated references into the VAE's latent representations. These latent representations are then cached, enabling their swift retrieval during the training phase.

A.2. VAE Training.

For the standard Teacher VAE, we first reduce the number of channels in all intermediate layers to 64 and remove all attention mechanisms, following [2]. Subsequently, all standard convolutional layers are substituted with depthwise separable convolutions (*SepConv*), in line with the methodology proposed by [6, 9].

We train the VAE encoder $\mathcal{E}_{tiny}$ by aligning latent space features using MSE loss. The training objective is defined as:

$$\mathcal{L}_{encoder} = \|\mathcal{E}_{tiny}(x_{\mathrm{LR}}) - \mathcal{E}_{pre}(x_{\mathrm{LR}})\|_2^2 \tag{A.1}$$

Here, x_{LR} represents low-quality data, $\mathcal{E}_{\mathrm{pre}}$ is the pre-trained encoder. Training is conducted for 100k steps using a batch size of 64 and a learning rate (AdamW optimizer) of 3e-4 for this phase.

We use LPIPS loss and GAN loss to train the VAE decoder $\mathcal{D}_{tiny}$:

$$\begin{aligned}\mathcal{L}_{decoder} = \lambda_1 \mathcal{L}_{\mathrm{LPIPS}}(\mathcal{D}_{tiny}(\mathcal{E}_{pre}(x_{LR})), x_{HR}) \\ +\lambda_2 \mathcal{L}_{\mathrm{GAN}}(\mathcal{D}_{tiny}(\mathcal{E}_{pre}(x_{LR})))\end{aligned} \tag{A.2}$$

Here, x_{HR} represents high-quality data. We set λ_1 to 3 and λ_2 to 1. We employ a learning rate of 5e-4 for the decoder and 1e-5 for the discriminator during training. We train the model for 200k iterations using 64 batch size setting. The random seed is set to 80 throughout the training.

A.3. Pruning Decision Training.

We initialize our model using the pre-trained weights of TSD-SR [7] for pruning training. We set the pruning rate to 50% to establish our baseline model. Following the TinyFusion [8] approach, we retained two out of every four layers and employed a dynamic block-wise activation mechanism between adjacent layers. Our masks are calculated via the Gumbel-Softmax operation [10]. During network propagation, calculation for a layer is bypassed if its associated mask value is 0. We optimize the network and probability parameters using SR's task loss and distillation loss aligned with the teacher features. Specifically, task loss is defined as LPIPS loss and L_1 loss is utilized for the distillation loss. The total loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_{pruning} =& \lambda_3 \mathcal{L}_{\mathrm{LPIPS}}(\mathcal{D}_{tiny}(z_{stu}), x_{HR}) \\ &+ \lambda_4 \|z_{stu} - z_{tea}\|_1 \\ where \quad & z_{stu} \sim x_{LR} - \epsilon_{stu}(\mathcal{E}_{tiny}(x_{LR}), t), \\ & z_{tea} \sim x_{LR} - \epsilon_{tea}(\mathcal{E}_{tea}(x_{LR}), t)\end{aligned} \tag{A.3}$$

ϵ_{stu} denotes the student's denoising network, while ϵ_{tea} represents the teacher's. t denotes timesteps, and $\mathcal{E}_{tea}$ denotes the teacher encoder. This encoder differs from the pre-trained version $\mathcal{E}_{pre}$ as it is fine-tuned by TSD-SR [7].

Training is conducted for 100k iterations across 8 NVIDIA V100 GPUs, employing a learning rate of 5e-5 (AdamW optimizer) and a global batch size of 8. We use LoRA training, with LoRA rank set to 64. λ_3 and λ_4 are both set to 1.

A.4. Restoration Training.

We perform depth pruning on the TSD-SR according to the pruning mask and discard the condition-related components to initialize our student network. To achieve rapid convergence, we divided the model's training into two stages. In the first stage, training is exclusively conducted within the latent space. We employ L_1 loss for teacher-student knowledge distillation to align features. The formulation of this distillation loss is consistent with that described in Equation (A.3):

$$\mathcal{L}_{stage_1} = \|z_{stu} - z_{tea}\|_1 \tag{A.4}$$

The meaning of z_{stu} and z_{tea} is the same as mentioned above. Training in the latent space enables us to use a larger global batch size (128) on 8 V100 GPUs. We set the learning rate to 1e-4 and the LoRA rank to 64. We iterate training 150k steps until convergence.

In stage 2, we further enhance the perceptual quality of the results in image space by fine-tuning the model directly within the image domain. We additionally incorporate LPIPS loss and GAN loss to enhance image restoration. The total loss is expressed as follows:

$$\begin{aligned}\mathcal{L}_{stage_2} =& \lambda_5 \|z_{stu} - z_{tea}\|_1 \\ &+ \lambda_6 \mathcal{L}_{\mathrm{LPIPS}}(\mathcal{D}_{tiny}(z_{stu}), x_{HR}) \\ &+ \lambda_7 \mathcal{L}_{\mathrm{GAN}}(\mathcal{D}_{tiny}(z_{stu}))\end{aligned} \tag{A.5}$$

The meaning of z_{stu}, z_{tea} and $\mathcal{D}_{tiny}$ is the same as mentioned above. λ_5, λ_6, and λ_7 are set to 5, 1, and 0.3, respectively. We fine-tune our model for 50k steps on 8 V100 GPUs, with a global batch size of 96, a learning rate of 1e-6 for student (5e-6 for discriminator), and a LoRA rank of 64. The random seed for the entire training process is set to 80. And all training is done on *fp16* precision.

B. More Comparisons on Benchmarks

B.1. More Quantitative Comparisons

We compared GAN-based and diffusion-based methods across various datasets (DIV2K-Val [1], DrealSR [18], RealSR [3]), with the results presented in Table A.1. We observe that traditional GAN-based approaches [5, 12, 16, 25] generally excel on full-reference metrics, particularly PSNR and SSIM. However, some studies indicate that PSNR and SSIM often do not accurately reflect fidelity under more complex degradation conditions [7, 21, 23]. In most perceptual quality metrics, such as NIQE [26], MUSIQ [11], MANIQA [22] and CLIPIQA [14], diffusion-based methods demonstrate superior performance compared to these GANs, highlighting their enhanced capability in generating natural textures. TinySR achieved competitive performance across most metrics, demonstrating comparable results to its teacher model, TSD-SR, and showcasing the robust recoverability of the pruning methods.

B.2. More Qualitative Comparisons

Figure B.1 presents a visual comparison between the GAN-based and diffusion-based methods. GAN-based methods often struggle to recover fine, high-frequency details, resulting in blurred textures. For instance, models such as BSRGAN, Real-ESRGAN, LDL, and FeMASR produce blurring on petal textures. Similarly, BSRGAN and LDL create overly smooth butterfly wings, while Real-ESRGAN and FeMASR fail to reconstruct crisp mushroom textures. This consistent lack of detail suggests a fundamental limitation in the ability of these GAN-based approaches to restore high-frequency information. Multi-step diffusion-based methods, such as StableSR, DiffBIR, SeeSR, and ResShift, can introduce artifacts when restoring natural textures like water and rocks, and may also produce blurred details. Notably, DiffBIR is particularly susceptible to over-generation, which can result in illogical or unnatural textures, as has been observed in the restoration of images containing mushrooms. Methods like OSEdiff, AdcSR, and SinSR can suffer from incomplete denoising and are prone to generating broken or fragmented textures during the super resolution process. Our model demonstrates highly competitive performance, excelling in both structural and texture recovery. Compared to other methods, it restores a greater degree of high-frequency detail while rigorously maintaining overall structural integrity.

C. More Ablation Studies.

C.1. Ablation Study on Prompt Condition

Table C.1 presents the results of token pruning of the teacher model. We found that pruning prompt information does not negatively impact certain full-referenced metrics. In fact, some metrics, such as LPIPS and DISTS, even show improvement at specific pruning rates. Token pruning primarily affects no-referenced metrics. However, we observe no significant performance degradation even at a 50% pruning ratio. Furthermore, performance degrades gracefully at higher pruning percentages without a sharp decline, which suggests that the contribution of textual information to the final image synthesis is limited. As shown in Figure C.1, although TP 90% token's output contains less fine-grained detail than the baseline, it effectively removes the noise from the low-quality input, resulting in an image with high visual quality.

C.2. Ablation Study on Pruning Ratio

Table C.2 compares the performance of our method against ShortGPT and TinyFusion across various metrics at token pruning ratios of 33%, 50%, and 67%. The results show our approach surpassing TinyFusion [8] and ShotGPT [13] at every pruning ratio, which demonstrates its robust ability to recover performance. Furthermore, the model exhibits only a slight degradation in performance as the pruning ratio is increased from 33% to 50%, indicating the continued presence of parameter redundancy at the 33% level. However, as the pruning rate increases from 50% to 67%, the model's performance on metrics such as DISTS and MANIQA declines sharply, indicating that excessive pruning leads to irreversible performance degradation.

C.3. Ablation Study of Knowledge Distillation

Table C.3 demonstrates the effectiveness of our knowledge distillation method under stage 1. We can draw the following conclusions: (1) Distillation employing GT (High-Quality) data consistently resulted in unsatisfactory performance, whether applied in the image space or the latent space. As illustrated in Figure C.2, distillation using GT data yields smooth, blurred results, whereas using the teacher produces clearer textures. (2) Distillation performed in the image space achieves better scores on full-reference metrics such as SSIM, LPIPS, and DISTS. Distillation in the latent space yields superior no-reference metrics (MUSIQ, CLIPIQA, TOPIQ and Q-Align), with most even matching those of the teacher model. However, a potential compromise in reference metrics necessitates a second stage of training, which we perform in the image space.

Table A.1. Quantitative comparison among different GAN-based and diffusion-based Real-ISR approaches on both synthetic and real-world benchmarks. "s" denotes the required number of sampling steps in the diffusion-based method. The best and second-best results are highlighted in **bold**, *italic*, respectively

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑	CLIPIQA ↑
DIV2K-Val	BSRGAN	*24.58*	0.6269	0.3502	0.2280	49.55	4.75	61.68	0.5071	0.5386
	Real-ESRGAN	24.02	**0.6387**	0.3150	0.2123	38.87	4.83	61.06	0.5401	0.5251
	LDL	23.83	*0.6344*	0.3256	0.2227	42.29	4.86	60.04	0.5350	0.5180
	FeMASR	23.06	0.5887	0.3126	0.2057	35.87	4.74	60.83	0.5074	0.5997
	StableSR-s200	23.27	0.5722	0.3111	0.2046	*24.95*	4.77	65.78	0.6164	0.6764
	DiffBIR-s50	23.13	0.5717	0.3469	0.2108	33.93	4.61	68.54	**0.6360**	0.7125
	SeeSR-s50	23.73	0.6057	0.3198	0.1953	25.81	4.83	68.49	*0.6198*	0.6899
	ResShift-s15	**24.71**	0.6234	0.3473	0.2253	42.01	6.36	60.63	0.5283	0.5962
	SinSR-s1	24.41	0.6018	0.3262	0.2069	35.55	6.00	62.95	0.5430	0.6501
	OSEDiff-s1	23.72	0.6109	0.2942	0.1975	26.34	4.71	67.31	0.6131	0.6681
	AdcSR-s1	23.74	0.6017	0.2853	0.1899	25.52	4.36	68.00	0.6090	0.6764
	TSD-SR-s1	23.02	0.5808	**0.2673**	**0.1821**	29.16	*4.32*	**71.69**	0.6192	**0.7416**
	TinySR-s1 (Ours)	22.76	0.5725	*0.2793*	*0.1883*	**24.44**	**4.15**	*69.90*	0.6083	*0.7201*
DRealSR	BSRGAN	**28.70**	0.8028	0.2858	0.2143	155.61	6.54	57.15	0.4847	0.5091
	Real-ESRGAN	28.61	*0.8051*	*0.2818*	**0.2088**	147.66	6.70	54.27	0.4888	0.4512
	LDL	28.20	**0.8124**	**0.2791**	*0.2127*	155.51	7.14	53.94	0.4894	0.4476
	FeMASR	26.87	0.7569	0.3156	0.2238	157.72	*5.91*	53.70	0.4413	0.5633
	StableSR-s200	28.04	0.7454	0.3279	0.2272	144.15	6.60	58.53	0.5603	0.6250
	DiffBIR-s50	25.93	0.6525	0.4518	0.2761	177.04	6.23	65.66	**0.6296**	0.6860
	SeeSR-s50	28.14	0.7712	0.3141	0.2297	146.95	6.46	64.74	*0.6022*	0.6893
	ResShift-s15	*28.69*	0.7874	0.3525	0.2541	176.77	7.88	52.40	0.4756	0.5413
	SinSR-s1	28.38	0.7499	0.3669	0.2484	172.72	6.96	55.03	0.4904	0.6412
	OSEDiff-s1	27.92	0.7836	0.2968	0.2162	135.51	6.45	64.69	0.5898	0.6958
	AdcSR-s1	28.10	0.7726	0.3046	0.2200	**134.05**	6.45	*66.26*	0.5927	0.7049
	TSD-SR-s1	27.77	0.7559	0.2967	0.2136	*134.98*	*5.91*	**66.62**	0.5874	**0.7344**
	TinySR-s1 (Ours)	27.48	0.7459	0.3116	0.2204	146.70	**5.67**	65.36	0.5804	*0.7094*
RealSR	BSRGAN	26.38	**0.7651**	**0.2656**	0.2121	141.24	5.64	63.28	0.5425	0.5114
	Real-ESRGAN	**26.65**	*0.7603*	*0.2726*	**0.2065**	136.29	5.85	60.45	0.5507	0.4518
	LDL	25.28	0.7565	0.2750	0.2119	142.74	5.99	60.92	0.5494	0.4559
	FeMASR	25.07	0.7356	0.2936	0.2285	141.01	5.77	59.05	0.4872	0.5405
	StableSR-s200	24.62	0.7041	0.3070	0.2156	128.54	5.78	65.48	0.6223	0.6198
	DiffBIR-s50	24.24	0.6650	0.3469	0.2300	134.56	5.49	68.35	**0.6544**	0.6961
	SeeSR-s50	25.21	0.7216	0.3003	0.2218	125.10	5.40	69.69	*0.6443*	0.6671
	ResShift-s15	*26.39*	0.7567	0.3158	0.2432	149.59	6.87	60.22	0.5419	0.5496
	SinSR-s1	26.27	0.7351	0.3217	0.2341	137.59	6.30	60.76	0.5418	0.6163
	OSEDiff-s1	25.15	0.7341	0.2920	0.2128	123.48	5.65	69.10	0.6326	0.6687
	AdcSR-s1	25.47	0.7301	0.2885	0.2129	118.41	5.35	*69.90*	0.6360	0.6731
	TSD-SR-s1	24.81	0.7172	0.2743	*0.2104*	**114.45**	*5.13*	**71.19**	0.6347	**0.7160**
	TinySR-s1 (Ours)	24.79	0.7171	0.2806	0.2123	*118.00*	**4.74**	69.78	0.6235	*0.7035*

Table C.1. Ablation study of prompt token pruning (TP) on DrealSR dataset. The best is highlighted in **bold**.

Method	PSNR ↑	LPIPS ↓	DISTS ↓	NIQE ↓	MUSIQ ↑	MANIQA ↑
Baseline	**27.77**	0.2967	0.2136	**5.9131**	**66.62**	**0.5927**
TP 10% token	27.66	0.2945	0.2135	5.9536	66.57	0.5870
TP 25% token	27.65	0.2947	0.2135	5.9510	66.53	0.5861
TP 50% token	27.66	0.2924	**0.2120**	5.9350	66.35	0.5801
TP 75% token	27.62	**0.2860**	0.2122	6.0142	65.79	0.5783
TP 90% token	27.59	0.2866	0.2144	6.1461	65.01	0.5721

Table C.2. Ablation study of pruning ratio on DIV2K-Val dataset. The best is highlighted in **bold**.

Method	Pruning Ratio	LPIPS ↓	DISTS ↓	NIQE ↓	MANIQA ↑	CLIPIQA ↑
ShortGPT	33%	0.3049	0.2150	5.1010	0.5727	0.7248
TinyFusion	33%	0.2808	0.1928	4.3013	0.5912	0.7274
Ours	33%	**0.2789**	**0.1917**	**4.1396**	**0.6071**	**0.7284**
ShortGPT	50%	0.2892	0.2034	4.8874	0.5681	0.7116
TinyFusion	50%	0.2799	0.1904	4.1705	0.5880	0.6995
Ours	50%	**0.2793**	**0.1883**	**4.1500**	**0.6083**	**0.7201**
ShortGPT	67%	0.3466	0.2476	5.5182	0.5257	0.7069
TinyFusion	67%	0.3071	0.2194	4.7256	0.5217	0.7056
Ours	67%	**0.2984**	**0.2110**	**4.2475**	**0.5389**	**0.7198**

C.4. Ablation Study of Losses in Stage 2

We conduct an ablation study on the Stage 2 training losses, as shown in Table C.4. The results indicate that Stage 2 training significantly improved image quality, particularly in terms of image fidelity. Specifically, we find that the inclusion of LPIPS loss is highly beneficial for improving

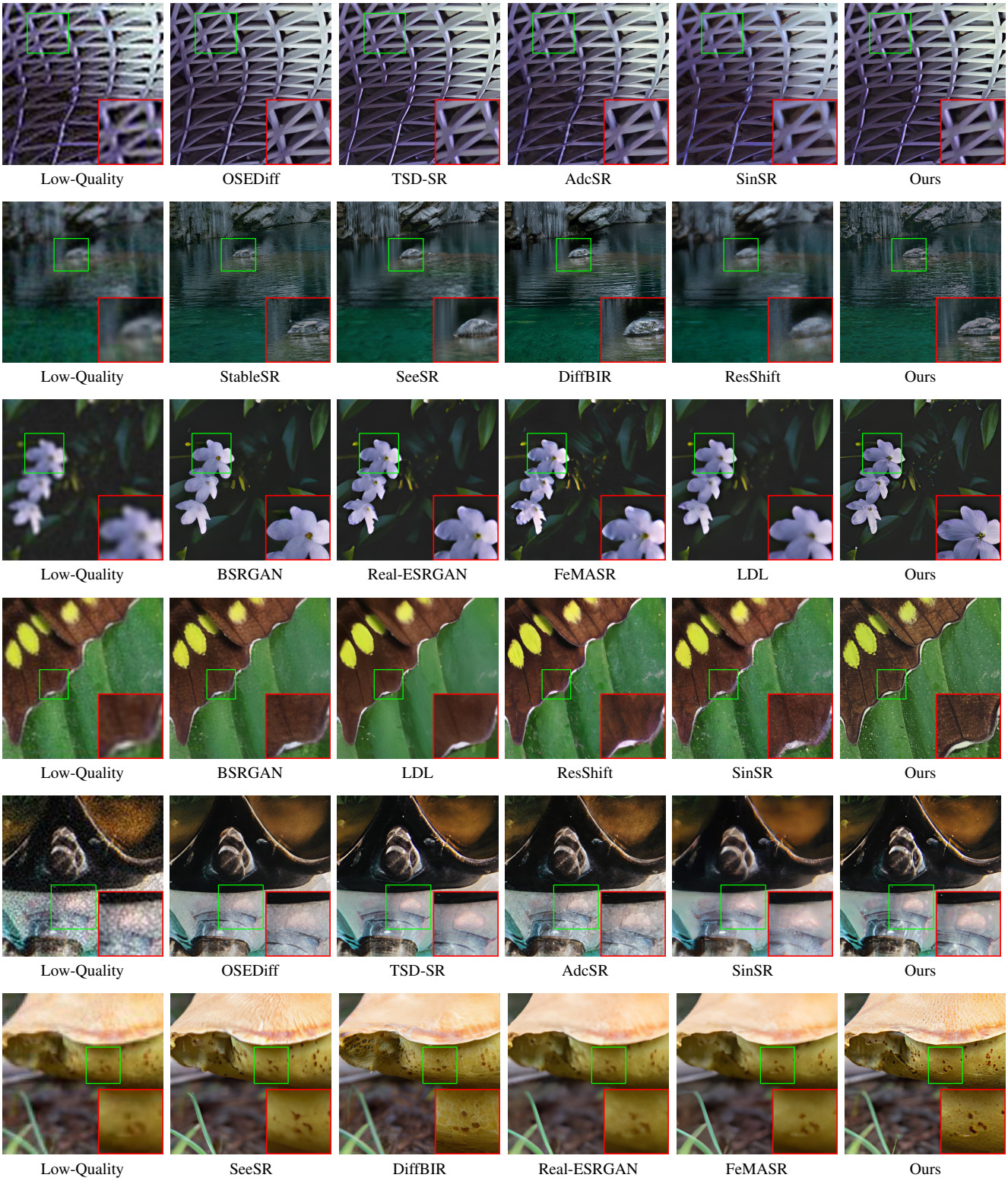
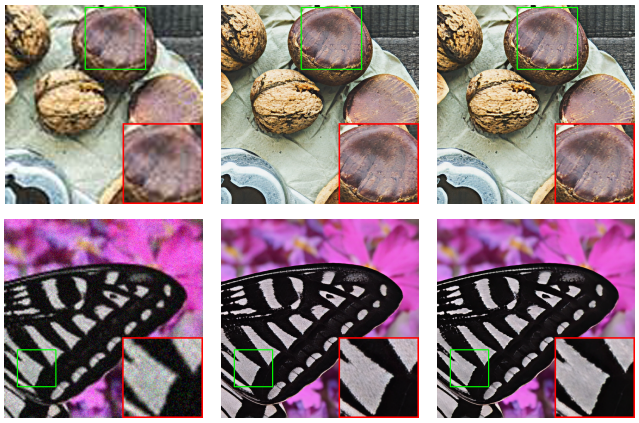


Figure B.1. Qualitative comparisons of GAN-based and diffusion-based Real-ISR methods. Please zoom in for a better view.

reference metrics such as DISTS and FID. The addition of GAN loss, in turn, is helpful for enhancing several no-

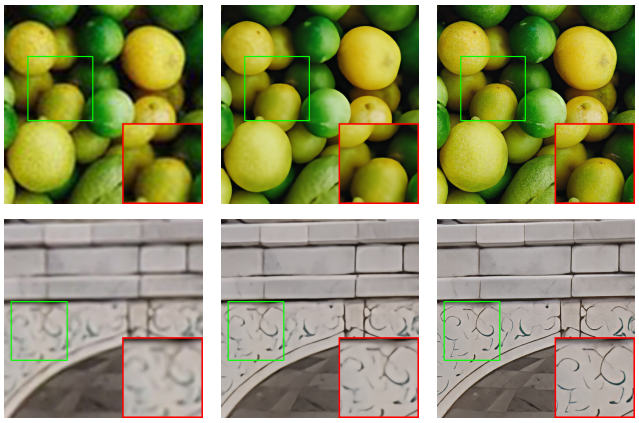
Table C.3. Ablation studies of Stage 1 distillation loss on DrealSR dataset. The best (**other than Teacher**) is highlighted in **bold**.

Method	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	MUSIQ ↑	CLIPIQA ↑	TOPIQ ↑	Q-Align ↑
Teacher TSD-SR Baseline	0.7559	0.2967	0.2136	134.98	66.62	0.7344	0.6177	3.6055
Distill HR in Image Space	**0.8480**	0.3082	0.2438	176.50	48.89	0.3456	0.3652	2.4315
Distill TEA. in Image Space	0.7904	**0.2819**	**0.2150**	154.16	64.74	0.6171	0.6020	3.2871
Distill HR in Latent Space	0.7814	0.4253	0.2988	190.11	48.41	0.4441	0.4726	2.3841
Distill TEA. in Latent Space (Ours)	0.7508	0.3316	0.2322	**148.63**	**66.57**	**0.7321**	**0.6211**	**3.5356**



(a) (a) Low-Quality (b) (b) Baseline (c) (c) 90% TP Result

Figure C.1. Applying 90% token pruning (TP) yields visually comparable results to the baseline with a slight quality drop, indicating the limited contribution of the default prompt.



(a) (a) Low-Quality (b) (b) Distill HR (c) (c) Distill Tea.

Figure C.2. Visual comparison of knowledge distillation: high-resolution ground truth versus teacher.

reference metrics, including NIQE and MANIQA. We ultimately weighted the two new losses to balance the trade-off between fidelity and the generative ability.

Table C.4. Ablation studies of Stage 2 training loss on RealSR dataset. The best is highlighted in **bold**.

Method	LPIPS ↓	DISTS ↓	NIQE ↓	MANIQA ↑	FID ↓
Stage 1 Baseline	0.3087	0.2302	5.1740	0.6045	132.53
w/ LPIPS loss & w/o GAN	**0.2702**	0.2180	5.0696	0.5850	123,43
w/ GAN loss & w/o LPIPS	0.2844	0.2173	**4.7203**	0.6073	124.22
w/ LPIPS & w/ GAN (Ours)	0.2806	**0.2123**	4.7400	**0.6235**	**118.01**

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. 2

[2] Ollin Boer Bohan. Tiny autoencoder for stable diffusion. https://github.com/madebyollin/taesd, 2023. 1

[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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3086–3095, 2019. 2

[4] Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. *arXiv preprint arXiv:2411.13383*, 2024. 1

[5] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1329–1338, 2022. 1, 2

[6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 1

[7] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23174–23184, 2025. 1, 2

[8] Gongfan Fang, Kunjun Li, Xinyin Ma, and Xinchao Wang. Tinyfusion: Diffusion transformers learned shallow. *arXiv preprint arXiv:2412.01199*, 2024. 1, 2

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical

reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 1

[11] 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. 2

[12] 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. 2

[13] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024. 2

[14] 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. 2

[15] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, pages 1–21, 2024. 1

[16] 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. 2

[17] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 1

[18] 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. 2

[19] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 1

[20] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25456–25467, 2024. 1

[21] Rui Xie, Ying Tai, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 1, 2

[22] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 2

[23] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25669–25680, 2024. 2

[24] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[25] 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. 2

[26] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 2