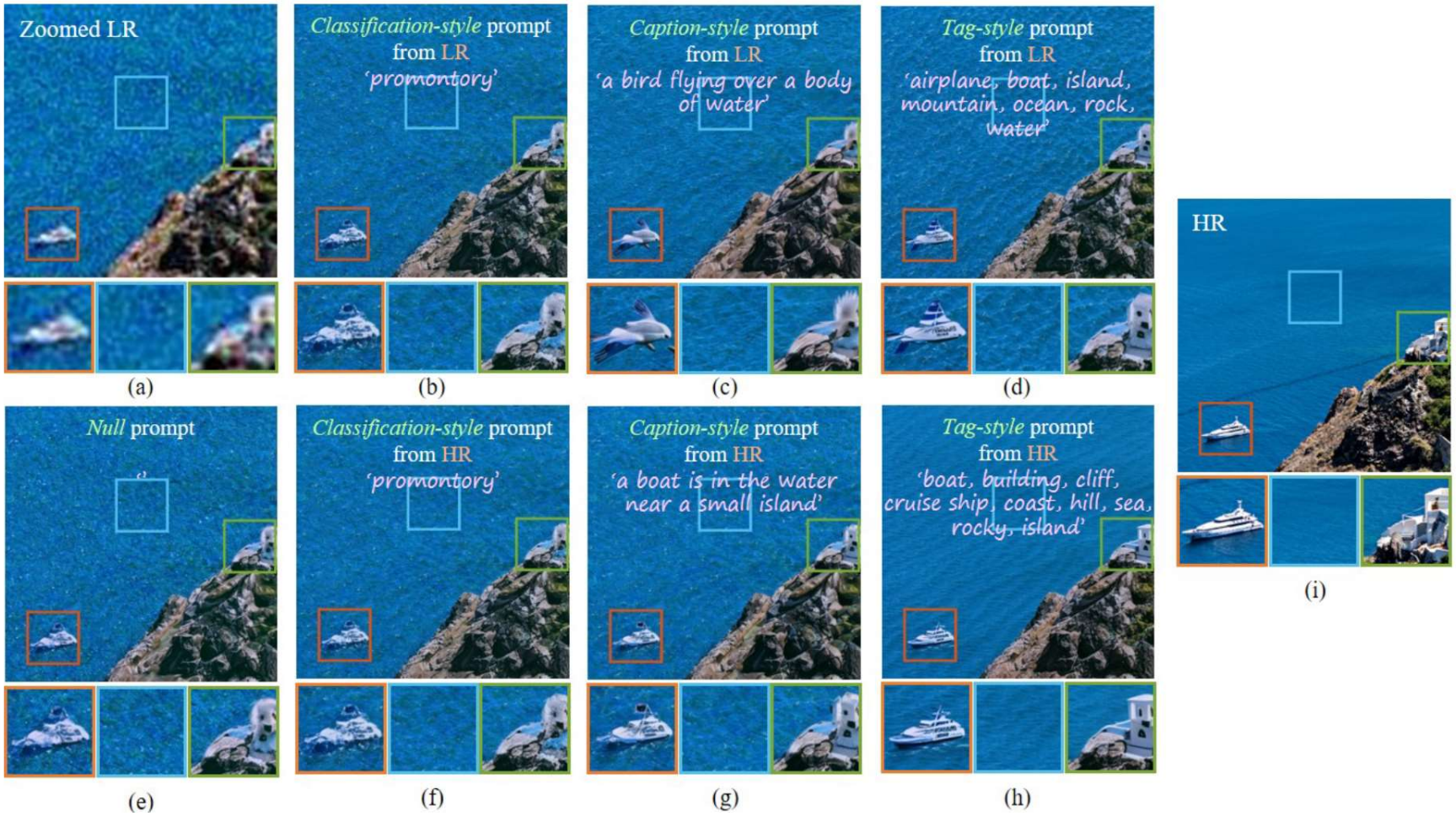
SeeSR: Towards Semantics-Aware Real-World Image Super-Resolution

Rongyuan Wu[1,2], Tao Yang[3], Lingchen Sun[1,2], Zhengqiang Zhang[1,2], Shuai Li[1,2], Lei Zhang[1,2,*]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute [3]ByteDance Inc.
{rong-yuan.wu, ling-chen.sun, zhengqiang.zhang, novak.li}@connect.polyu.hk
yangtao9009@gmail.com, cslzhang@comp.polyu.edu.hk

Wu, Rongyuan, et al. "Seesr: Towards semantics-aware real-world image super-resolution." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024.

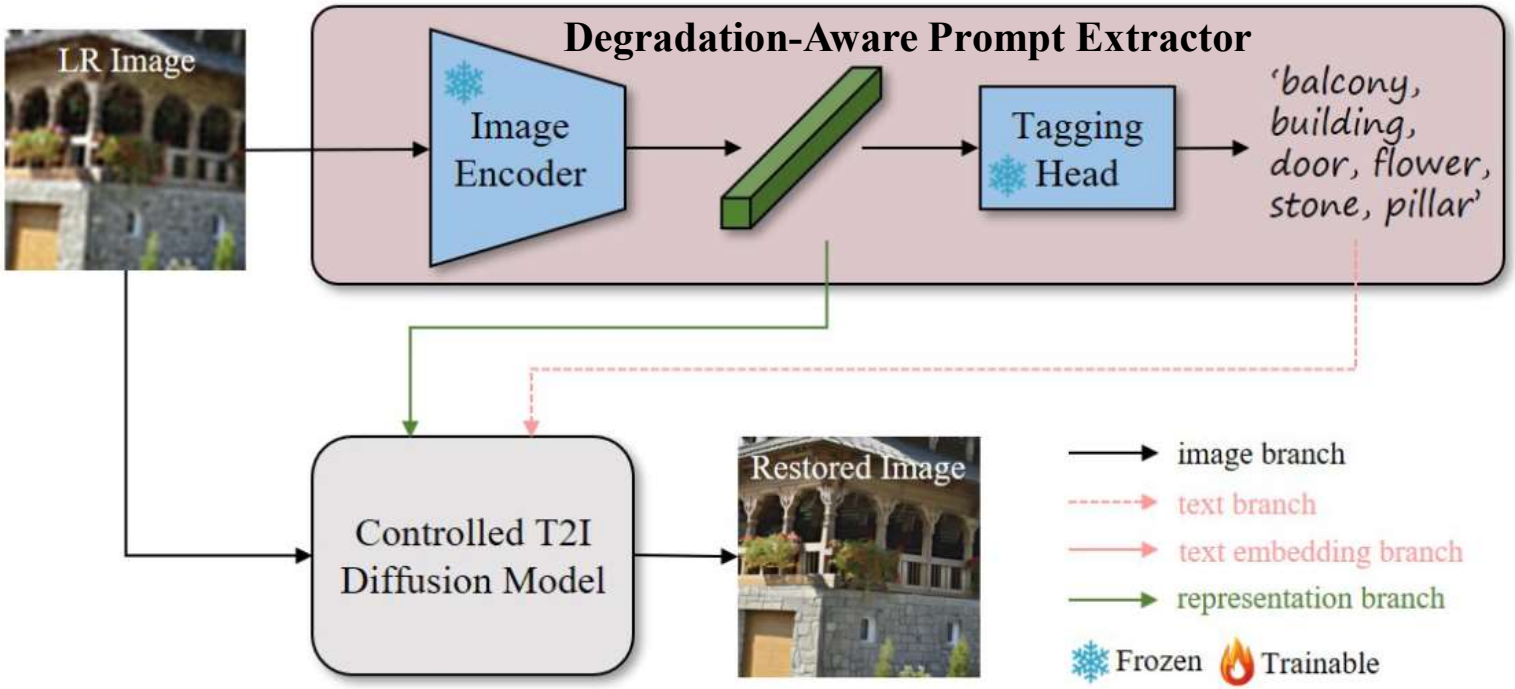


He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

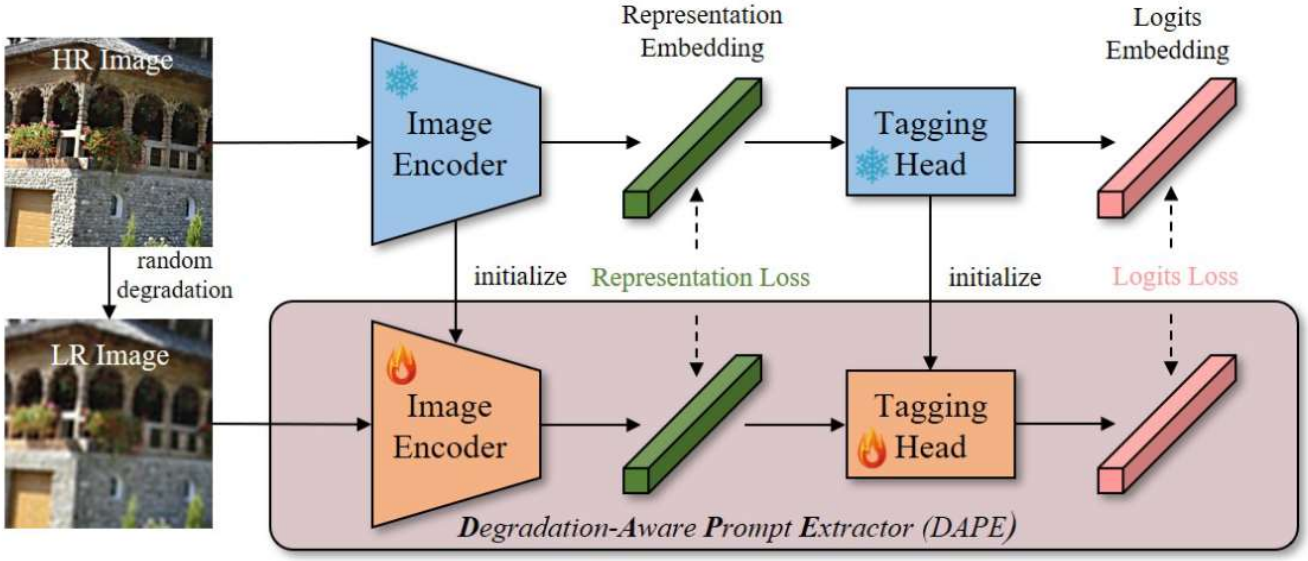
Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International conference on machine learning. PMLR, 2022.

Zhang, Youcai, et al. "Recognize anything: A strong image tagging model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

Yang, Tao, et al. "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization." arXiv preprint arXiv:2308.14469 (2023).



Real-ISR with DAPE



(a) Degradation-aware prompt extractor

balance parameter

representation embdedding

logits embdedding

$$\mathcal{L}_{DAPE} = \mathcal{L}_r(f_y^{rep}, f_x^{rep}) + \lambda \mathcal{L}_l(f_y^{logits}, f_x^{logits}), \quad (1)$$

MSE loss

cross-entropy loss

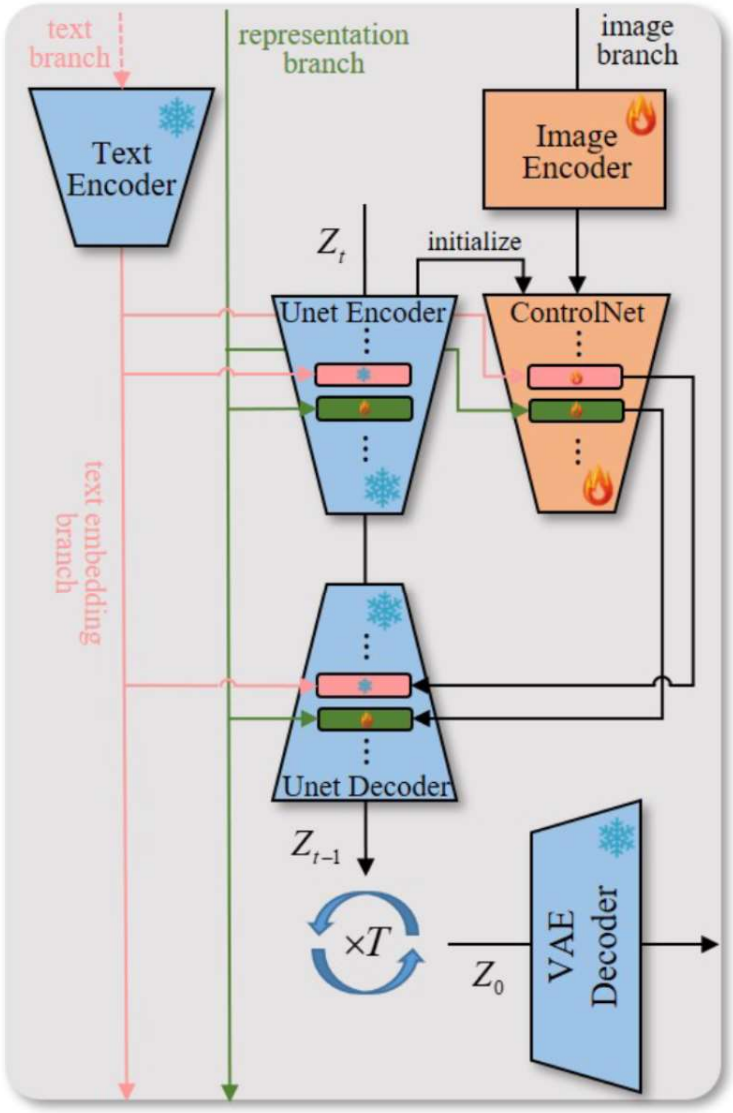
Table 1. Comparison of different prompt styles.

	Rich Objects	Concise Description	Degradation Aware
Classification-style	✗	✓	✓
Caption-style	✓	✗	✗
Tag-style	✓ ✓	✓	✗
Our DAPE	✓ ✓	✓	✓

Zhang, Youcai, et al. "Recognize anything: A strong image tagging model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

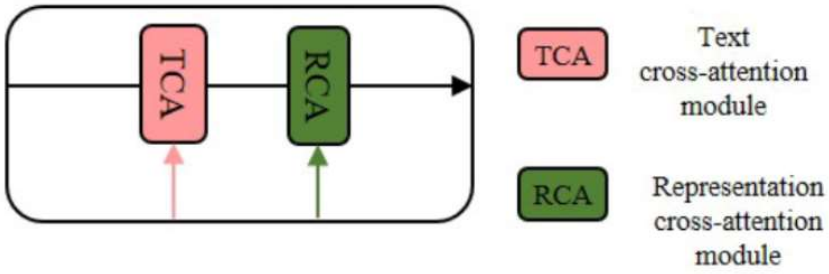
Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Stable Diffusion



(c) Controlled T2I diffusion model

ControlNet



$$\mathcal{L} = \mathbb{E}_{z_0, z_{lr}, t, p_h, p_s, \epsilon \sim \mathcal{N}} \left[\| \epsilon - \epsilon_\theta \left(\mathbf{z}_t, z_{lr}, t, p_h, p_s \right) \|_2^2 \right]$$

z_t: noisy latent

z_{lr}: LR latent

t: current timestep

p_h: hard prompts

p_s: soft prompts

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

LR Embedding in Inference

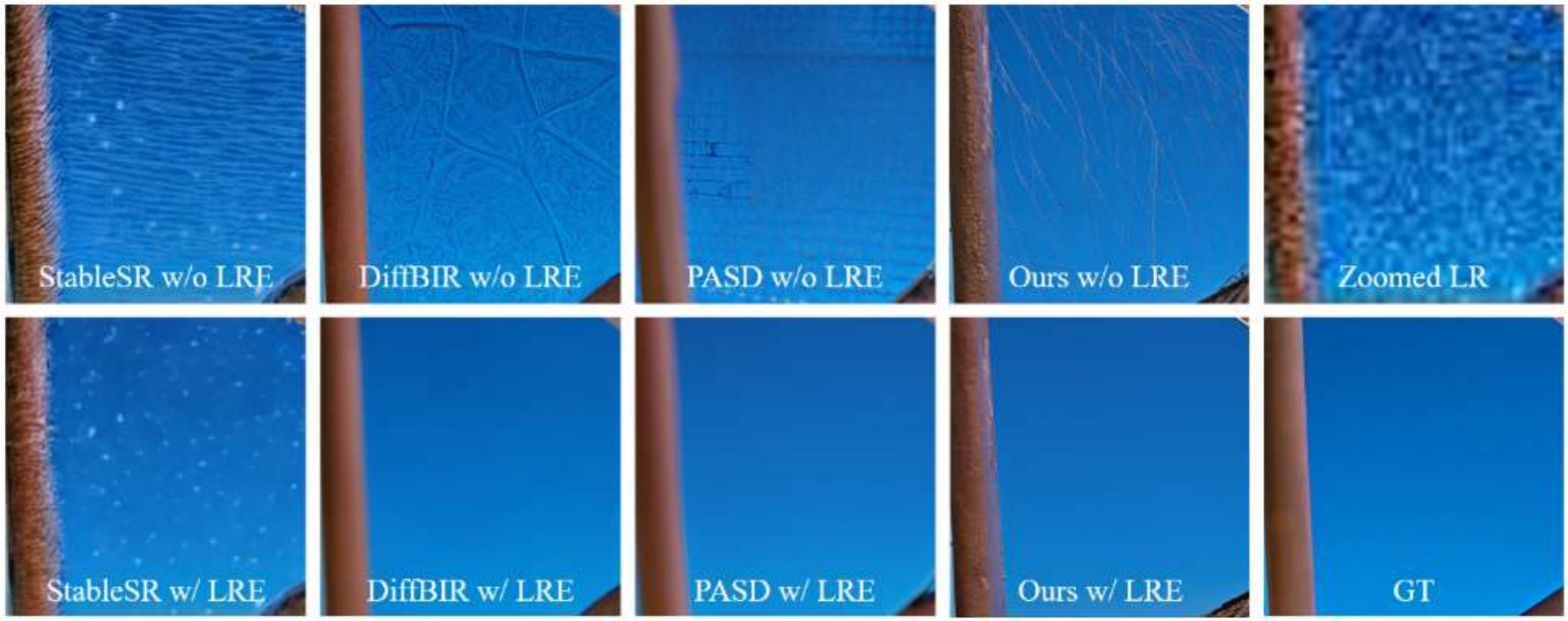


Figure 3. Effectiveness of the LR embedding (LRE) strategy in alleviating the discrepancy between training and inference of SD-based Real-ISR methods [41, 57, 67]. Top row: results without using LRE. Bottom row: results with LRE. We see that many falsely generated details in the sky area are removed.

		GAN-based					Diffusion-based					
Datasets	Metrics	BSRGAN [72]	Real- [59] ESRGAN	LDL [35]	DASR [36]	FeMaSR [5]	LDM [47]	StableSR [57]	ResShift [68]	PASD [67]	DiffBIR [41]	SeeSR
DIV2K-Val	PSNR ↑	21.87	**21.94**	21.52	21.72	20.85	21.26	20.84	21.75	20.77	20.94	21.19
	SSIM ↑	0.5539	**0.5736**	0.5690	0.5536	0.5163	0.5239	0.4887	0.5422	0.4958	0.4938	0.5386
	LPIPS ↓	0.4136	0.3868	0.3995	0.4266	0.3973	0.4154	0.4055	0.4284	0.4410	0.4270	**0.3843**
	DISTS ↓	0.2737	0.2601	0.2688	0.2688	0.2428	0.2500	0.2542	0.2606	0.2538	0.2471	**0.2257**
	FID ↓	64.28	53.46	58.94	67.22	53.70	41.93	36.57	55.77	40.77	40.42	**31.93**
	NIQE ↓	4.7615	4.9209	5.0249	4.8596	**4.5726**	6.4667	4.6551	6.9731	4.8328	4.7211	4.9275
	MANIQA ↑	0.4834	0.5251	0.5127	0.4346	0.4869	0.5237	0.5914	0.5232	0.6049	**0.6205**	0.6198
	MUSIQ ↑	59.11	58.64	57.90	54.22	58.10	56.52	62.95	58.23	66.85	65.23	**68.33**
	CLIPIQA ↑	0.5183	0.5424	0.5313	0.5241	0.5597	0.5695	0.6486	0.5948	0.6799	0.6664	**0.6946**
RealSR	PSNR ↑	26.39	25.69	25.28	**27.02**	25.07	25.48	24.70	26.31	24.29	24.77	25.18
	SSIM ↑	0.7654	0.7616	0.7567	**0.7708**	0.7358	0.7148	0.7085	0.7421	0.6630	0.6572	0.7216
	LPIPS ↓	**0.2670**	0.2727	0.2766	0.3151	0.2942	0.3180	0.3018	0.3460	0.3435	0.3658	0.3009
	DISTS ↓	0.2121	**0.2063**	0.2121	0.2207	0.2288	0.2213	0.2135	0.2498	0.2259	0.2310	0.2223
	FID ↓	141.28	135.18	142.71	132.63	141.05	132.72	128.51	141.71	129.76	128.99	**125.55**
	NIQE ↓	5.6567	5.8295	6.0024	6.5311	5.7885	6.5200	5.9122	7.2635	**5.3628**	5.5696	5.4081
	MANIQA ↑	0.5399	0.5487	0.5485	0.3878	0.4865	0.5423	0.6221	0.5285	**0.6493**	0.6253	0.6442
	MUSIQ ↑	63.21	60.18	60.82	40.79	58.95	58.81	65.78	58.43	68.69	64.85	**69.77**
	CLIPIQA ↑	0.5001	0.4449	0.4477	0.3121	0.5270	0.5709	0.6178	0.5444	0.6590	0.6386	**0.6612**
DrealSR	PSNR ↑	28.75	28.64	28.21	**29.77**	26.90	27.98	28.13	28.46	27.00	26.76	28.17
	SSIM ↑	0.8031	0.8053	0.8126	**0.8264**	0.7572	0.7453	0.7542	0.7673	0.7084	0.6576	0.7691
	LPIPS ↓	0.2883	0.2847	**0.2815**	0.3126	0.3169	0.3405	0.3315	0.4006	0.3931	0.4599	0.3189
	DISTS ↓	0.2142	**0.2089**	0.2132	0.2271	0.2235	0.2259	0.2263	0.2656	0.2515	0.2749	0.2315
	FID ↓	155.63	147.62	155.53	155.58	157.78	156.01	148.98	172.26	159.24	166.79	**147.39**
	NIQE ↓	6.5192	6.6928	7.1298	7.6039	5.9073	7.1677	6.5354	8.1249	**5.8595**	6.2935	6.3967
	MANIQA ↑	0.4878	0.4907	0.4914	0.3879	0.4420	0.5043	0.5591	0.4586	0.5850	0.5923	**0.6042**
	MUSIQ ↑	57.14	54.18	53.85	42.23	53.74	53.73	58.42	50.60	64.81	61.19	**64.93**
	CLIPIQA ↑	0.4915	0.4422	0.4310	0.3684	0.5464	0.5706	0.6206	0.5342	0.6773	0.6346	**0.6804**
RealLR200	NIQE ↓	4.3817	4.2048	4.3845	4.3360	4.6357	-	4.2516	6.2878	4.1715	4.9330	**4.1620**
	MANIQA ↑	0.5462	0.5582	0.5519	0.4877	0.5295	-	0.5841	0.5417	0.6066	0.5902	**0.6254**
	MUSIQ ↑	64.87	62.94	63.11	55.67	64.14	-	63.30	60.18	68.20	62.06	**69.71**
	CLIPIQA ↑	0.5679	0.5389	0.5326	0.4659	0.6522	-	0.6068	0.6486	0.6797	0.6509	**0.6813**

Figure 4. Qualitative comparisons of different Real-ISR methods. Please zoom in for a better view.

Table 3. The comparison of semantic restoration performance among different Real-ISR methods.

Metrics	GT	Zoomed LR	BSRGAN	Real-ESRGAN	LDL	DASR	FeMaSR	LDM	StableSR	ResShift	PASD	DiffBIR	SeeSR
Panoptic Segmentation (PQ)	52.5	9	16.2	19.4	17.8	15.5	15.6	18.7	26.8	21.4	23.7	27.2	**30.0**
Object Detection (AP)	49.1	5	10.5	13.1	11.9	9.9	10.1	11.4	18.3	14.3	15.5	18.9	**21.1**
Instance Segmentation (AP)	43.8	4	9.2	11.4	10.3	8.6	8.8	9.9	16.2	12.4	13.4	16.5	**18.5**
Semantic Segmentation (mIOU)	62.0	12	21.7	26.0	24.2	20.5	20.4	25.3	34.5	30.4	33.3	37.7	**41.3**

We resize the original images from COCO-Val (5K images) to 512 × 512 as GT, and then degrade them to generate LR images as in training. We employ OpenSeeD trained on COCO as the detector and segmentor since it is a strong transformer-based unified model for segmentation and detection tasks.

Zhang, Hao, et al. "A simple framework for open-vocabulary segmentation and detection." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

Table 4. Results of user study on synthetic and real-world data.

Methods	Confusion rates on synthetic data	Best rates on real-world data
Real-ESRGAN	5.4%	0%
StableSR	5.4%	14.3%
ResShift	3.6%	0%
PASD	10.7%	13.4%
DiffBIR	12.5%	15.2%
SeeSR	**38.6%**	**57.1%**

在合成数据上，参与者每次都会看到放置在两个 HR 图像之间的 LR 图像：一个是 GT，另一个是一个模型的 Real-ISR 输出。他们被要求确定“哪个 HR 图像更适合 LR 图像？”在做出决定时，参与者被要求考虑两个因素：HR 图像的感知质量及其与 LR 图像的语义相似性。然后可以计算混淆率，这表明参与者对 GT 或 Real-ISR 输出的偏好。

根据真实数据，向参与者提供了 LR 图像以及所有 Real-ISR 输出，并要求他们回答“哪张图像是 LR 图像的最佳 SR 结果？”表示模型被选择的概率。

我们邀请 20 名参与者测试六种代表性方法（Real-ESRGAN、StableSR、PASD、DiffBIR、ResShift 和 SeeSR）。有 16 个综合测试集和 16 个真实世界测试集。合成数据从 DIV2K-Val 中随机采样，真实世界数据从 RealLR200 中随机采样。 20 名参与者每人被要求做出 112 个选择（16×6+16）。