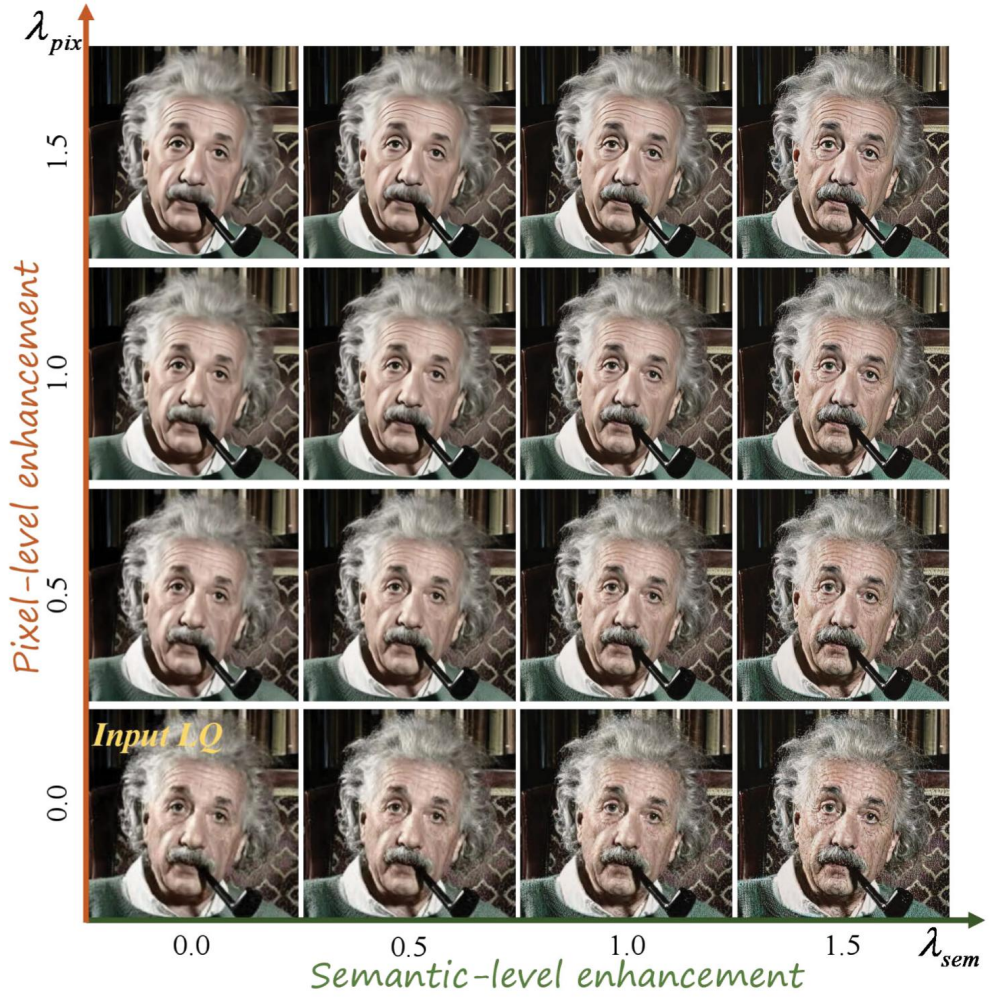
Pixel-level and Semantic-level Adjustable Super-resolution: A Dual-LoRA Approach

Lingchen Sun[1,2], Rongyuan Wu[1,2], Zhiyuan Ma[1], Shuaizheng Liu[1,2], Qiaosi Yi[1,2], Lei Zhang[1,2*]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute

Accepted by CVPR 2025

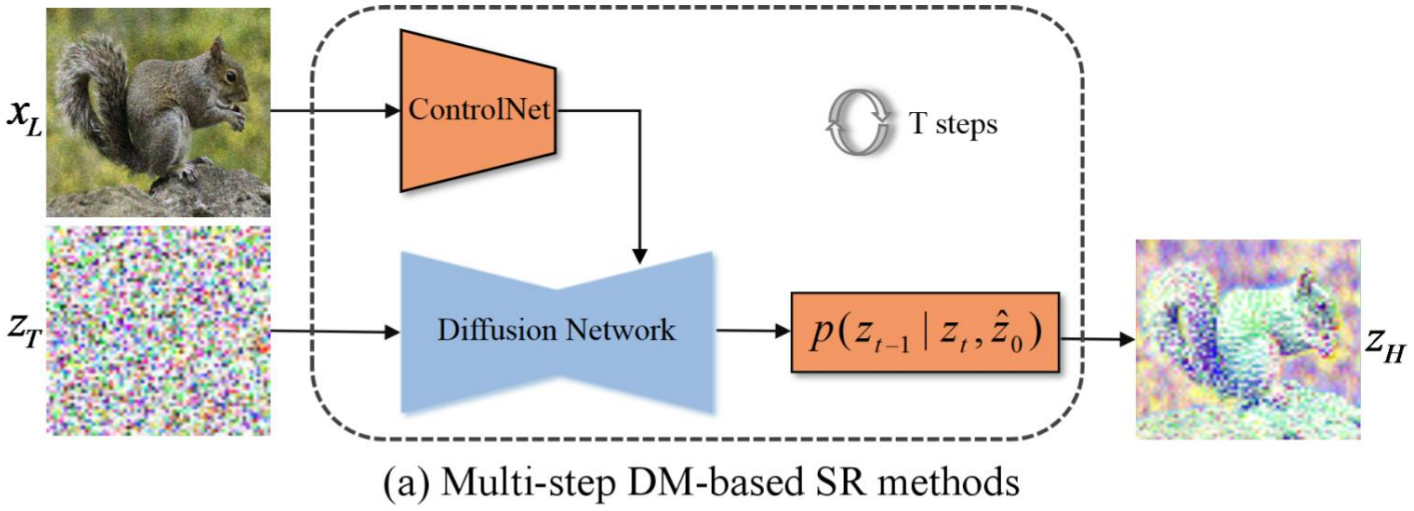
一、介绍



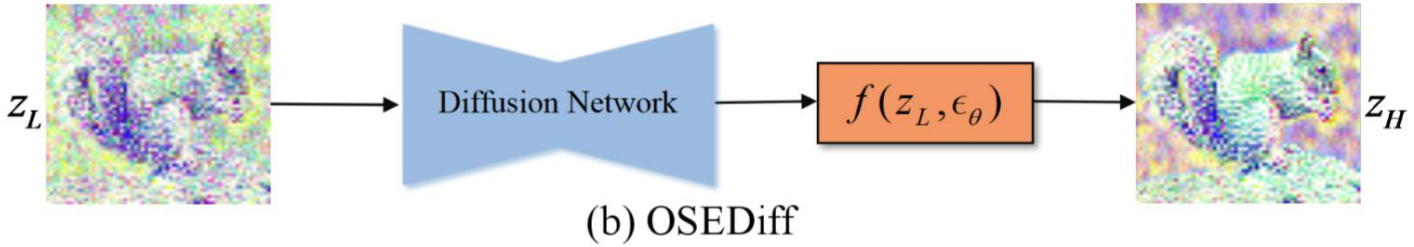
- 通过增加像素级引导尺度 λ_{pix}，可以逐渐消除图像的退化，如噪声和压缩伪影
  - 然而，λ_{pix}过高将使SR图像结果过度平滑
- 通过增加语义级引导尺度 λ_{sem}，SR图像具有更多的语义细节
  - 然而，过高的 λ_{sem} 会产生视觉伪影

- 现有基于Stable Diffusion的SR方法在感知上表现出了更真实可信的结果，但它们通常会混淆像素级保真度和语义级增强的目标，在训练过程中将像素级和语义级的SR目标纠缠在一起，在保证像素级保真度和感知质量之间难以取得平衡。
- 用户在实际应用中对SR结果的偏好也各不相同
  - 一些用户更注重内容保真度而非细节生成
  - 另一些用户则偏好丰富的语义细节而非像素 Level 的保真度。

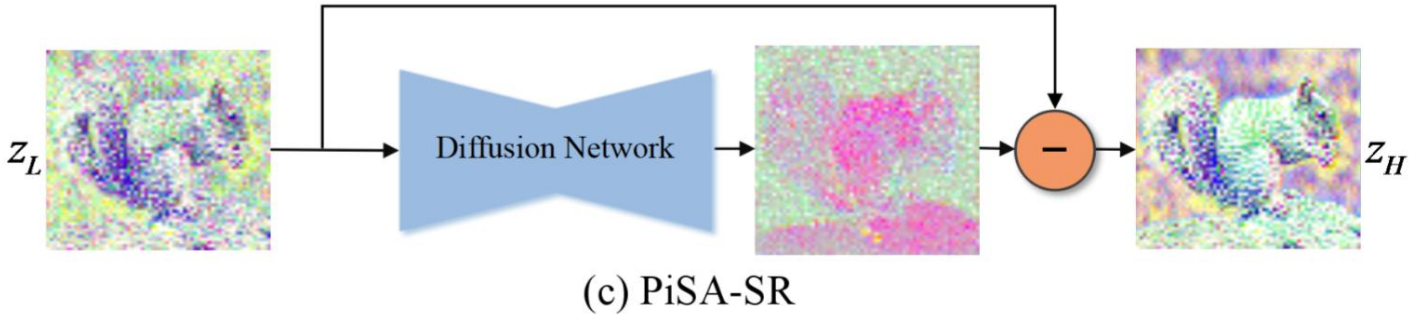
二、模型描述



(a) Multi-step DM-based SR methods



(b) OSEDiff



(c) PiSA-SR

Comparison of the pipeline of different DM-based SR methods.

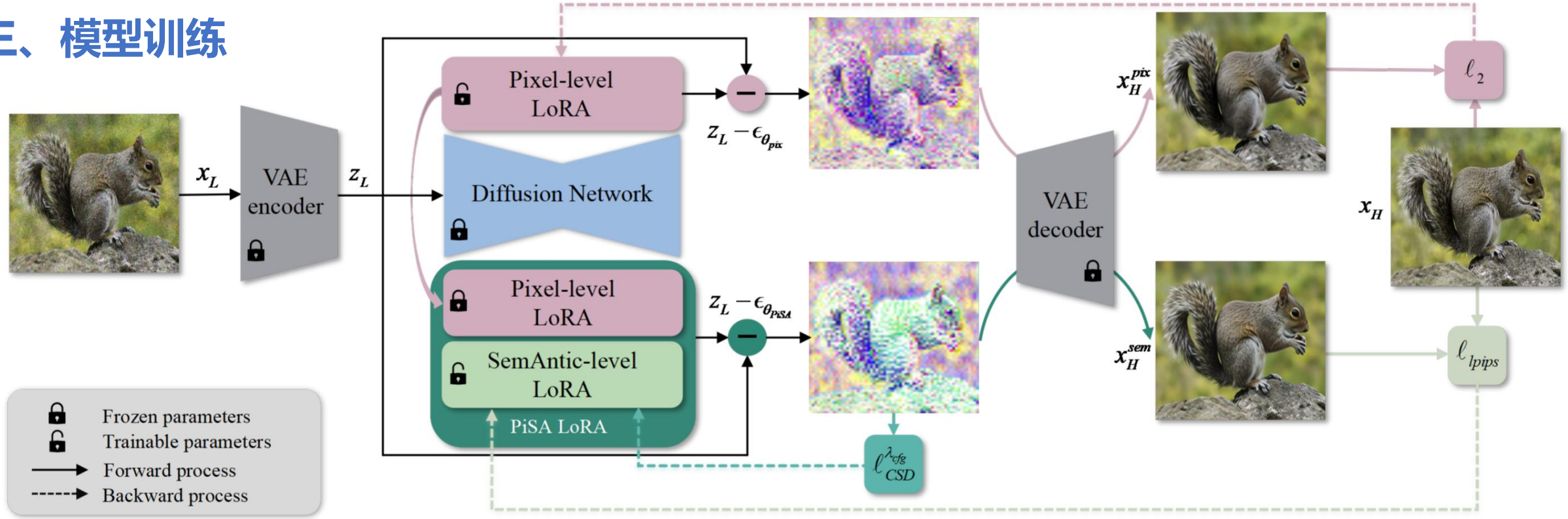
- 多步去噪框架
- 单步去噪框架

$$z_H = f(z_L, \epsilon_\theta) = \frac{z_L - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_L)}{\sqrt{\bar{\alpha}_t}}.$$

- PiSA-SR，基于SD的SR，定义为学习LQ的latent z_L 和 HQ的latent z_H 之间的残差

$$z_H = z_L - \lambda\epsilon_\theta(z_L).$$

三、模型训练

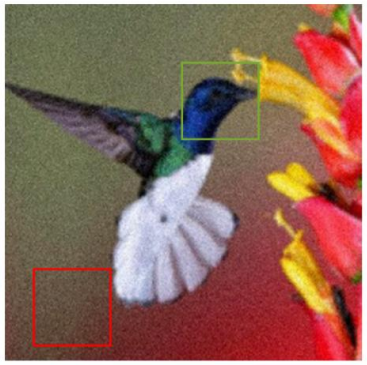


(a) Training process of the PiSA-SR

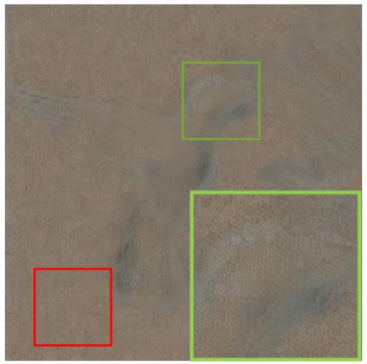
- 作者提出了一个解耦的训练方法，利用两个LoRA模块分别针对**像素级**和**语义级**增强进行SR任务
- 由于LQ图像受到噪声、模糊和下采样等退化的影响，作者首先优化**像素级**的LoRA以降低这些退化的影响，随后再优化**语义级**的LoRA

- 像素级训练阶段，完整参数集表示为 $\theta_{pix} = \{\theta_{sd}, \Delta\theta_{pix}\}$
  高质量的latent $z_H^{pix} = z_L - \epsilon_{\theta_{pix}}(z_L)$
- 语义级训练阶段，冻结像素级LoRA，$\theta_{PiSA} = \{\theta_{sd}, \Delta\theta_{pix}, \Delta\theta_{sem}\}$
  高质量的latent $z_H^{sem} = z_L - \epsilon_{\theta_{PiSA}}(z_L)$

三、模型训练

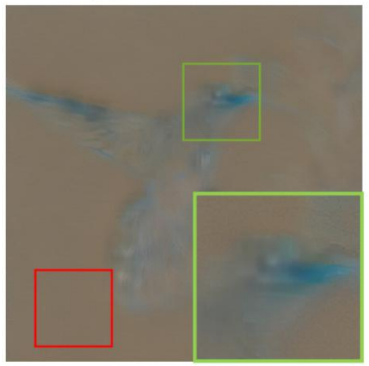


x_L



$D\left(\epsilon_{\theta_{pix}}(z_L)\right)$

$\ell_{lpips} + \ell_{CSD}^{\lambda_{cfg}=7.5}$



$D\left(\epsilon_{\theta_{PiSA}}(z_L) - \epsilon_{\theta_{pix}}(z_L)\right)$

- 对于像素级LoRA，使用 l_2 损失来训练
  - 能够有效地去除退化并增强边缘
  - 但不足以生成语义级细节，从而导致平滑的SR输出

- 对于语义级LoRA，使用 l_{lpips} 和 l_{CSD} 进行训练
  - l_{lpips}：通过预训练的VGG分类网络提取特征，将超分图像与GT图像在高层语义空间中的相似性最小化
  - 但是VGG是在有限类别上训练的，泛化能力有限

- CSD Loss（Classifier Score Distillation）
  - 借鉴于 Classifier-Free Guidance 理论， SD可以用于形成一个隐式的分类器来建模后验分布
  - CSD损失通过最大化条件概率 $p(c \mid x_t)$ 的梯度，提升语义内容的一致性。

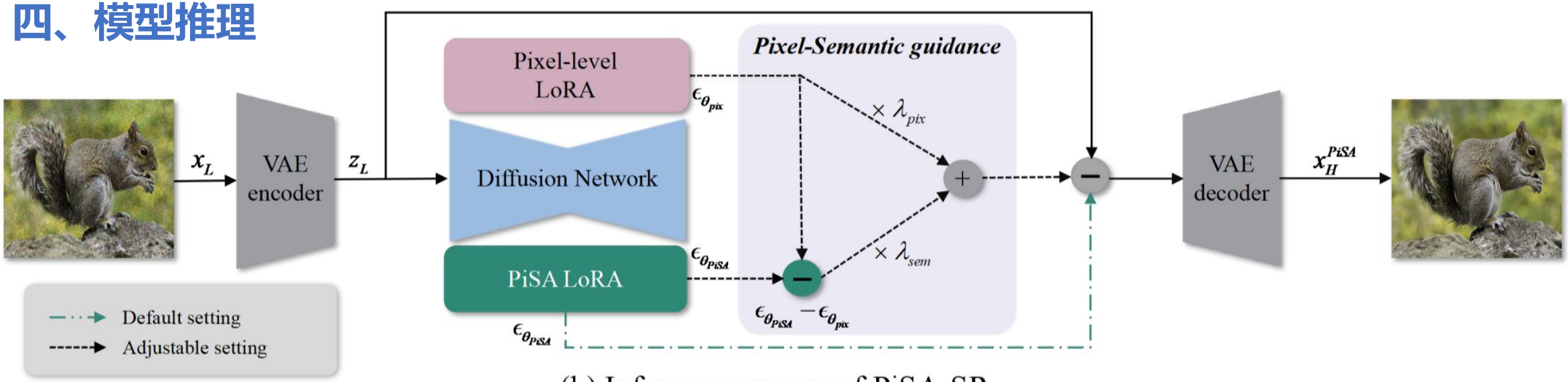
$$\begin{aligned}\nabla_{x_t} \log p_\theta(c \mid x_t) &= \nabla_{x_t} \log p_\theta(x_t \mid c) - \nabla_{x_t} \log p_\theta(x_t) \\ &= -(\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t))/\sigma_t,\end{aligned}$$

- CSD Loss（Classifier Score Distillation）
  - 为了提高超分效果，本文在潜在空间而不是噪声域中表示CSD的梯度

$$\nabla \ell_{CSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[w_t\left(f(z_t, \epsilon_{real}) - f(z_t, \epsilon_{real}^{\lambda_{cfg}})\right)\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right],$$

$$\epsilon_{real}^{\lambda_{cfg}} = \epsilon_{real}(z_t, t) + \underbrace{\lambda_{cfg}\left(\epsilon_{real}(z_t, t, c) - \epsilon_{real}(z_t, t)\right)}_{\epsilon_{real}^{cls}(z_t, t, c)}.$$

四、模型推理



(b) Inference process of PiSA-SR

- 默认设置：使用合并了像素级和语义级LoRA模块的PiSA-LoRA与预训练的SD模型共同处理输入

- 附加设置：为了实现具有多样用户偏好的灵活超分辨率，引入了一对像素级引导因子 λ_{pix} 和语义级引导因子 λ_{sem}

$$\epsilon_\theta(z_L) = \lambda_{pix}\epsilon_{\theta_{pix}}(z_L) + \lambda_{sem}(\epsilon_{\theta_{PiSA}}(z_L) - \epsilon_{\theta_{pix}}(z_L)).$$

  - 像素级LoRA输出：$\epsilon_{\theta_{pix}}(z_L)$
  - 语义级LoRA输出：$\epsilon_{\theta_{PiSA}}(z_L) - \epsilon_{\theta_{pix}}(z_L)$

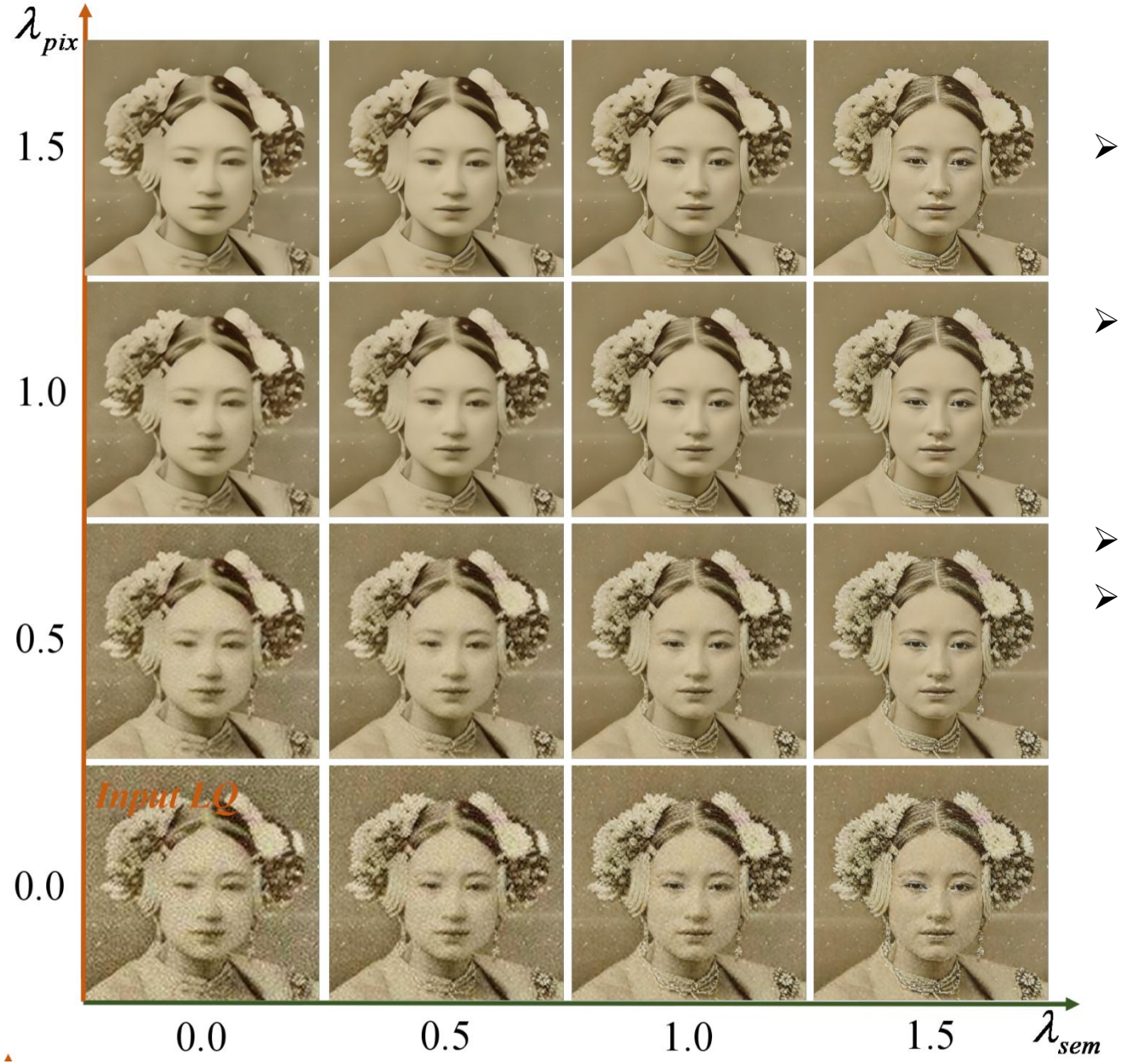
Table 1. Results of PiSA-SR with different pixel-semantic guidance scales on the RealSR test dataset.

λ_{pix}	λ_{sem}	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
0.0	1.0	25.96	0.3426	0.4129	46.45
0.2	1.0	26.48	0.3042	0.4868	54.05
0.5	1.0	26.75	0.2646	0.5705	63.82
0.8	1.0	26.18	0.2612	0.6292	68.95
1.0	1.0	25.50	0.2672	0.6702	70.15
1.2	1.0	24.76	0.2723	0.6746	70.33
1.5	1.0	23.74	0.2769	0.6305	69.23
1.0	0.0	26.92	0.3018	0.3227	49.62
1.0	0.2	26.95	0.2784	0.3591	53.64
1.0	0.5	26.77	0.2476	0.4322	58.76
1.0	0.8	26.20	0.2465	0.5806	66.33
1.0	1.0	25.50	0.2672	0.6702	70.15
1.0	1.2	24.59	0.3000	0.7015	71.60
1.0	1.5	23.08	0.3541	0.6835	71.76

不同像素语义引导尺度下PiSA-SR在RealSR测试数据集上的结果

- 增加像素级因子 λ_{pix} 会导致无参考指标CLIPIQA和MUSIQ的持续改进
  - 消除图像退化和增强边缘
- 增加语义级因子 λ_{sem} 也会导致CLIPIQA和MUSIQ的持续改善

- 但是过高的 λ_{sem} 会导致 PSNR 和 LPIPS 指标衰退
- 这是因为过多的语义细节可能会引起图像内容的变化，降低像素级的保真度，增加过多冗余细节



- 增加像素级因子 λ_{pix} 会导致无参考指标CLIPIQA和MUSIQ的持续改进
  - 消除图像退化和增强边缘
- 增加语义级因子 λ_{sem} 也会导致CLIPIQA和MUSIQ的持续改善

- 但是过高的 λ_{sem} 会导致 PSNR 和 LPIPS 指标衰退
- 这是因为过多的语义细节可能会引起图像内容的变化，降低像素级的保真度，增加过多冗余细节

Table 2. Quantitative comparison among the state-of-the-art DM-based SR methods on synthetic and real-world test datasets. 'S' denotes the number of diffusion steps. The best and the second-best results are highlighted in **red** and **blue**, respectively.

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	CLIPIQA↑	MUSIQ↑	MANIQA↑
DIV2K	ResShift-S15	**24.69**	**0.6175**	0.3374	0.2215	36.01	6.82	0.6089	60.92	0.5450
	StableSR-S200	23.31	0.5728	0.3129	0.2138	**24.67**	4.76	0.6682	65.63	0.6188
	DiffBIR-S50	23.67	0.5653	0.3541	0.2129	30.93	4.71	0.6652	65.66	0.6204
	PASD-S20	23.14	0.5489	0.3607	0.2219	29.32	**4.40**	0.6711	**68.83**	**0.6484**
	SeeSR-S50	23.71	0.6045	0.3207	**0.1967**	25.83	4.82	**0.6857**	68.49	0.6239
	SinSR-S1	**24.43**	0.6012	0.3262	0.2066	35.45	6.02	0.6499	62.80	0.5395
	OSEDiff-S1	23.72	**0.6108**	**0.2941**	0.1976	26.32	4.71	0.6683	67.97	0.6148
	PiSA-SR-S1	23.87	0.6058	**0.2823**	**0.1934**	**25.07**	**4.55**	**0.6927**	**69.68**	**0.6400**
RealSR	ResShift-S15	**26.31**	**0.7411**	0.3489	0.2498	142.81	7.27	0.5450	58.10	0.5305
	StableSR-S200	24.69	0.7052	0.3091	0.2167	127.20	5.76	0.6195	65.42	0.6211
	DiffBIR-S50	24.88	0.6673	0.3567	0.2290	124.56	5.63	0.6412	64.66	0.6231
	PASD-S20	25.22	0.6809	0.3392	0.2259	**123.08**	**5.18**	0.6502	68.74	**0.6461**
	SeeSR-S50	25.33	0.7273	0.2985	0.2213	125.66	**5.38**	0.6594	**69.37**	0.6439
	SinSR-S1	**26.30**	0.7354	0.3212	0.2346	137.05	6.31	0.6204	60.41	0.5389
	OSEDiff-S1	25.15	0.7341	**0.2921**	**0.2128**	**123.50**	5.65	**0.6693**	69.09	0.6339
	PiSA-SR-S1	25.50	**0.7417**	**0.2672**	**0.2044**	124.09	5.50	**0.6702**	**70.15**	**0.6560**
DrealSR	ResShift-S15	**28.45**	0.7632	0.4073	0.2700	175.92	8.28	0.5259	49.86	0.4573
	StableSR-S200	28.04	0.7460	0.3354	0.2287	147.03	6.51	0.6171	58.50	0.5602
	DiffBIR-S50	26.84	0.6660	0.4446	0.2706	167.38	**6.02**	0.6292	60.68	0.5902
	PASD-S20	27.48	0.7051	0.3854	0.2535	157.36	**5.57**	0.6714	64.55	**0.6130**
	SeeSR-S50	28.26	0.7698	0.3197	0.2306	149.86	6.52	0.6672	64.84	0.6026
	SinSR-S1	**28.41**	0.7495	0.3741	0.2488	177.05	7.02	0.6367	55.34	0.4898
	OSEDiff-S1	27.92	**0.7835**	**0.2968**	**0.2165**	**135.29**	6.49	**0.6963**	**64.65**	0.5899
	PiSA-SR-S1	28.31	**0.7804**	**0.2960**	**0.2169**	**130.61**	6.20	**0.6970**	**66.11**	**0.6156**

Table 3. The inference time and the number of parameters of DM-based SR methods.

	StableSR	ResShift	DiffBIR	PASD	SeeSR	SinSR	OSEDiff	PiSA-SR-def.	PiSA-SR-adj.
Inference Steps	200	15	50	20	50	1	1	1	2
Inference time(s)/Image	10.03	0.76	2.72	2.80	4.30	0.13	0.12	0.09	0.13
#Params(B)	1.56	0.18	1.68	2.31	2.51	0.18	1.77	1.30	1.30

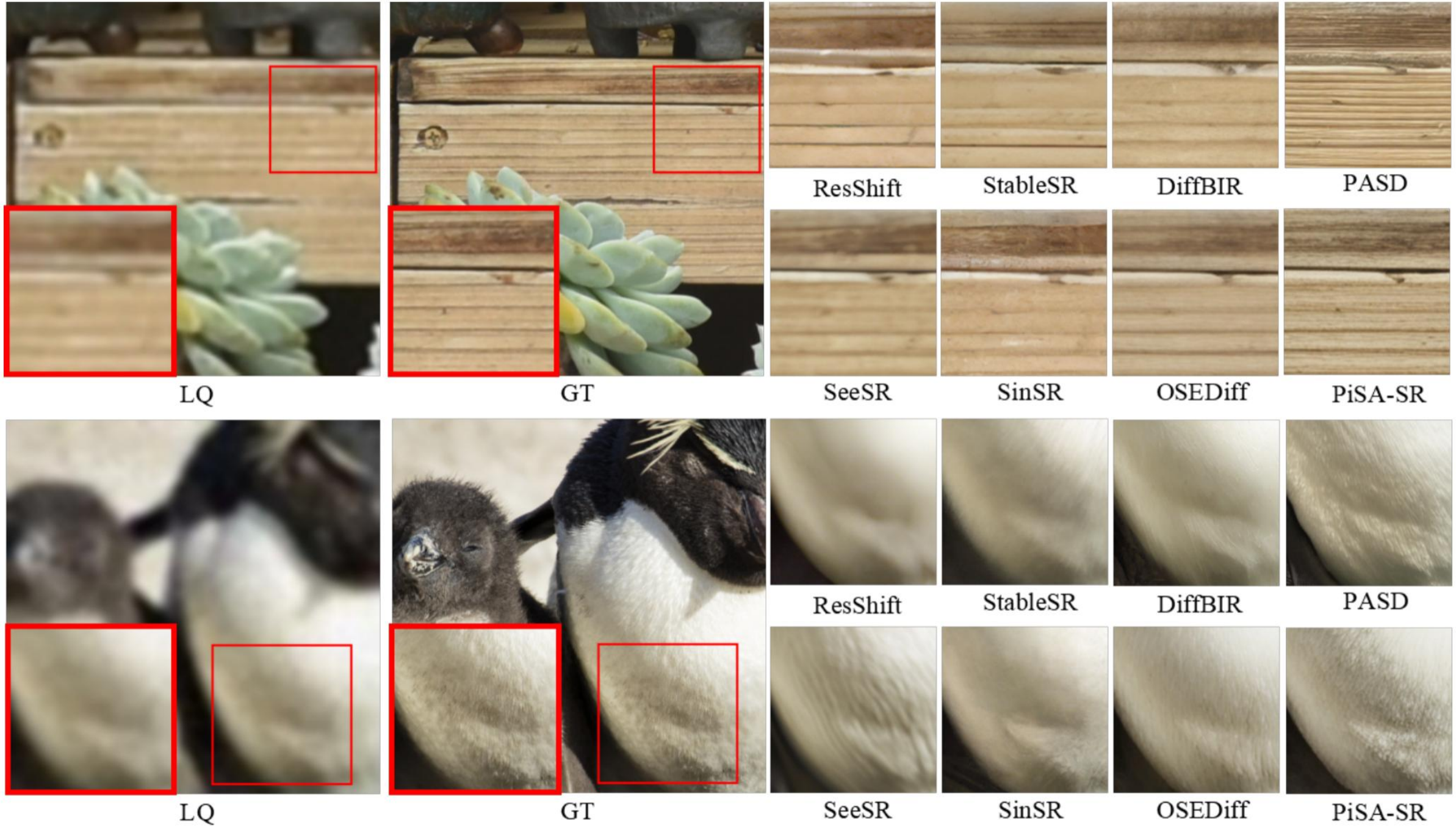
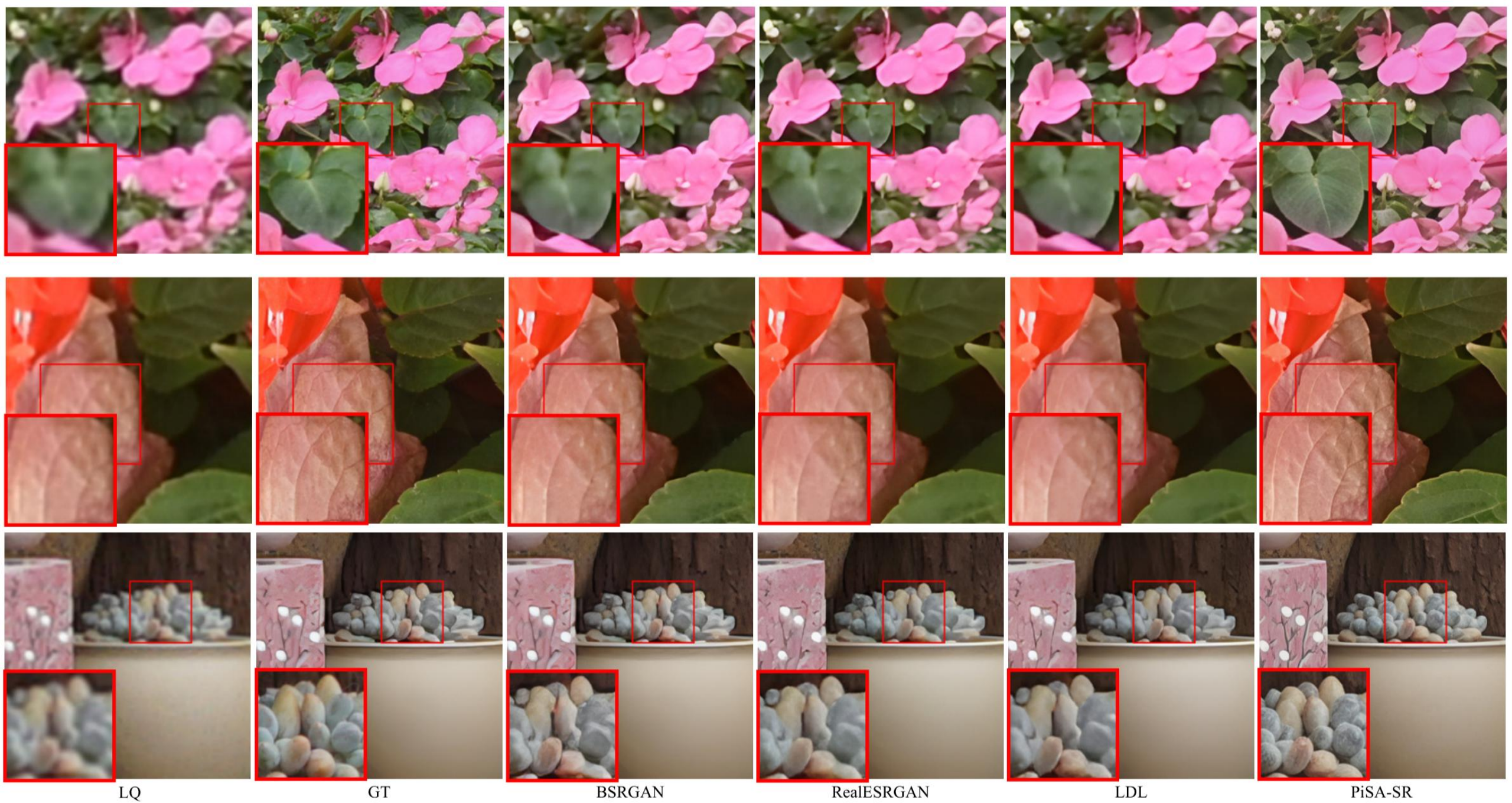


Figure 5. Visual comparisons of different DM-based SR methods. Please zoom in for a better view.



PiSA-SR 与不同 GAN-SR 方法的视觉比较

Table 6. Ablation studies on pixel-level and semantic-level LoRA ranks on RealSR dataset.

Methods	Pixel-level LoRA rank	Semantic-level LoRA rank	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
PiSA-SR	4	4	25.50	0.7417	0.2672	0.6702	0.6560	70.15
R1	4	8	25.40	0.7401	0.2719	0.6774	0.6584	69.93
R2	4	16	25.36	0.7398	0.2726	0.6777	0.6603	70.15
R3	4	32	25.40	0.7394	0.2713	0.6784	0.6634	70.13
R4	8	4	25.39	0.7422	0.2628	0.6663	0.6600	69.86
R5	16	4	25.54	0.7511	0.2624	0.6603	0.6636	69.77
R6	32	4	26.01	0.7565	0.2564	0.6318	0.6409	68.30

- 在默认设置中，像素级和语义级LoRA的秩都设置为4
- 通过将一个LoRA的排名固定为4，并改变另一个来进行实验，以观察RealSR测试数据集上的性能变化

- 增加语义级别的LoRA rank可以增强语义细节
  - 反映在无参考的CLIPIQA、MANIQA、MUSIQ指标中
  - 然而，会降低图像保真度，导致较低的基于参考的分数

- 增加像素级别的LoRA rank，使图像过于平滑而保真度较高
  - 反映在PSNR、SSIM、LPIPS指标中
  - 然而，超分结果缺乏细节，导致较低的无参考评估指标

Thank You