RealSR

背景概述

- Real SR相较于普通SR，旨在处理具有真实世界具有复杂退化的输入
- 因为具有复杂退化，一般的网络难以有效拟合，常常导致细节模糊过度平滑
- 许多研究者使用生成方法来使图片具有较好的视觉效果
- 自从StableSR引入SD进入real-SR任务以来，基于SD的超分越来越流行，强大的细节生成能力，避免以前SR方法的过度平滑
- SUPIR DiffBIR SeeSR……等多步方法能够生成良好的细节，但是推理成本较大
- 目前许多方法已经转向单步SDSR，主要是基于LoRA微调SD

指标问题

输入

一般方法（CNN、Transformer）

23.4099/0.5622

SD扩散超分方法

20.8640/0.4606

GT

可以看到，SD方法能够生成更多精致的细节
但是以往SR的主要指标：PSNR/SSIM，却很低
这两个指标主要是考虑像素保真度，也就是直接与GT计算
单纯比较PSNR/SSIM并不公平，因为现实中是没有GT去比较的。因此引入了许多主观指标
CLIPIQA、NIQE……

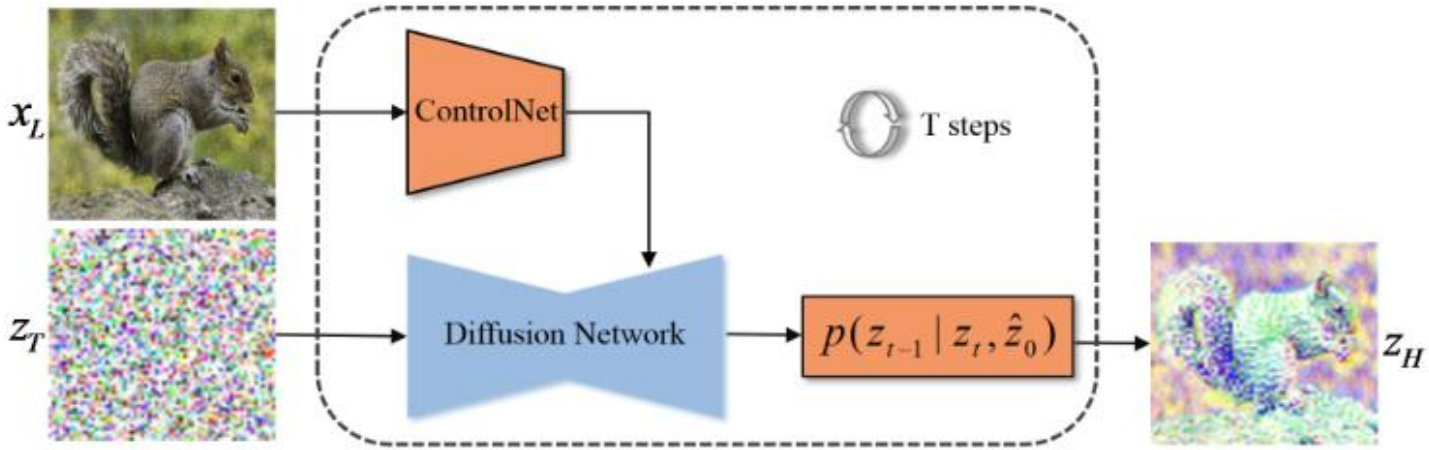
但是，即使不考虑GT的存在，结果的像素保真度依然也是重要的
目前的SD方法在像素保真度方面表现都劣于一般方法

一步式高效真实扩散超分

One-Step Effective Diffusion Network for Real-World Image Super-Resolution NeurIPS 24

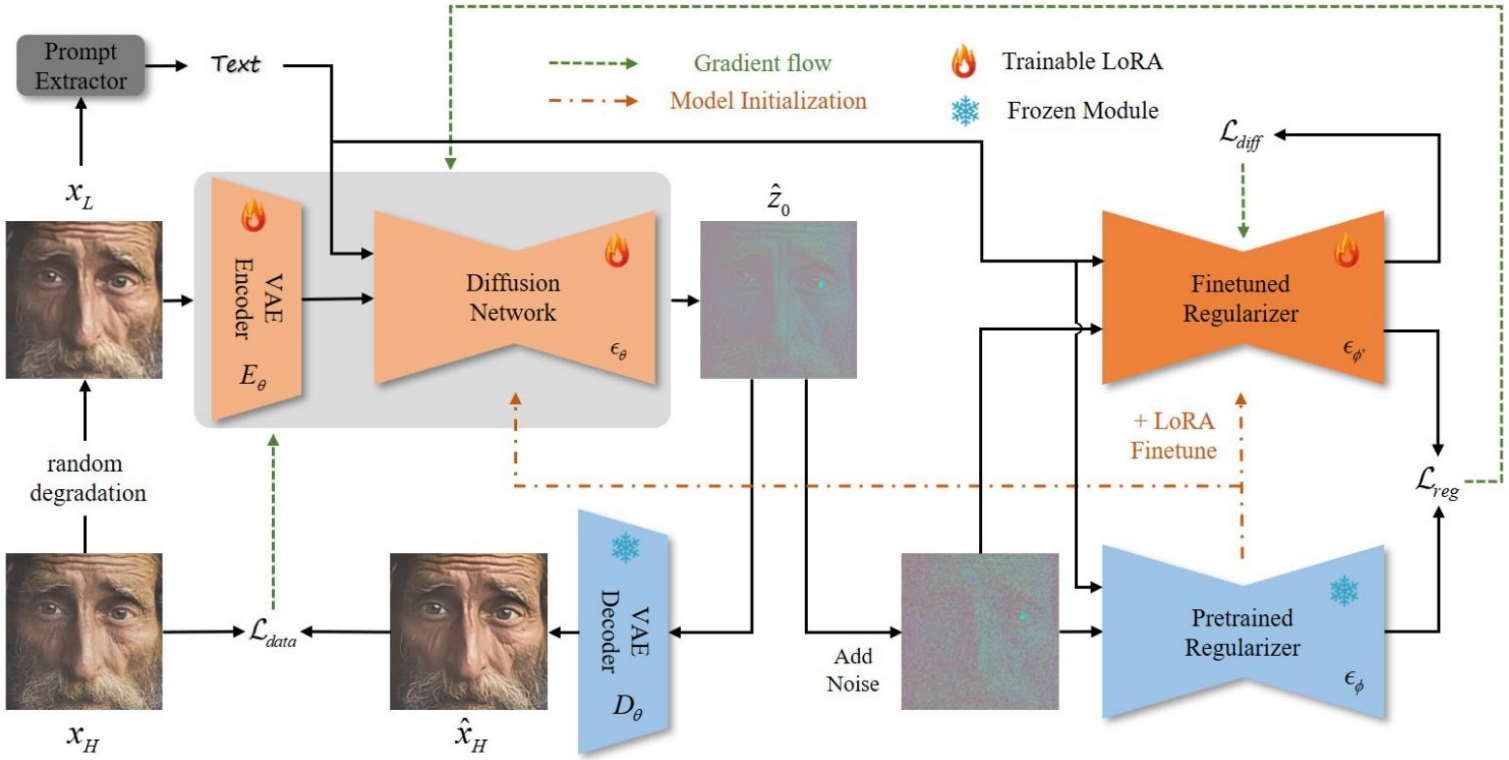
动机

- 现有扩散超分方法(StableSR， SeeSR， DiffBIR……)都都是多步的范式，不论是推理时间还是训练成本都不便宜
  - 训练一个ControlNet，本质上还是让扩散模型在生成图片，而不是一个SR模型
  - 多步的扩散



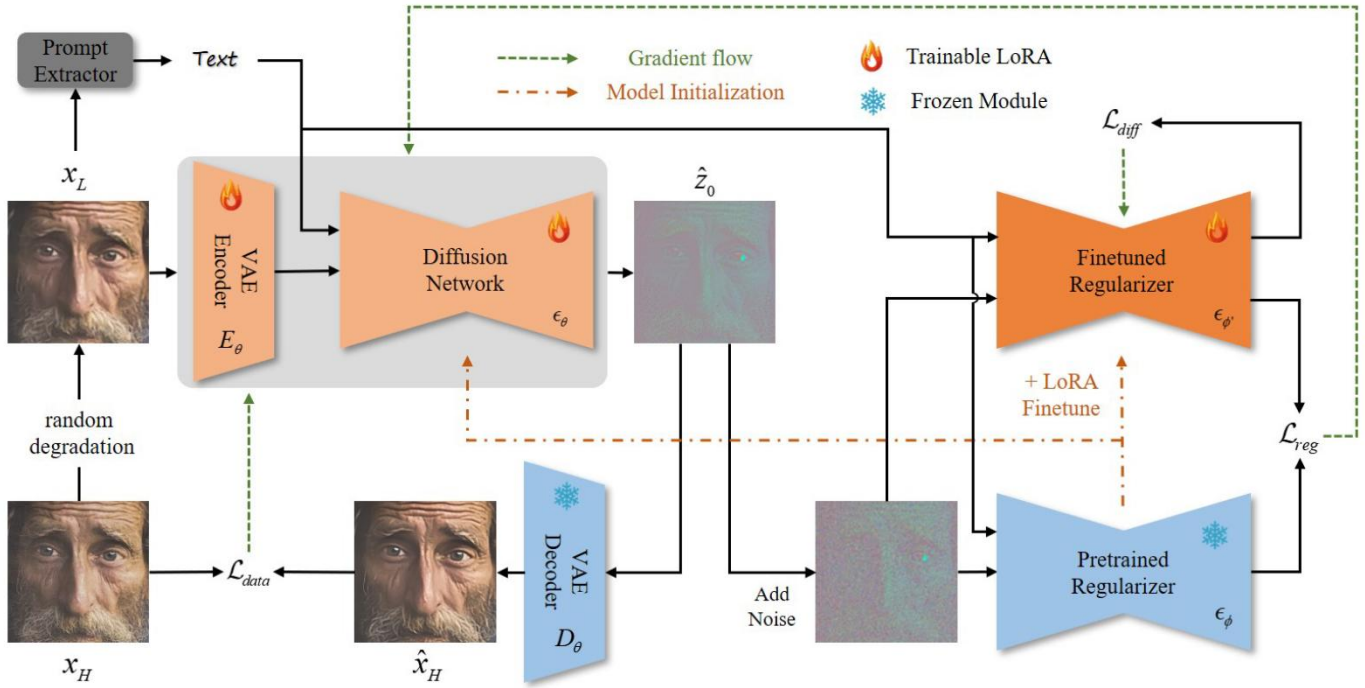
(a) Multi-step DM-based SR methods

单步超分



- 直接LoRA微调SD的VAE Encoder和Unet→相较于ControlNet来说训练参数小很多，并且可以单步
- L_data = L2 + LPIPS
- 但是微调SD可能会导致原有的丰富扩散先验损失
- 加入VSD Loss以蒸馏扩散先验
- DAPE for Prompt Extractor

VSD Loss



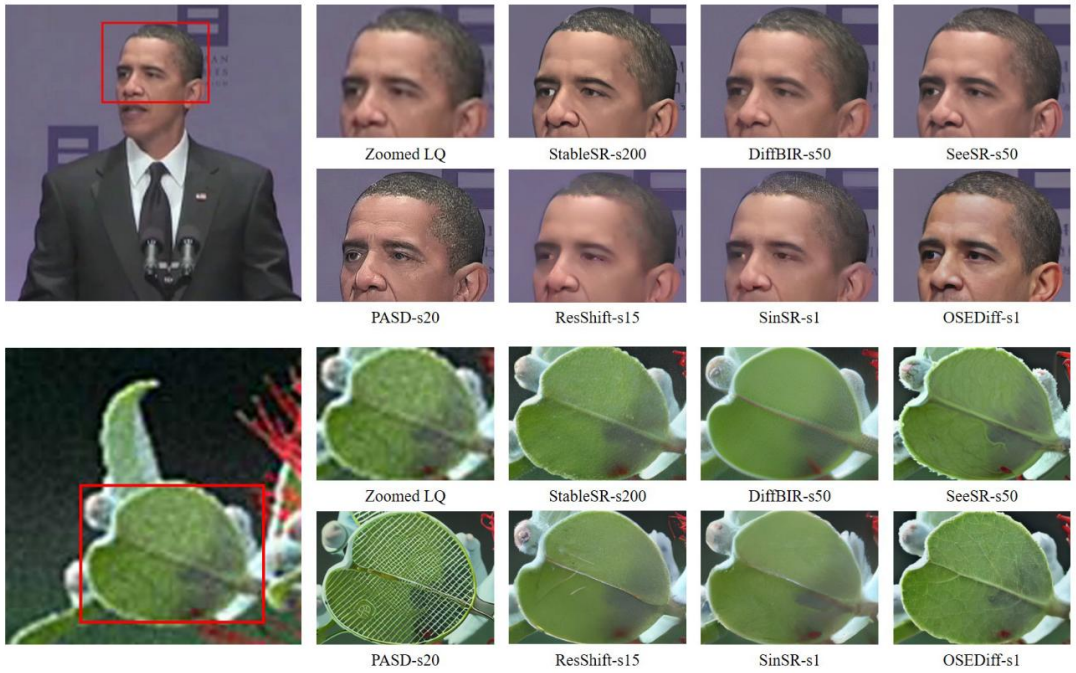
- 从原始SD初始化出一个学生模型
- 微调该学生模型，就发挥原先的扩散功能（逐步去噪）
- 将超分结果加噪，学生模型进行去噪，通过L_Diff来使得学生模型拟合超分结果的分布
- L_reg：使得学生模型的输出分布拟合 教师模型对加噪的超分结果的输出分布（SD先验）
- 通过这样相隔的方式来将SD先验蒸馏到超分模型上

结果

Datasets	Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑
DIV2K-Val	StableSR-s200	23.26	0.5726	**0.3113**	0.2048	**24.44**	4.7581	65.92	0.6192	0.6771
	DiffBIR-s50	23.64	0.5647	0.3524	0.2128	30.72	**4.7042**	65.81	0.6210	0.6704
	SeeSR-s50	23.68	0.6043	0.3194	**0.1968**	**25.90**	4.8102	**68.67**	**0.6240**	**0.6936**
	PASD-s20	23.14	0.5505	0.3571	0.2207	29.20	**4.3617**	**68.95**	**0.6483**	**0.6788**
	ResShift-s15	**24.65**	**0.6181**	0.3349	0.2213	36.11	6.8212	61.09	0.5454	0.6071
	SinSR-s1	**24.41**	0.6018	0.3240	0.2066	35.57	6.0159	62.82	0.5386	0.6471
	OSEDiff-s1	23.72	**0.6108**	**0.2941**	**0.1976**	26.32	4.7097	67.97	0.6148	0.6683
DrealSR	StableSR-s200	28.03	0.7536	0.3284	**0.2269**	148.98	6.5239	58.51	0.5601	0.6356
	DiffBIR-s50	26.71	0.6571	0.4557	0.2748	166.79	**6.3124**	61.07	0.5930	0.6395
	SeeSR-s50	28.17	**0.7691**	**0.3189**	0.2315	**147.39**	6.3967	**64.93**	**0.6042**	0.6804
	PASD-s20	27.36	0.7073	0.3760	0.2531	156.13	**5.5474**	**64.87**	**0.6169**	**0.6808**
	ResShift-s15	**28.46**	0.7673	0.4006	0.2656	172.26	8.1249	50.60	0.4586	0.5342
	SinSR-s1	**28.36**	0.7515	0.3665	0.2485	170.57	6.9907	55.33	0.4884	0.6383
	OSEDiff-s1	27.92	**0.7835**	**0.2968**	**0.2165**	**135.30**	6.4902	64.65	0.5899	**0.6963**
RealSR	StableSR-s200	24.70	0.7085	0.3018	0.2288	128.51	5.9122	65.78	0.6221	0.6178
	DiffBIR-s50	24.75	0.6567	0.3636	0.2312	128.99	5.5346	64.98	0.6246	0.6463
	SeeSR-s50	25.18	0.7216	**0.3009**	**0.2223**	125.55	**5.4081**	**69.77**	**0.6442**	0.6612
	PASD-s20	25.21	0.6798	0.3380	0.2260	**124.29**	**5.4137**	68.75	**0.6487**	**0.6620**
	ResShift-s15	**26.31**	**0.7421**	0.3460	0.2498	141.71	7.2635	58.43	0.5285	0.5444
	SinSR-s1	**26.28**	**0.7347**	0.3188	0.2353	135.93	6.2872	60.80	0.5385	0.6122
	OSEDiff-s1	25.15	0.7341	**0.2921**	**0.2128**	**123.49**	5.6476	**69.09**	0.6326	**0.6693**

Table 2: Complexity comparison among different methods. All methods are tested with an input image of size 512×512, and the inference time is measured on an A100 GPU.

	StableSR	DiffBIR	SeeSR	PASD	ResShift	SinSR	OSEDiff
Inference Step	200	50	50	20	15	1	1
Inference Time (s)	11.50	2.72	4.30	2.80	0.71	0.13	0.11
MACs (G)	79940	24234	65857	29125	5491	2649	2265
# Total Param (M)	1410	1717	2524	1900	119	119	1775
# Trainable Param (M)	150.0	380.0	749.9	625.0	118.6	118.6	8.5



消融实验

Table 3: Comparison of different losses on the RealSR benchmark.

	PSNR↑	LPIPS↓	MUSIQ↑	CLIPIQA↑
w/o VSD loss	25.42	0.2934	65.23	0.5876
GAN loss	25.00	0.2760	67.29	0.6254
VSD loss in image domain	25.05	0.2759	67.90	0.6256
OSEDiff	25.15	0.2921	69.09	0.6693

Table 4: Comparison of different text prompt extractors on the DrealSR benchmark.

Prompt Extraction Methods	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FID↓	NIQE↓	MUSIQ↑	MANIQA↑	CLIPIQA↑	Prompt Extraction Time (s)
Null	28.51	0.7910	0.2896	0.2080	126.59	6.6436	62.13	0.5782	0.6599	0
DAPE	27.92	0.7835	0.2968	0.2165	135.30	6.4902	64.65	0.5899	0.6963	0.02
LLaVA-v1.5	27.72	0.7735	0.3109	0.2249	149.45	6.4119	65.70	0.6038	0.7033	3.53

Table 5: Comparison of LoRA in VAE encoder with different ranks.

Rank	PSNR↑	DISTS↓	MUSIQ↑	NIQE↓
2	-	-	-	-
4	25.15	0.2128	69.09	5.6479
8	24.86	0.2134	68.37	5.8184

Table 6: Comparison of LoRA in UNet with different ranks.

Rank	PSNR↑	DISTS↓	MUSIQ↑	NIQE↓
2	25.28	0.2154	68.89	5.7171
4	25.15	0.2128	69.09	5.6479
8	24.87	0.2115	68.39	5.8184

Table 7: Ablation studies on finetuning the VAE encoder and decoder on the RealSR benchmark.

	Train VAE Encoder	Train VAE Decoder	PSNR↑	DISTS↓	LPIPS↓	CLIPIQA↑	MUSIQ↑	NIQE↓
(1)	×	×	25.27	0.1966	0.2656	0.5303	58.99	6.5496
(2)	×	✓	25.30	0.2049	0.2829	0.5604	65.83	6.6291
(3)	✓	✓	25.59	0.2141	0.3017	0.5778	65.92	6.9845
OSEDiff	✓	×	25.15	0.2128	0.2921	0.6693	69.09	5.6479

- LOSS消融
  - VSD Loss确实起到了蒸馏SD先验的作用
  - GAN某种程度上也有用
- Prompt消融
- LoRA秩消融
- VAE LoRA消融

结论

- 第一个单步SDSR的方法
- 主要贡献是VSD Loss：蒸馏SD先验
  - 相较于某些不用这个Loss的单步SDSR方法，未必就是真的关键
  - 根据使用场景
  - 训练成本较大
- 性能不弱于多步的方法

双lora方法实现像素级和语义级可调SR

Pixel-level and Semantic-level Adjustable Super-resolution: A Dual-LoRA Approach CVPR25

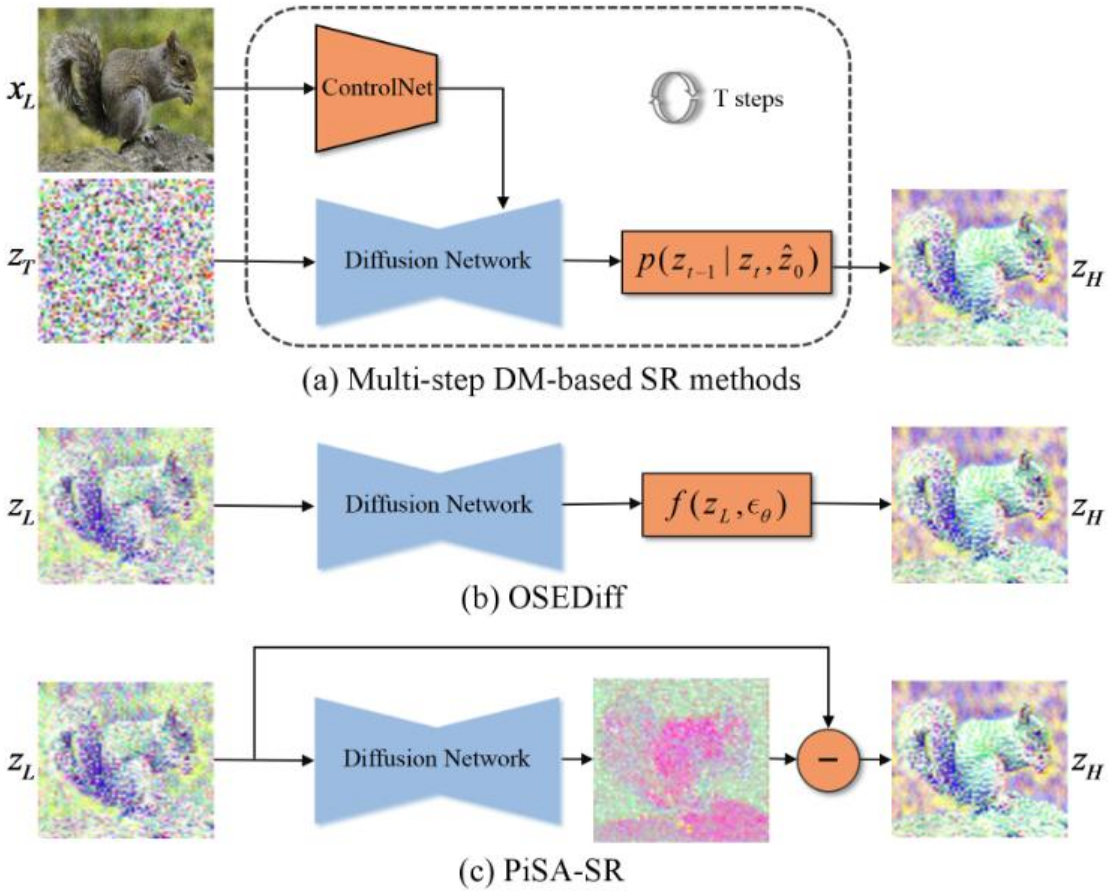
Motivation

- 现有SD方法难以良好平衡像素保真度（PSNR/SSIM）和语义级别的感知程度（其他主观指标）

RealSR	StableSR-s200	24.70	0.7085
	DiffBIR-s50	24.75	0.6567
	SeeSR-s50	25.18	0.7216
	PASD-s20	25.21	0.6798
	ResShift-s15	**26.31**	**0.7421**
	SinSR-s1	**26.28**	**0.7347**
	OSEDiff-s1	25.15	0.7341

- 用户偏好难以便利调节
  - 有些人希望能保留更多原始内容（PSNR/SSIM），有些人希望人眼感官更好（主观指标）

基础结构

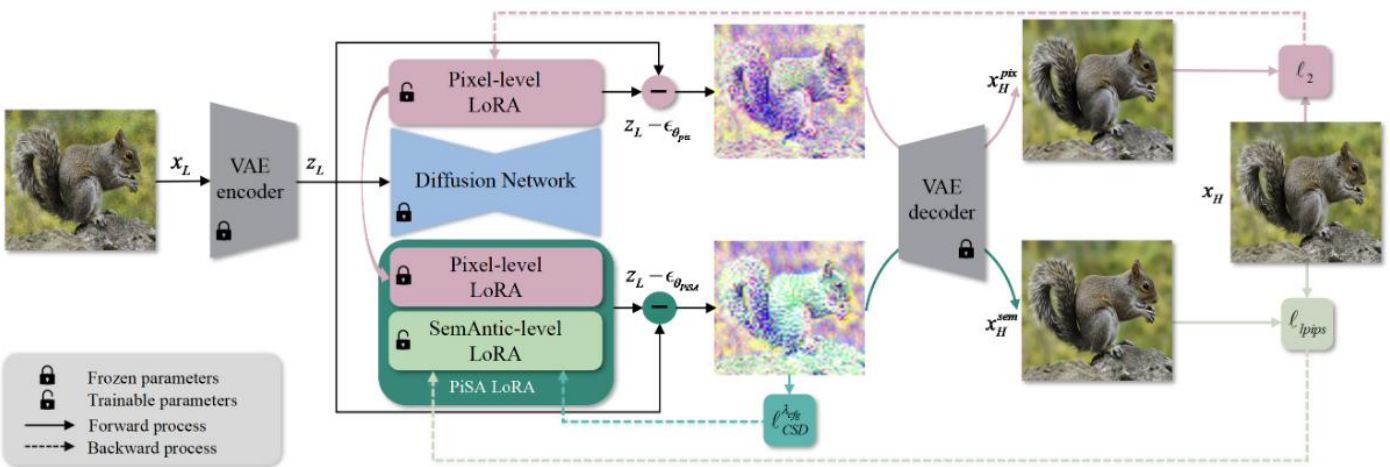


单步SD方法OSEDiff实现直接从低质量Zl通过Unet直接增强到Zh

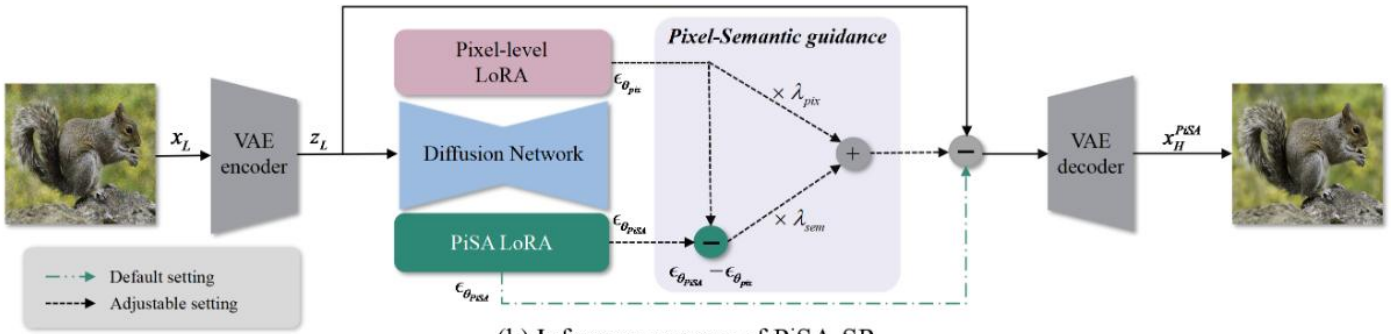
$$z_H = f(z_L, \epsilon_\theta) = \frac{z_L - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_L)}{\sqrt{\bar{\alpha}_t}}$$

单步SD结构使得应用残差成为可能：

$$z_H = z_L - \lambda\epsilon_\theta(z_L) \text{ 。}$$



(a) Training process of the PiSA-SR



(b) Inference process of PiSA-SR

- 从原始SD初始化出一个学生模型
- 两个LoRA，Pix LoRA和Sem LoRA
- 先微调Pix LoRA：L2 Loss
- 然后将Pix LoRA融进原Unet，训Sem LoRA：LPIPS + CSD_Loss
- 推理：默认模式下，只需要1步（Pix和Sem的权重都是1，也就是直接将微调好的Sem LoRA融进去）；用户偏好模式下，还需要再读取原先Pix LoRA的权重再走一步，然后根据权重来

CSD Loss

- 不同于VSD Loss，CSD Loss目的是使得超分结果和prompt的语义空间相近，并且不需要train一个学生模型
- 将VSD Loss拓展到有CFG的形式

$$\nabla \ell_{VSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[w_t\left(f(z_t,\epsilon_{fake}) - f(z_t,\epsilon_{real}^{\lambda_{cfg}})\right)\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right], \tag{7}$$

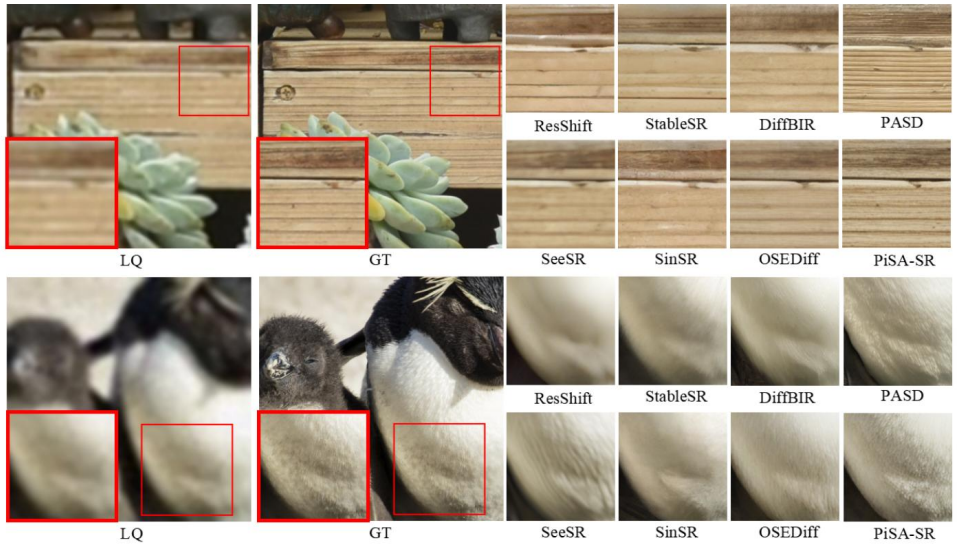
$$\epsilon_{real}^{\lambda_{cfg}} = \epsilon_{real}(z_t,t) + \underbrace{\lambda_{cfg}\left(\epsilon_{real}(z_t,t,c) - \epsilon_{real}(z_t,t)\right)}_{\epsilon_{real}^{cls}(z_t,t,c)}.$$

$$\nabla \ell_{CSD}^{\lambda_{cfg}} = \mathbb{E}_{t,\epsilon,z_t,c}\left[w_t\left(f(z_t,\epsilon_{real}) - f(z_t,\epsilon_{real}^{\lambda_{cfg}})\right)\frac{\partial z_H^{sem}}{\partial \theta_{PiSA}}\right], \tag{5}$$

结果

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**

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



消融

Table 5. Ablation studies on the dual-LoRA training approach on RealSR dataset.

Methods	Pixel-level LoRA	Semantic-level LoRA	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	MUSIQ↑
V1	✓	×	27.28	0.7975	0.3090	0.3130	0.3995	49.02
V2	×	✓	24.13	0.7290	0.2803	0.6711	0.6614	70.69
PiSA-SR	✓	✓	25.50	0.7417	0.2672	0.6702	0.6560	70.15

消融了Pix 和Sem LoRA

Table 8. Comparisons of CFG and the proposed semantic-level guidance on RealSR dataset.

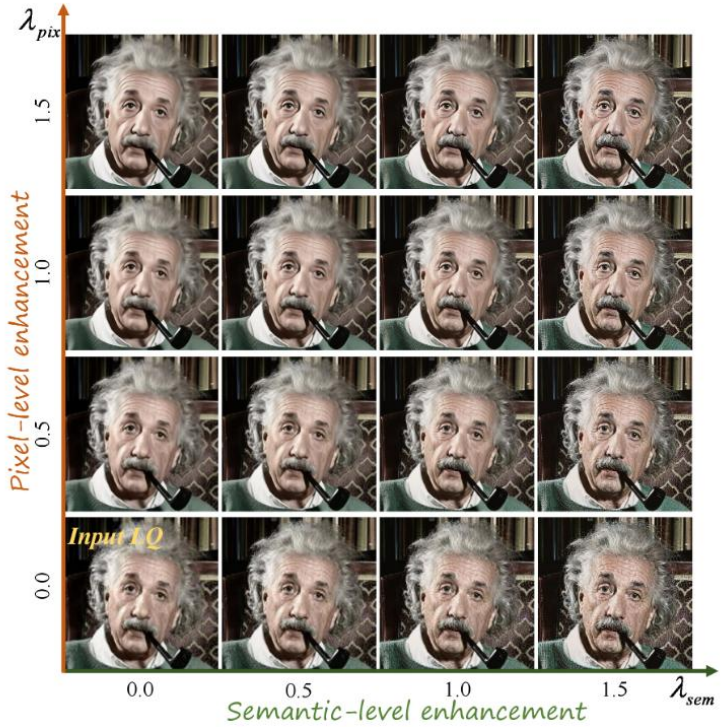
λ_{cfg}	λ_{sem}	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑	Inference time(s)/Image
1.0	1.0	25.50	0.2672	0.6702	70.15	0.09
1.2	×	25.38	0.2684	0.6708	70.23	0.15
1.5	×	25.30	0.2698	0.6708	70.29	0.15
3.0	×	24.61	0.2834	0.6540	70.14	0.15
×	1.2	24.59	0.3000	0.7015	71.60	0.13
×	1.5	23.08	0.3541	0.6835	71.76	0.13

消融了CFG和Sem LoRA

总结

- 主要贡献在于双LoRA的设计，实现像素-语义便利调节
- 但是这里的双LoRA权重调节缺少实际的物理意义
- CSD Loss讲的不是很清楚，其实际贡献在实验没有充分体现（对比其他Loss），代码也没给
- 有一点没有深入探讨：残差结构的好处

Figure 10. Visual comparisons of SR results from OSEDiff and PiSA-SR across 1 to 2000 training iterations.



关于SD-SR

- 可能是目前Low-Level最火热的任务
- 训练成本其实还好
- 还有许多可以做的事:
  - Pixel-Semantic平衡
  - 推理加速
  - 文字问题

- 谢谢!