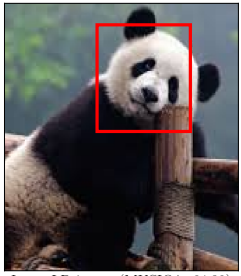


Input LR image (MUSIQ↑: 64.39)

Input (43.90)

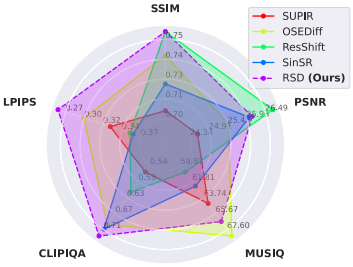
SUPIR-50 (71.14)

ResShift-15 (64.68)

OSEDiff-1 (69.29)

SinSR-1 (69.22)

Ours-1 (**73.19**)



One-Step Residual Shifting Diffusion for Image Super-Resolution via Distillation

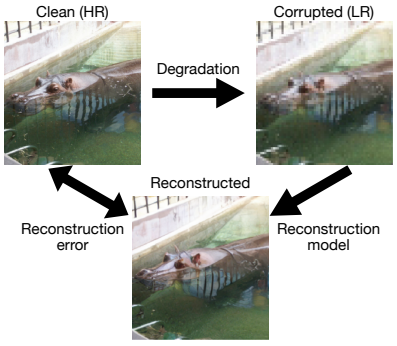
Daniil Selikhanovych* **David Li*** **Aleksei Leonov*** **Nikita Gushchin***
Sergey Kushneryuk **Alexander Filippov** **Evgeny Burnaev** **Iaroslav Koshelev**
Alexander Korotin

**Equal contribution*

2026

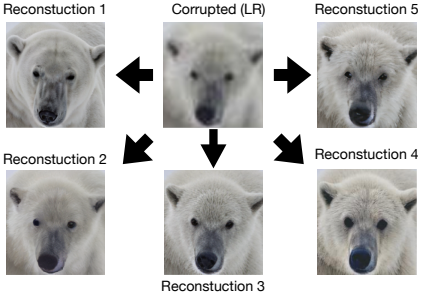
The Problem: Real-World Image Super-Resolution[1]

The problem: to reconstruct a high-resolution (HR) image from a degraded low-resolution (LR) observation.



Real-world setting:

1. Complex and unknown degradations.
2. Real-ISR is an ill-posed problem.



Which one to choose?

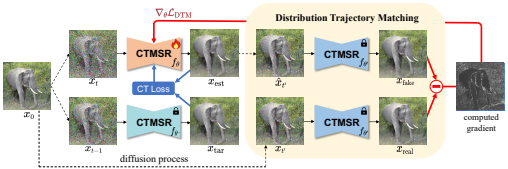
[1]Daniel Glasner, Shai Bagon, and Michal Irani (2009). **"Super-resolution from a single image".** In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 349–356.

Diffusion Models for Real-ISR

Diffusion models have been shown to outperform GANs in perceptual quality for the Real-ISR problem[2]. They fall into two categories depending on the use of T2I priors.

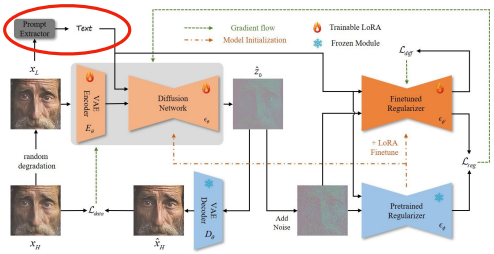
No T2I prior (CTMSR).

The diffusion process starts from the LR image instead of Gaussian noise. The diffusion model is trained without weight initialization.



Use the pre-trained T2I prior (OSEDiff).

LR conditioning is based on controllers; prompts are extracted from the LR image.



[2] Jianyi Wang et al. (2024). **"Exploiting Diffusion Prior for Real-World Image Super-Resolution".** In: *International Journal of Computer Vision*.

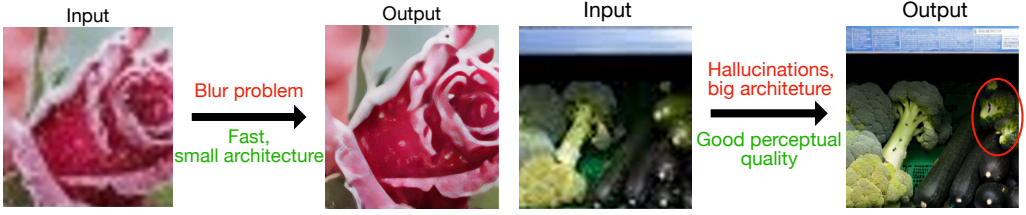
Limitations of Diffusion Models and Contribution

No T2I prior.

+ SinSR and CTMSR have light architectures and fast inference.
- But they are known to produce **blurred results** compared to T2I-based models.

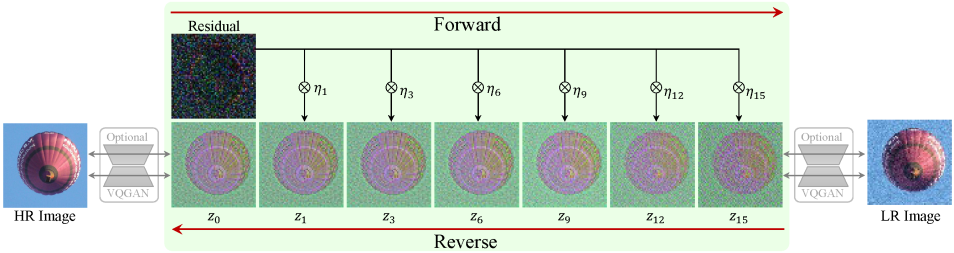
Use the pre-trained T2I prior.

+ T2I-based models have high perceptual quality.
- But they can **hallucinate** and require high computational costs.



We propose to unite the best of two worlds:
Good perceptual-distortion trade-off using low computational resources.

ResShift Model[3]



Core idea	Why it is a good teacher
Gradually shifts the HR latent toward the LR image and learns the reverse restoration path.	Diffusion in latent space leads to a high perceptual-distortion trade-off.

We propose to distill ResShift teacher with 15 denoising steps into one-step student.

[3]Zongsheng Yue et al. (2023). **"ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting".** In: *Advances in Neural Information Processing Systems*.

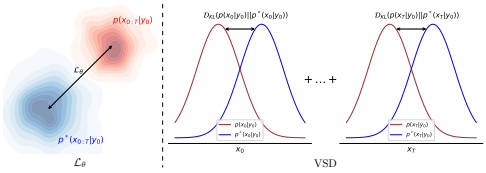
Motivation and Idea of the Method

Objective for student G_θ:

- Parameterization: predict HR from LR y_0 as $\widehat{x}_0 = G_\theta(x_T, y_0, \epsilon)$ with $\epsilon \sim \mathcal{N}(0, I)$.
- Assumption: if a ResShift model trained on generated pairs f_{G_θ} matches the teacher f^*, then generated and real LR-HR pairs should match.

$$f_{G_\theta} \approx f^* \quad \Rightarrow \quad p_\theta(y_0, x_0) \approx p_{data}(y_0, x_0)$$

This assumption holds if the teacher is ideal.



Representation with KL divergence: in contrast to variational score distillation (VSD), we align full teacher trajectories:

$$L_\theta = \mathbb{E}_{p(y_0)} \mathcal{D}_{\mathrm{KL}}(p(x_{0:T}|y_0) \;||\; p^*(x_{0:T}|y_0))$$

RSD loss:

$$L_\theta = \sum_{t=1}^{T} w_t \, \mathbb{E}_{p_\theta(\hat{x}_0, y_0, x_t)} \, \| f_{G_\theta}(x_t, y_0, t) - f^*(x_t, y_0, t) \|_2^2$$

Backpropagation Problem and its Solution

Why the direct objective is impractical:

$$\frac{dL_\theta}{d\theta} = \underbrace{\frac{\partial L}{\partial \theta}}_{\text{direct}} + \frac{\partial L}{\partial f_{G_\theta}} \underbrace{\frac{\partial f_{G_\theta}}{\partial \theta}}_{\text{implicit}}$$

$$f_{G_\theta} = \arg\min_f \sum_{t=1}^{T} w_t \, \mathbb{E}_{p_\theta} \, \| f(x_t, y_0, t) - \widehat{x}_0 \|_2^2$$

- f_{G_θ} is the ResShift model that would be retrained on current student outputs.
- Unrolling this training loop is computationally expensive.

This objective has a tractable form:

$$L_\theta = -\min_\phi \sum_{t=1}^{T} w_t \, \mathbb{E}_{p_\theta} \Big[-\| f^* \|_2^2 + \| f_\phi \|_2^2 - 2 \langle f_\phi - f^*, \widehat{x}_0 \rangle \Big]$$

- f_ϕ is trained as a **fake ResShift** on generated data.
- Generator updates use teacher f^* and fake model f_ϕ outputs, without backpropagating through an argmin.

We propose a novel objective for 1-step distillation of diffusion SR models in **discrete time** and derive its tractable version.

Latent Space, Supervised Losses, and Multistep Training

Latent-space implementation

- Frozen VAE encodes (y_0, x_0) into (z_y, z_0).
- We compute L_θ and GAN losses in latent space, following DMD2:

$$L_{\text{GAN}} = \mathbb{E}_{p_{\text{data}}} \log D(x_0|y_0) - \mathbb{E}_{p_\theta} \log D(\hat{x}_0|y_0)$$

Multistep training, one-step inference

- Train $G_\theta(z_{t_n}, y_0, t_n, \epsilon)$ over N timesteps t_n.
- Use only $t = T$ at inference.
- $N = 4$ provides the best perception-distortion trade-off.

Correcting teacher predictions with supervised losses

Loss	Space	Role
L_θ	latent	distill teacher trajectory
L_{LPIPS}	pixels	improve perceptual features
L_{GAN}	latent	match HR distribution

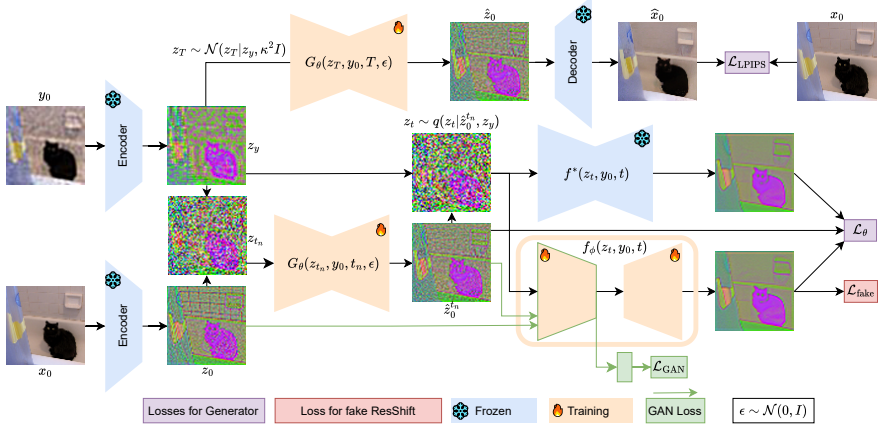
Total generator objective:

$$L_\theta + \lambda_1 L_{\text{LPIPS}} + \lambda_2 L_{\text{GAN}}.$$

Effect of supervised losses on RealSR:

Model	PSNR↑	LPIPS↓	MUSIQ↑
Distill only	24.92	0.355	66.43
Supervised losses	25.91	0.273	65.86

Training Framework



1. Encode (LR, HR) pair into latents (z_y, z_0) with VAE.

2. Train G_θ so that f_ϕ fitted on its noised outputs agrees with frozen f^*.

3. Compute GAN and LPIPS losses in latent and pixel spaces, respectively.

Evaluation Setup and Baselines

We follow the training protocol of our closest competitor, SinSR[4], and use the same teacher.

Training protocol

- HR crops: 256×256 from ImageNet.
- LR generation: Real-ESRGAN degradations with a scale factor $\times 4$.

Main evaluation datasets

- SinSR protocol: RealSR, RealSet65, ImageNet-Test.
- OSEDiff protocol: 512×512 crops from DIV2K-Val, DRealSR, RealSR.

Compared methods and metrics

Group	Methods
Small architectures	ResShift, SinSR, CTMSR
T2I-based models	OSEDiff, AdcSR, PiSA-SR, TSDSR

- Fidelity: PSNR, SSIM.
- Full-reference perceptual: LPIPS, DISTS.
- No-reference quality: CLIPIQA, MUSIQ, NIQE, MANIQA.

[4]Yufei Wang et al. (2024). **"SinSR: Diffusion-Based Image Super-Resolution in a Single Step".** In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

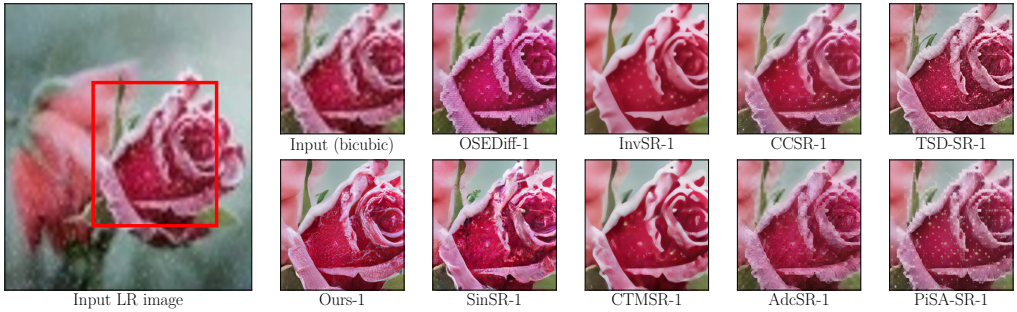
Quantitative Results

Perceptual-distortion trade-off on RealSR and inference complexity on 64×64 LR with $\times 4$ SR.

Model	T2I prior	PSNR↑	LPIPS↓	CLIPIQA↑	Params (M)↓	GPU memory (MB)↓
OSEDiff	yes	25.25	0.299	0.677	1775	3651
AdcSR	yes	25.63	0.300	0.703	456	3940
ResShift	no	**26.49**	0.360	0.596	174	1167
CTMSR	no	26.18	0.294	0.645	**172**	904
SinSR	no	25.83	0.365	0.689	174	570
RSD	no	25.91	**0.273**	**0.706**	174	**539**

Our model improves perception-distortion trade-off among diffusion models without T2I prior. Compared to T2I-based SR models, our model has competitive perceptual results while using much lower computational resources.

Visual Results



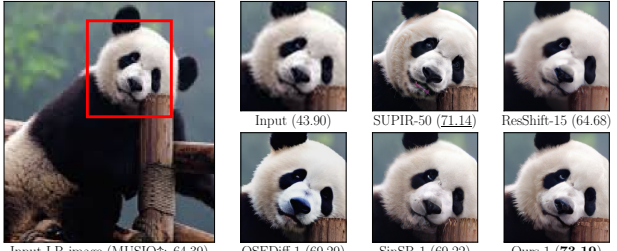
T2I-based SR models add rich details, but they may hallucinate.

ResShift, SinSR, and CTMSR are more conservative and may leave blurred textures.

RSD targets a better trade-off: one-step restoration with sharper details and low computational costs.

One-Step Residual Shifting Diffusion for Image Super-Resolution via Distillation (RSD)

A novel 1-step distillation of diffusion SR models in discrete time.



GitHub

https://github.com/Daniil-Selikhanovych/RSD