DRIFTSR: ONE-STEP REAL-WORLD IMAGE SUPER-RESOLUTION VIA DISTRIBUTION DRIFTING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step real-world image super-resolution (Real-ISR) offers efficient inference, but recovering realistic and perceptually rich details often relies on score distillation or adversarial learning, introducing additional trainable components and making optimization more cumbersome. To this end, we propose DriftSR, a one-step Real-ISR framework that leverages pretrained diffusion priors through distribution drifting. Specifically, we perform drifting in the frozen intermediate representation space of a pretrained diffusion model, without introducing an additional task-specific feature encoder. Building on this space, we introduce Spatial Feature Drifting, which treats spatial features rather than entire images as distributional samples, enabling denser supervision for distribution alignment. To mitigate structural deviations, we further introduce Structure-Modulated Guidance, which adaptively refines drifting guidance according to local structural consistency with the LQ input. Consequently, DriftSR optimizes only the one-step generator, without auxiliary distillation branches or adversarial discriminators. Extensive experiments on three real-world benchmarks demonstrate that DriftSR delivers high-quality super-resolution reconstruction with efficient one-step inference.

1 INTRODUCTION

Real-world image super-resolution (Real-ISR) Zhang et al. (2021); Wang et al. (2021); Zhang et al. (2018a; 2020); Zheng et al. (2026) aims to recover high-quality (HQ) images from low-quality (LQ) observations affected by unknown and complex degradations. The mapping from LQ observations to HQ images is inherently ambiguous, and pixel-level reconstruction objectives often lead to over-smoothed results Ledig et al. (2017); Sajjadi et al. (2016). Therefore, recovering natural image details while maintaining content and structural fidelity to the input remains a central challenge in Real-ISR Wu et al. (2024b); Wang et al. (2024a).

Recently, large-scale pretrained diffusion models have significantly advanced Real-ISR by providing strong generative priors for natural images Wang et al. (2024a); Yu et al. (2024); Wu et al. (2024c); Dong et al. (2025). As the field shifts from iterative diffusion-based restoration toward efficient one-step generation, a key challenge is to recover realistic and perceptually rich details in a single forward pass. To address this challenge, existing diffusion-based one-step SR methods Wu et al. (2024b); Wang et al. (2026a); Zhang et al. (2024); Li et al. (2025) often employ score distillation or adversarial learning to enhance perceptual details. Score distillation approaches Wang et al. (2024b); Wu et al. (2024b); Dong et al. (2025) typically rely on an additional teacher model and trainable LoRA branches, with the LoRA parameters optimized under diffusion-based supervision during training. Adversarial approaches Goodfellow et al. (2014); Zhang et al. (2024); Li et al. (2025) instead introduce an additional discriminator and may suffer from unstable optimization. Despite their effectiveness in improving perceptual quality, both paradigms require auxiliary trainable components and additional objectives, increasing the complexity of the training pipeline. This motivates a simpler alternative that optimizes only the one-step generator while effectively leveraging generative priors.

The recently proposed Drifting Model Deng et al. (2026b) introduces an alternative to score distillation and adversarial learning for distributional generative optimization. It models the relation between source and target distributions in a pretrained feature space and constructs a drifting field guiding generated samples toward the target distribution. By operating on sample distributions without an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

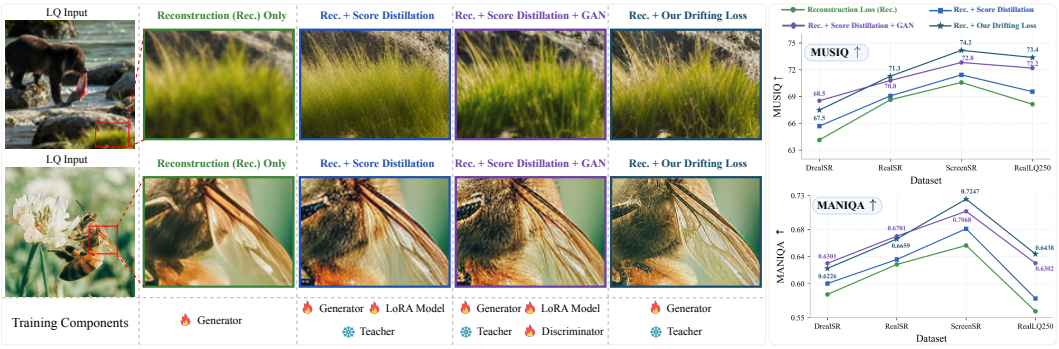


Figure 1: Qualitative and quantitative comparison of different training objectives. **Left**: Reconstruction-only training with MSE and LPIPS produces over-smoothed results. Adding variational score distillation Wang et al. (2024b) improves perceptual details, while further incorporating the GAN objective from OMGSR Wu et al. (2025) enhances texture realism at the cost of additional trainable components. Our drifting loss achieves comparable or better perceptual quality without introducing additional trainable modules. **Right**: MANIQA and MUSIQ scores on four real-world datasets, where drifting remains competitive with joint score distillation and adversarial training.

additional discriminator or a separately optimized distillation branch, distribution drifting presents a potential alternative for one-step Real-ISR.

Nevertheless, directly extending distribution drifting to one-step Real-ISR is non-trivial. ***First***, the effectiveness of the drifting field depends strongly on the feature space used to measure sample relationships, while obtaining a suitable representation may require additional design for Real-ISR. ***Second***, drifting at the image level constructs negative samples by excluding the current sample and therefore requires additional samples under the same condition. Such a requirement can be naturally satisfied in image generation conditioned on class labels or text prompts, where multiple images can share the same condition. In contrast, Real-ISR provides only one paired HQ target for each LQ input, leaving no additional sample for negative construction and thus preventing direct application of drifting at the image level. ***Third***, Real-ISR requires both high perceptual quality and fidelity to the LQ input. Excessive drifting may encourage visually plausible textures that are inconsistent with the input, resulting in incorrect details and structural distortions.

To address these challenges, we propose **DriftSR**, a one-step Real-ISR framework that adapts distribution drifting for perceptual detail enhancement. Inspired by Teacher-Feature Drifting (TFD) Zhang et al. (2026), we employ the pretrained diffusion model as a frozen feature extractor and construct the drifting space using its intermediate representations. This design eliminates the need for an additional task-specific feature network while providing a representation space naturally aligned with the diffusion prior used by the generator. We further introduce **Spatial Feature Drifting**, which reformulates distribution modeling from the image level to the spatial feature level. By treating each spatial location as an individual sample, this formulation enables negative sample construction within each image and provides denser observations for modeling local texture distributions. To balance perceptual enhancement and input fidelity, we propose **Structure-Modulated Guidance**. Based on local structural consistency with the LQ input, this guidance adaptively modulates drifting supervision across spatial locations, thereby suppressing visually plausible but structurally inconsistent details. As shown in Figure 1, these designs collectively enable DriftSR to leverage the pretrained diffusion prior for perceptual detail enhancement while preserving structural fidelity, with optimization restricted to the one-step generator. Our contributions are summarized as follows:

- We propose DriftSR, a one-step Real-ISR framework that introduces Spatial Feature Drifting to model distributions over spatial features and enhance perceptual details without additional trainable components.
- We introduce Structure-Modulated Guidance to adaptively modulate drifting supervision based on local structural consistency with the LQ input, suppressing structural inconsistencies while preserving perceptually rich textures.
- Extensive experiments on real-world benchmarks demonstrate that our DriftSR achieves superior perceptual quality while maintaining efficient one-step inference.

2 RELATED WORK

One-Step Diffusion-based Real-ISR. To reduce diffusion model inference costs in Real-ISR, recent studies have increasingly explored one-step generative models replacing iterative generation with a single forward pass Yue et al. (2023); You et al. (2025); Wu et al. (2024b). These methods aim to improve inference efficiency while preserving high perceptual restoration quality. Existing one-step Real-ISR methods Wang et al. (2026b); You et al. (2025); Dong et al. (2025) commonly adopt distribution or trajectory alignment objectives during training. Among them, distribution alignment is often achieved through score distillation, which constructs supervision from the score discrepancy between a frozen diffusion teacher and a trainable LoRA branch. Some methods further incorporate adversarial learning Zhang et al. (2024); Li et al. (2025); Zhu et al. (2026) to enhance perceptual quality, introducing an additional discriminator. Although these strategies achieve strong restoration performance, the auxiliary trainable components and additional objectives increase the complexity of the training pipeline. Therefore, we explore a simpler alternative that leverages pretrained diffusion priors to guide one-step Real-ISR, without auxiliary distillation branches or discriminators.

Drifting Model. Drifting Model Deng et al. (2026b) introduces a new paradigm for one-step generation by moving distribution transport from inference to training. Unlike diffusion or flow-based models Rombach et al. (2022); Esser et al. (2024); Black Forest Labs that rely on iterative inference, the drifting model learns a one-step generator whose output distribution is progressively updated during training under a drifting field defined by a kernel. The drifting field attracts generated samples toward target samples while repelling them from the current distribution, thereby progressively aligning the two distributions. Since distribution transport is performed during training, inference requires only a single forward pass. The effectiveness of the drifting model depends on the representation space used to measure sample relationships Deng et al. (2026b). For high-dimensional images, distances measured in pixel or latent space may not adequately capture perceptual and semantic similarity. The drifting model therefore uses pretrained representations to construct informative feature spaces, making the choice of feature space important for effective distribution alignment.

3 METHOD

3.1 PRELIMINARY

Drifting Model. Drifting Model formulates generative modeling as the gradual evolution of the model distribution during training. Let f_θ denote a one-step generator that maps a noise variable $\epsilon \sim p_\epsilon$ to a generated sample:

$$x = f_\theta(\epsilon), \qquad x \sim q_\theta, \tag{1}$$

where $q_\theta = (f_\theta)_{\#} p_\epsilon$ denotes the pushforward distribution induced by f_θ. Unlike diffusion or flow-based models, which transport samples along an iterative trajectory at inference time, drifting models progressively update q_θ by optimizing the generator during training. Consequently, the learned generator supports one-step generation through a single forward pass. To guide the evolution of the generated distribution, Drifting Model defines a vector field determined by the target distribution p and the current model distribution q:

$$V_{p,q}(x) = V_p^+(x) - V_q^-(x). \tag{2}$$

The attraction and repulsion components are defined as

$$V_p^+(x) = \frac{\mathbb{E}_{y^+\sim p}\left[k(x,y^+)(y^+ - x)\right]}{\mathbb{E}_{y^+\sim p}\left[k(x,y^+)\right]}, \quad V_q^-(x) = \frac{\mathbb{E}_{y^-\sim q}\left[k(x,y^-)(y^- - x)\right]}{\mathbb{E}_{y^-\sim q}\left[k(x,y^-)\right]}, \tag{3}$$

where $k(\cdot,\cdot)$ is a similarity kernel, typically defined using Laplace distance logits of the form $-\|y - x\|/\tau$, with τ denoting a temperature parameter. The normalized kernel weighting makes $V_p^+(x)$ and $V_q^-(x)$ local mean-shift directions determined by nearby real and generated samples, respectively. Their difference therefore provides a drifting direction that moves the current model distribution toward the target distribution. When $p = q$, the two local mean-shift fields coincide, causing the drifting field to vanish and yielding the equilibrium condition:

$$p = q \quad \Rightarrow \quad V_{p,q}(x) = 0. \tag{4}$$

The drifting field is used to construct a transported target for each generated sample $x = f_\theta(\epsilon)$:

$$\tilde{x} = x + V_{p,q_\theta}(x). \tag{5}$$

The generator is then optimized to regress toward this target using a stop-gradient operation:

$$\mathcal{L}_{\text{drift}} = \mathbb{E}_{\epsilon\sim p_\epsilon}\left[\left\|f_\theta(\epsilon) - \text{sg}\left(f_\theta(\epsilon) + V_{p,q_\theta}\left(f_\theta(\epsilon)\right)\right)\right\|^2\right], \tag{6}$$

The effectiveness of drifting depends on the feature representation used to measure sample similarity. For high-dimensional images, conventional distances may fail to capture semantic similarity. Thus, drifting is performed in an informative feature space extracted by a pretrained model.

Problem Formulation. Real-ISR aims to reconstruct an HQ image x_H from an LQ input x_L. Given a restoration model G_θ, Real-ISR can be formulated as the following optimization problem:

$$\min_\theta \left(\mathcal{L}_{\text{rec}}(G_\theta(x_L), x_H),\ \mathcal{L}_{\text{reg}}(G_\theta(x_L))\right), \tag{7}$$

where $\mathcal{L}_{\text{rec}}$ enforces consistency between the prediction $\hat{x}_H = G_\theta(x_L)$ and target x_H via L_2 or LPIPS. The regularizer $\mathcal{L}_{\text{reg}}$ incorporates natural image priors by aligning generated and real image distributions. Common choices include Variational Score Distillation (VSD) Wang et al. (2024b) and GAN-based objectives Arjovsky et al. (2017). However, VSD typically requires a trainable score model and a pretrained diffusion teacher, while GAN-based regularization introduces a discriminator and may destabilize adversarial training. We instead adopt drifting as the sole distribution regularizer, avoiding auxiliary trainable components while aligning generated and real image distributions.

3.2 Overview of DriftSR

As illustrated in Figure 2, given an LQ image x_L, the generator G_θ predicts the restored image $\hat{x}_H$ under reconstruction loss $\mathcal{L}_{\text{rec}}$, with the paired HQ image x_H as the target. A frozen pretrained teacher extracts multilevel latent features F and $\hat{F}$ from x_H and $\hat{x}_H$, respectively. At each selected layer, spatial features serve as positive samples P and negative samples Q to construct the drifting field $V_{p,q}$ (Sec. 3.3). The same procedure is applied to DWT Mallat (2002) low-frequency components x_H^{low} and $\hat{x}_H^{\text{low}}$ to construct the low-frequency drifting field $V_{p,q}^{\text{low}}$. A structure modulation map $\mathbf{M}$ adaptively combines the two drifting fields into $V_{p,q}^{\text{SMG}}$, which constructs the layer-wise drifting objective. The final drifting loss $\mathcal{L}_{\text{drift}}$ aggregates the objectives across all selected teacher layers (Sec. 3.4). During training, G_θ is optimized using both $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{drift}}$.

3.3 Spatial Feature Drifting

The effectiveness of drifting strongly depends on the feature space for measuring sample similarity. Inspired by TFD Zhang et al. (2026), we employ a frozen pretrained diffusion model as the feature extractor and construct the drifting field in its feature space. This design offers two advantages. First, it avoids a handcrafted or separately trained feature extractor tailored to Real-ISR. Second, the generator is initialized from the same pretrained diffusion model, inheriting a compatible representation hierarchy. This shared initialization aligns the drifting feature space with the generator's inherited representations, enabling more consistent sample similarity measurement during optimization.

Concretely, for the reconstructed image $\hat{x}_H$ and its HQ image x_H, where $\hat{x}_H, x_H \in \mathbb{R}^{C\times H\times W}$, let ψ denote the parameters of the frozen teacher, and let E_ψ and F_ψ denote its VAE encoder and denoising network, respectively. Both images are first encoded into the latent space:

$$\hat{z} = E_\psi(\hat{x}_H), \qquad z = E_\psi(x_H). \tag{8}$$

The resulting latent features are then perturbed using the same Gaussian noise at noise strength t:

$$\hat{z}_t = \hat{z} + t\epsilon, \quad z_t = z + t\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{9}$$

The perturbed latents are fed into the frozen denoising network. Let $h_{F_\psi}^\ell(\cdot)$ denote the intermediate representation extracted from selected layer ℓ. The corresponding teacher features are extracted as

$$\hat{F}^\ell = h_{F_\psi}^\ell(\hat{z}_t, t, c), \qquad F^\ell = h_{F_\psi}^\ell(z_t, t, c), \tag{10}$$

where $\hat{F}^\ell, F^\ell \in \mathbb{R}^{C_\ell\times H_\ell\times W_\ell}$, and c denotes the textual information extracted from the input image.

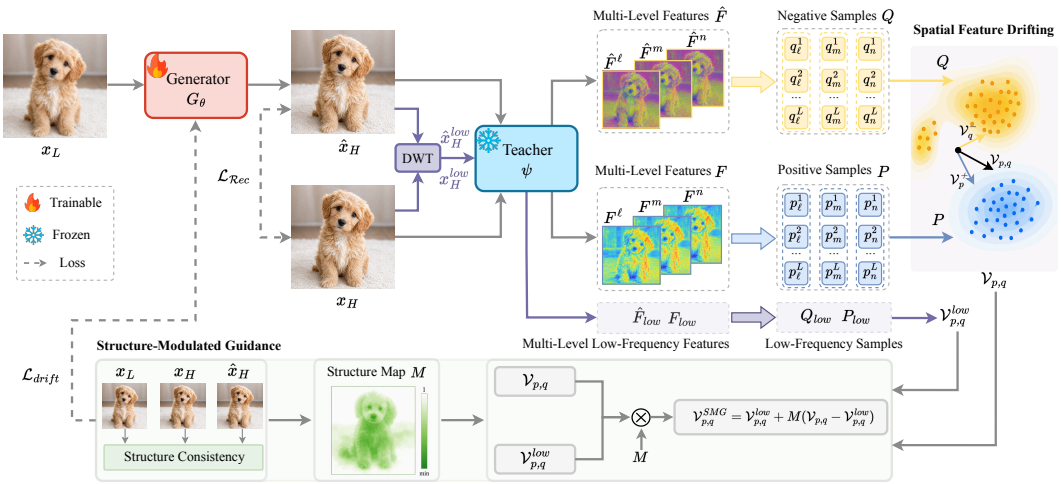


Figure 2: Overview of DriftSR training pipeline. Given an LQ image x_L, G_θ produces $\hat{x}_H$ under reconstruction supervision from the corresponding HQ image x_H. A frozen pretrained teacher extracts multi-level latent features from x_H and $\hat{x}_H$. These features are converted into positive and negative spatial samples to construct the original drifting field. The same procedure is applied to their DWT low-frequency components to construct a low-frequency drifting field. A structure map M adaptively combines the two drifting fields to obtain a structure modulated drifting field for computing the drifting loss. During training, only G_θ is optimized using the reconstruction and drifting losses.

Direct image-level drifting treats each image representation as a single sample, which limits its applicability to Real-ISR because each LQ input is associated with only one paired HQ reference during training. To this end, we exploit the spatial structure in teacher features and treat spatial feature vectors as individual drifting samples. Specifically, at each selected layer ℓ, we unfold the spatial dimensions of the feature maps such that each spatial feature vector forms an individual sample:

$$Q^\ell \in \mathbb{R}^{N_\ell \times C_\ell}, \qquad P^\ell \in \mathbb{R}^{N_\ell \times C_\ell}, \tag{11}$$

where $N_\ell = H_\ell W_\ell$ denotes the number of spatial samples, and Q^ℓ and P^ℓ are constructed from the spatial features $\hat{F}^\ell$ and F^ℓ, respectively. This spatial formulation converts each training pair into N_ℓ feature samples, allowing positive and negative sample sets to be constructed within a single pair. It also provides denser observations for distribution modeling and captures local feature variations that are important for perceptual restoration. Following the drifting formulation introduced in the preliminaries, P^ℓ and Q^ℓ jointly define the drifting field acting on Q^ℓ in the feature space of layer ℓ:

$$V^\ell = V_{P^\ell, Q^\ell}\left(Q^\ell\right). \tag{12}$$

By modeling spatial feature rather than entire image representations, this formulation enables distribution drifting within each paired training example despite limited image-level references.

3.4 Structure-Modulated Guidance

Spatial feature drifting aligns the distribution of reconstructed features with the HQ feature distribution, thereby improving perceptual quality. However, Real-ISR also requires the restored image to preserve the structural information contained in the LQ input. Unconstrained drifting may introduce visually plausible details that are inconsistent with the input structure. To address this issue, we propose Structure-Modulated Guidance (SMG), which estimates local structural consistency and adaptively balances perceptual enhancement and structural fidelity across spatial locations.

We first characterize local structure through image gradients. For an image $\mathbf{x}$, we convert it to luminance $\mathbf{Y}(\mathbf{x})$ and compute its gradient magnitude:

$$\mathbf{\Gamma}(\mathbf{x}) = \sqrt{\left(K_x * \mathbf{Y}(\mathbf{x})\right)^2 + \left(K_y * \mathbf{Y}(\mathbf{x})\right)^2}, \tag{13}$$

where K_x and K_y denote horizontal and vertical Sobel filters, respectively. We then measure the local structural consistency between two spatially aligned images $\mathbf{x_1}$ and $\mathbf{x_2}$ as

$$\mathcal{C}(\mathbf{x_1}, \mathbf{x_2}) = \frac{\mathcal{A}\left[\mathbf{\Gamma}(\mathbf{x_1}) \odot \mathbf{\Gamma}(\mathbf{x_2})\right]}{\sqrt{\mathcal{A}\left[\mathbf{\Gamma}(\mathbf{x_1})^2\right] \odot \mathcal{A}\left[\mathbf{\Gamma}(\mathbf{x_2})^2\right]}}, \tag{14}$$

Table 1: A comprehensive comparison with state-of-the-art methods on three real-world datasets. For VOSR, we report the one-step variant with 1.4B parameters. The best and second-best results under each metric are marked in **red** and **blue** respectively.

Datasets	Metrics	DiT4SR	StableSR	SeeSR	ResShift	CTMSR	OSEDiff	PiSA-SR	TSD-SR	HYPIR	VOSR	FiDeSR	StrSR	DriftSR
T2I Prior		*SD3.5*	*SD2*	*SD2*	–	–	*SD2.1*	*SD2.1*	*SD3.5*	*SD2.1*	–	*SD2.1*	*FLUX*	*SD2.1*
DRealSR	PSNR ↑	25.7966	28.030	28.073	28.693	28.682	27.915	28.318	27.769	25.932	27.824	**29.034**	25.0242	27.393
	SSIM ↑	0.6822	0.7536	0.7684	0.7875	0.7839	0.7833	0.7804	0.7559	0.7197	0.7427	**0.7942**	0.7297	0.7584
	LPIPS ↓	0.3677	0.3284	0.3173	0.3525	0.3236	0.2968	0.2960	0.2967	0.3371	0.3259	**0.2811**	0.3052	0.3379
	CLIPIQA ↑	0.6624	0.6356	0.6909	0.5404	0.6517	0.6974	0.6971	0.7344	0.6392	0.5914	0.6927	0.6435	**0.7442**
	MUSIQ ↑	64.3893	58.512	65.080	52.373	59.766	64.691	66.113	66.599	61.069	66.025	65.359	64.086	**67.490**
	MANIQA ↑	0.6218	0.5594	0.6051	0.4760	0.4853	0.5903	0.6160	0.5874	0.6053	0.6052	0.6184	0.6129	**0.6226**
	TOPIQ ↑	0.5758	0.5323	0.6575	0.4669	0.5960	0.6002	0.6333	0.6182	0.5742	0.6022	0.6043	0.5782	**0.6970**
	QALIGN ↑	3.3353	3.0614	3.5882	2.8515	3.0299	3.5450	3.5838	3.6041	3.5204	3.4570	3.6243	3.5902	**3.6391**
RealSR	PSNR ↑	23.4421	24.645	25.149	**26.382**	25.992	25.148	25.503	24.806	22.826	25.229	26.159	22.927	24.357
	SSIM ↑	0.6647	0.7080	0.7211	0.7347	**0.7546**	0.7338	0.7418	0.7172	0.6783	0.7175	0.7484	0.7065	0.7016
	LPIPS ↓	0.3225	0.3002	0.3007	0.3159	0.2896	0.2920	0.2672	0.2743	0.3087	0.2732	**0.2611**	0.2670	0.3287
	CLIPIQA ↑	0.6404	0.6234	0.6699	0.5505	0.6355	0.6687	0.6699	0.7162	0.6390	0.5625	0.6821	0.6430	**0.7334**
	MUSIQ ↑	68.4939	65.883	69.819	60.225	64.415	69.087	70.147	71.179	66.258	70.584	69.579	68.501	**71.283**
	MANIQA ↑	0.6635	0.6230	0.6450	0.5402	0.5268	0.6337	0.6551	0.6347	0.6436	0.6443	0.6631	0.6502	**0.6659**
	TOPIQ ↑	0.6057	0.5748	0.6890	0.5191	0.5998	0.6251	0.6374	0.6580	0.6069	0.6392	0.6142	0.6228	**0.7275**
	QALIGN ↑	3.4216	3.2767	3.7190	3.1645	3.1608	3.6915	3.6355	3.7149	3.6672	3.5115	3.6898	3.7385	**3.8463**
RealLQ250	CLIPIQA ↑	0.7155	0.5156	0.6998	0.6241	0.6701	0.6723	0.7056	0.7369	0.6899	0.5260	0.7030	0.6539	**0.7947**
	MUSIQ ↑	71.7407	57.485	70.161	61.193	68.018	69.553	71.245	73.217	69.049	62.276	71.624	70.044	**73.386**
	MANIQA ↑	0.6301	0.5116	0.5807	0.5027	0.5080	0.5782	0.6053	0.5924	0.6031	0.4851	0.6095	0.6136	**0.6438**
	TOPIQ ↑	0.6427	0.4464	0.6592	0.5111	0.6340	0.6075	0.6506	0.6468	0.6355	0.4907	0.6447	0.5942	**0.7299**
	QALIGN ↑	4.2144	3.5499	4.1325	3.5917	3.7199	4.2486	4.2061	4.1772	4.2066	2.8910	4.3877	4.2909	**4.3183**

where $\mathcal{A}(\cdot)$ denotes local average pooling over a fixed neighborhood, and $\odot$ denotes element-wise multiplication. We then measure the structural consistency between two spatially aligned images:

$$\mathbf{s}_{\text{pred}} = \mathcal{C}(\hat{\mathbf{x}}_H, \mathbf{x}_L), \quad \mathbf{s}_{\text{gt}} = \mathcal{C}(\mathbf{x}_H, \mathbf{x}_L). \tag{15}$$

The difference between these two maps captures the local structural deviation of the reconstruction from the HQ reference. We use this difference to define the spatial modulation map:

$$\mathbf{M} = \text{clip}\left(1 + \alpha\left(\mathbf{s}_{\text{pred}} - \mathbf{s}_{\text{gt}}\right), m_{\min}, 1\right), \tag{16}$$

where α controls the sensitivity to structural deviation, and $m_{\min}$ specifies the lower bound of the modulation coefficient. When the reconstructed structure exhibits similar consistency with the LQ input as the HQ reference, $\mathbf{M}$ remains close to 1. When the reconstructed structure becomes less consistent with the LQ than the HQ reference, the corresponding modulation coefficient decreases.

Since G_θ starts directly from the LQ input without introducing noise corruption and is trained with reconstruction objectives, the overall content and coarse structure are largely preserved. However, the original drifting field may still introduce high-frequency details that are inconsistent with the input structure. To provide additional structural guidance, we construct a low-frequency drifting field that emphasizes coarse image content while reducing the influence of high-frequency details. Specifically, DWT is applied to $\hat{x}_H$ and x_H to obtain $\hat{x}_H^{\text{low}}$ and x_H^{low}. The same teacher feature extraction and spatial drifting procedure then yields Q_{low}^ℓ and P_{low}^ℓ. The low-frequency drifting field is defined as

$$V_{\text{low}}^\ell = V_{P_{\text{low}}^\ell, Q_{\text{low}}^\ell}\left(Q_{\text{low}}^\ell\right). \tag{17}$$

With the original and low-frequency drifting fields, we employ the structure modulation map $\mathbf{M}$ to adaptively balance their contributions across spatial locations. For each selected layer ℓ, we resize $\mathbf{M}$ to match the spatial resolution of the corresponding teacher features, yielding $\mathbf{m}^\ell$. The structure-modulated drifting field is formulated as

$$V_{\text{SMG}}^\ell = \mathbf{m}^\ell \odot V^\ell + \left(1 - \mathbf{m}^\ell\right) \odot V_{\text{low}}^\ell. \tag{18}$$

Regions with larger modulation values receive stronger guidance from the original drifting field for perceptual detail refinement, whereas regions with smaller values rely more on the low-frequency drifting field to suppress structurally inconsistent details. Finally, we incorporate the structure-modulated drifting field into a unified drifting objective and aggregate it over all selected layers:

$$\mathcal{L}_{\text{drift}} = \sum_{\ell \in \mathcal{S}} \frac{1}{N_\ell C_\ell} \left\| Q^\ell - \text{sg}\left(Q^\ell + V_{\text{SMG}}^\ell\right) \right\|^2. \tag{19}$$

This formulation adaptively balances perceptual detail refinement and structural guidance. By modulating the drifting direction before constructing the optimization target, SMG integrates perceptual enhancement and structural guidance within a single drifting objective.

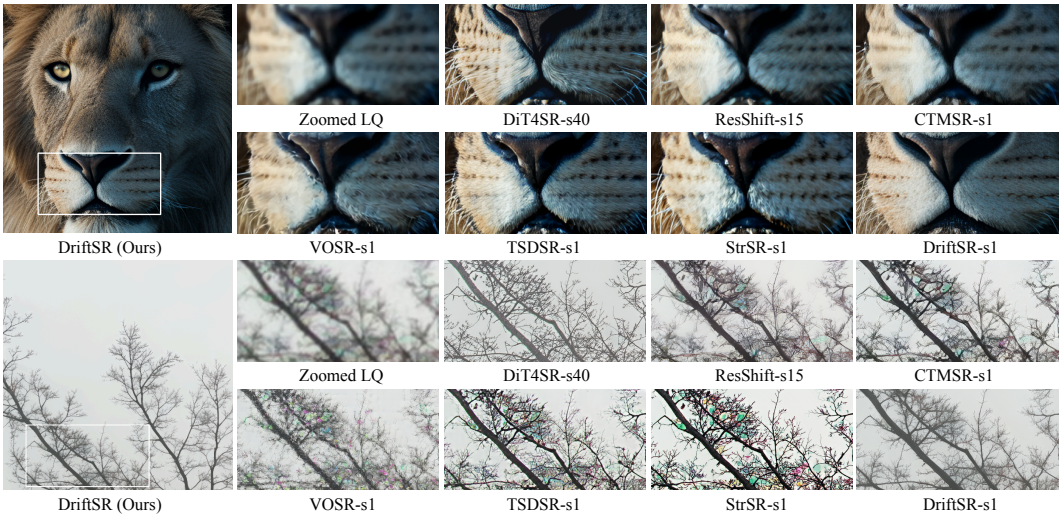


Figure 3: Visual comparison of different methods. s denotes the number of inference steps.

3.5 Training Objective

During training, only the generator G_θ is updated and optimized under the joint supervision of the reconstruction loss $\mathcal{L}_{\text{rec}}$ and the drifting loss $\mathcal{L}_{\text{drift}}$:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{drift}}. \tag{20}$$

The data fidelity term $\mathcal{L}_{\text{rec}}$ is composed of the MSE loss and LPIPS loss Zhang et al. (2018b):

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{MSE}}(\hat{x}_H, x_H) + \lambda_2 \mathcal{L}_{\text{LPIPS}}(\hat{x}_H, x_H), \tag{21}$$

where λ_1 and λ_2 control drifting and perceptual reconstruction objective contributions, respectively.

4 Experiments

4.1 Experimental Settings

Training and Evaluation Datasets. We train DriftSR on LSDIR Li et al. (2023) and construct synthetic LR-HR pairs using the Real-ESRGAN Wang et al. (2021). Degradation-aware prompt extraction Wu et al. (2024c) is adopted to generate descriptions. For evaluation, we conduct experiments on three real-world benchmarks, including RealSR Cai et al. (2020), DRealSR Wei et al. (2020) and RealLQ250 Ai et al. (2024). Implementation details are provided in the **supplementary material**.

Compared Methods. We compare DriftSR with state-of-the-art diffusion-based methods, including one-step methods OSEDiff Wu et al. (2024b), CTMSR You et al. (2025), PiSA-SR Sun et al. (2025), TSD-SR Dong et al. (2025), HYPIR Lin et al. (2025), VOSR Wu et al. (2026), FiDeSR Kim et al. (2026), and StrSR Wang et al. (2026a), as well as multi-step methods StableSR Wang et al. (2024a), ResShift Yue et al. (2023), DiT4SR Duan et al. (2025), and SeeSR Wu et al. (2024c). Specifically, DiT4SR and TSD-SR use SD3.5-Medium Esser et al. (2024), whereas StrSR adopts FLUX.2-klein-4B Black Forest Labs. ResShift, CTMSR, and VOSR are developed without T2I priors.

Evaluation Metrics. We comprehensively evaluate all methods using full-reference and no-reference metrics. Specifically, PSNR and SSIM Wang et al. (2004) measure reconstruction fidelity, while LPIPS Zhang et al. (2018b) assesses perceptual similarity. For no-reference perceptual quality assessment, we report MUSIQ Ke et al. (2021), MANIQA Yang et al. (2022), CLIPIQA Wang et al. (2023), TOPIQ Chen et al. (2024), and QALIGN Wu et al. (2024a). In particular, QALIGN leverages a large multimodal model for image quality assessment and exhibits stronger alignment with human perceptual judgments. We also report user study results in the **supplementary material**.

4.2 Comparison with State-of-the-Art Methods

Quantitative Comparisons. Table 1 compares DriftSR with existing diffusion-based methods on three real-world datasets. The results reveal several observations. (1) DriftSR achieves the best overall

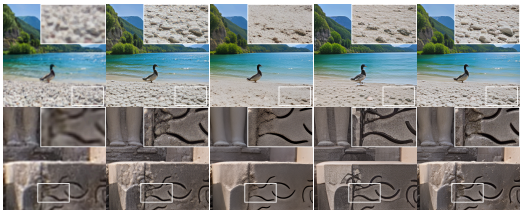
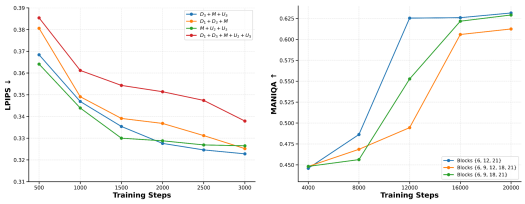


Figure 4: Qualitative representation comparison.



(a) U-Net feature layer. (b) DiT feature layer.

Figure 5: Ablation study on teacher feature layer.

Table 2: Comparison of efficiency among different Real-ISR methods on 512×512 inputs.

	Multi-step Real-ISR Methods				*One-step Real-ISR Methods*							
	StableSR	DiT4SR	SeeSR	ResShift	CTMSR	OSEDiff	PiSA-SR	TSDSR	HYPIR	VOSR-1.4B	FiDeSR	DriftSR
Inference Step	200	40	50	15	1	1	1	1	1	1	1	1
Params (M)	1,563	2,716	2,514	119	171	1,775	1,302	2,207	1,549	1,745	1,291	1,777
Time (s)	14.154	6.947	6.365	1.471	1.312	0.191	0.212	0.122	0.275	0.174	0.171	0.195

performance on no-reference perceptual metrics across all three benchmarks, indicating that DriftSR more effectively leverages the rich generative priors of pretrained diffusion models to recover realistic perceptual details. (2) ResShift, CTMSR, and VOSR employ purely visual architectures without T2I diffusion priors. Although these methods achieve competitive performance on full-reference metrics, they generally underperform DriftSR on no-reference perceptual metrics, suggesting limitations in recovering fine-grained high-frequency details. Meanwhile, DriftSR does not consistently achieve the best full-reference metrics, while maintaining strong perceptual performance. This observation is consistent with the recognized perception-distortion trade-off Blau & Michaeli (2018); Deng et al. (2019). As richer perceptual details are recovered, full-reference metrics may not fully reflect perceptual quality and often favor smoother reconstructions Yu et al. (2024); Wu et al. (2024c).

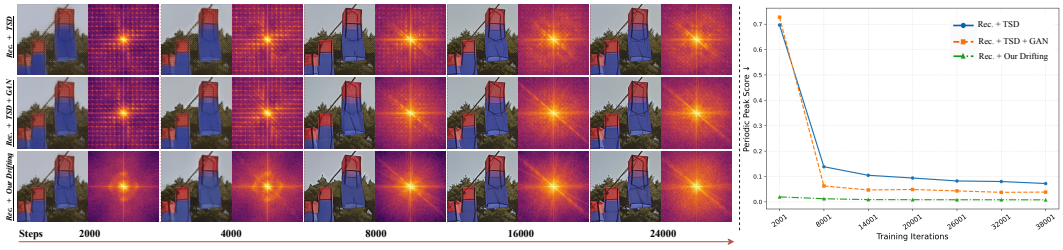
Qualitative Comparisons. Figure 3 presents visual comparisons among different methods. In the first row, DriftSR reconstructs more natural and perceptually realistic textures, particularly fine animal fur. Although ResShift, CTMSR, and VOSR achieve competitive PSNR and SSIM scores, their results exhibit noticeable over-smoothing. In the second row, DriftSR suppresses degradation artifacts and recovers clearer structures with consistent colors, whereas other one-step methods show limited fine-structure recovery and fail to correct color distortions. These comparisons further demonstrate the stronger capability of DriftSR to leverage generative priors.

Complexity Comparisons. We further evaluate the efficiency of Real-ISR methods in Table 2, considering inference steps, runtime, and parameters. We measure time on the $\times 4$ SR task with 128×128 LQ inputs, applying bicubic upsampling before inference on a single NVIDIA 5880 Ada GPU. DriftSR reduces inference cost compared with multi-step methods. It is more than 72 times faster than StableSR, 35 times faster than DiT4SR, over 32 times faster than SeeSR, and 7 times faster than ResShift. Compared with existing one-step models, DriftSR matches OSEDiff in inference speed and is 6 times faster than CTMSR. Despite being slower than VOSR, TSD-SR, and FiDeSR, DriftSR achieves a favorable balance between computational efficiency and restoration quality.

4.3 Ablation Study

Effectiveness of Drifting and SMG. We conduct an ablation study to evaluate the contribution of drifting and Structure-Modulated Guidance (SMG), as shown in Table 3. Introducing drifting improves all no-reference perceptual metrics, demonstrating its effectiveness in enhancing perceptual quality. Compared with the variant without SMG, DriftSR achieves lower LPIPS on both datasets, showing that SMG effectively regulates the drifting process and improves perceptual similarity.

Effect of Feature Space. We ablate the feature space used to compute the drifting loss by replacing the teacher representation with VGG Simonyan & Zisserman (2015), DINOv2 Oquab et al. (2023), and DINOv3-ConvNeXt Siméoni et al. (2025), with three layers selected for each variant. As shown in Table 4, DINOv3-ConvNeXt provides strong perceptual representations but yields worse LPIPS scores of 0.4171 and 0.4222 on RealSR and DRealSR, respectively. Visual comparisons in Figure 4 also show larger structural deviations from the LQ input, suggesting that stronger semantic representations are not necessarily well aligned with the structural constraints of image restoration. In



(a) Qualitative results and corresponding 2D Fourier spectra over training steps. (b) Periodic Peak Score over training steps.

Figure 6: Periodic artifact suppression with the pretrained SD3.5-Medium prior. (a) Comparison across training steps using the reconstruction objective with TSD, additional GAN supervision, and our Drifting loss. (b) Periodic Peak Score across training steps, computed following OP4KSR Deng et al. (2026a), where lower scores indicate fewer periodic artifacts.

Table 3: Ablation study of Spatial Feature Drifting (SFD) and Structure-Modulated Guidance (SMG).

Dataset	Method	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
RealSR	w/o SFD	0.3189	0.6816	68.64	0.6284	0.6739
	w/o SMG	0.3395	0.7407	70.23	0.6521	0.7127
	DriftSR	0.3287	0.7334	71.28	0.6659	0.7275
DRealSR	w/o SFD	0.3247	0.7015	64.13	0.5842	0.6502
	w/o SMG	0.3493	0.7538	68.07	0.6160	0.6852
	DriftSR	0.3379	0.7442	67.49	0.6226	0.6970

Table 4: Ablation study on different feature spaces.

Dataset	Space	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
RealSR	VGG	0.3312	0.7005	69.42	0.6420	0.7018
	DINOv2	0.3213	0.6993	69.92	0.6588	0.7086
	DINOv3	0.4171	0.7380	71.17	0.6584	0.7420
	Teacher	0.3287	0.7334	71.28	0.6659	0.7275
DRealSR	VGG	0.3343	0.7110	65.53	0.5955	0.6756
	DINO2	0.3294	0.7199	66.64	0.6173	0.6900
	DINOv3	0.4222	0.7373	68.41	0.6332	0.7149
	Teacher	0.3379	0.7442	67.49	0.6226	0.6970

contrast, the teacher feature space consistently outperforms VGG and DINOv2 on perceptual metrics across both benchmarks, indicating that it provides a more suitable representation space for drifting and a better balance between perceptual quality and structural fidelity.

Effect of Teacher Layer Configurations. We ablate the teacher layers used to compute the Drifting loss in the U-Net of SD2.1-Base and MMDiT of SD3.5-Medium, as shown in Figure 5. We report LPIPS for U-Net and MANIQA for MMDiT, as they show clearer trends across layer configurations. U-Net features from different stages operate at different spatial scales and encode distinct representations. As shown in Figure 5 (a), D_i and U_i denote the `downsampling` and `upsampling` blocks indexed by i, respectively, while M denotes the `bottleneck` block. Different combinations of three layers yield minor performance variations, whereas increasing the number of selected layers from three to five degrades DriftSR performance. For MMDiT, all transformer blocks operate at the same spatial resolution and differ mainly in depth. Prior work Kim et al. (2025) shows that intermediate DiT blocks can exhibit stronger class semantics, motivating our investigation of layer selection in DriftSR. Figure 5 (b) shows that including intermediate `Block 12` provides no performance benefit. Instead, performance appears more sensitive to the number of selected layers.

Drifting Loss Suppresses Checkerboard Artifacts. With the pretrained SD3.5-Medium prior, MMDiT patch tokenization may cause checkerboard artifacts. Figure 6 (a) shows restoration results and corresponding 2D Fourier spectra across training steps. Combining reconstruction loss (Rec.) with TSD Dong et al. (2025) produces persistent periodic structures in the frequency domain. Adding GAN Wu et al. (2025) supervision gradually reduces these artifacts but fails to prevent their early emergence. In contrast, our drifting loss suppresses dense periodic patterns throughout training. Following the autocorrelation formulation in OP4KSR Deng et al. (2026a), we compute the Periodic Peak Score in Figure 6 (b), confirming the suppression of periodic artifacts by the drifting loss. Periodic Peak Score computation details are provided in the **supplementary material**.

5 Conclusion

In this paper, we presented DriftSR, a one-step Real-ISR framework that exploits pretrained diffusion priors through distribution drifting, with only the one-step generator optimized during training. We introduced Spatial Feature Drifting, which reformulates distribution drifting at the spatial feature level by treating each spatial location as a distributional sample. We further introduced Structure-Modulated Guidance, which adaptively refines the drifting guidance based on local structural consistency with the LQ input. Extensive experiments on three real-world benchmarks demonstrate that DriftSR achieves strong perceptual quality with efficient one-step inference, demonstrating the effectiveness of distribution drifting for Real-ISR.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript to improve grammar, wording, and readability. We did not use generative AI tools for research ideation, hypothesis formulation, methodology or experimental design, implementation, data analysis, interpretation of results, or the development of theoretical or mathematical claims. All AI-assisted edits were reviewed by the authors to ensure that they preserved the intended technical meaning and accurately reflected the authors' original work. We take responsibility for the final content of this work, including text produced with the aid of generative AI.

REFERENCES

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 214–223, 2017.

Black Forest Labs. Flux.2-klein-4b. `https://huggingface.co/black-forest-labs/FLUX.2-klein-4B`.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

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 *2019 IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 2020.

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via wasserstein gradient flows of kde-approximated divergences, 2026.

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. 2022.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, pp. 2404–2418, 2024.

Chengyan Deng, Pengbin Yu, Zhentao Chen, Wei Shen, Kai Zhang, Meng Li, Lunxi Yuan, Xue Zhou, and Li Yu. Op4ksr: One-step patch-free 4k super-resolution with periodic artifact suppression, 2026a.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026b.

Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3076–3085, 2019.

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*, pp. 23174–23184, 2025.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *International Conference on Computer Vision*, 2021.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. Fidesr: High-fidelity and detail-preserving one-step diffusion super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38270–38280, 2026.

Dahye Kim, Xavier Thomas, and Deepti Ghadiyaram. Revelio: Interpreting and leveraging semantic information in diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4659–4669, 2025.

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of score-based and drifting models, 2026.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Unleashing the power of one-step diffusion based image super-resolution via a large-scale diffusion discriminator. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 133581–133600, 2025.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1775–1787, 2023.

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*, pp. 5657–5666, 2022.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44:1–21, 2025.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 2002.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, June 2022.

Mehdi S. M. Sajjadi, Bernhard Scholkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. *2017 IEEE International Conference on Computer Vision*, pp. 4501–4510, 2016.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*, 2015.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. 2025.

Erkan Turan, Nicolas Dufour, and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective, 2026.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C.K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *International Journal of Computer Vision*, pp. 5929–5949, 2024a.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and trajectory regularization for diffusion transformer super-resolution. In *European Conference on Computer Vision*, 2026a.

Ruiqing Wang, Kai Zhang, Yuanzhi Zhu, Hanshu Yan, Shilin Lu, and Jian Yang. Mfsr: Meanflow distillation for one step real-world image super resolution, 2026b.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024b.

Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*, 13(4):600–612, 2004.

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*, pp. 101–117, 2020.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54015–54029, 2024a.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pp. 92529–92553. Curran Associates, Inc., 2024b.

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*, pp. 25456–25467, 2024c.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16311–16321, June 2026.

Zhiqiang Wu, Zhaomang Sun, Tong Zhou, Bingtao Fu, Ji Cong, Yitong Dong, Huaqi Zhang, Xuan Tang, Mingsong Chen, and Xian Wei. Omgsr: You only need one mid-timestep guidance for real-world image super-resolution, 2025.

Sidi Yang, Tianhe Wu, Shu Shi, Shan Gong, Ming Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1190–1199, 2022.

Weiyi You, Mingyang Zhang, Leheng Zhang, Xingyu Zhou, Kexuan Shi, and Shuhang Gu. Consistency trajectory matching for one-step generative super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12747–12756, October 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 13294–13307. Curran Associates, Inc., 2023.

Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors, 2024.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018a.

Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3214–3223, 2020.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *IEEE International Conference on Computer Vision*, pp. 4791–4800, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018b.

Yuan Zhang, Chenyi Li, Haodong Yu, Guoqing Ma, Jiajun Zha, Yuanming Yang, Bo Wang, Wei Tang, Wenbo Li, Haoyang Huang, and Nan Duan. Teacher-feature drifting: One-step diffusion distillation with pretrained diffusion representations, 2026.

Yu Zheng, Kai Zhang, Wei Zhu, Qingguo Liu, Xiantao Hu, Jun Li, and Jian Yang. Dvar: Dynamic visual autoregressive modeling for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23378–23387, June 2026.

Wei Zhu, Kai Zhang, Yu Zheng, Lei Luo, Yong Guo, and Jian Yang. Noise-started one-step real-world super-resolution via lr-conditioned splitmeanflow and gan refinement, 2026.

A HYPERPARAMETER OVERVIEW

We list the main hyperparameters in Table 5.

B JOINT OPTIMIZATION WITH DRIFTING, TSD, AND GAN.

Under the SD3.5-Medium setting, we evaluate different combinations of drifting, TSD Dong et al. (2025), and GAN Wu et al. (2025) on the RealSR dataset. As shown in Table 6, drifting alone consistently improves perceptual quality. Adding TSD or GAN on top of drifting maintains the performance gains of drifting, while providing only limited additional benefit. Similarly, combining all three objectives shows no substantial advantage over Drifting alone. These results suggest that the objectives are compatible, but their benefits are not strongly additive, indicating limited complementarity among them.

To better understand why the additional benefit is limited, we analyze GAN and drifting from the perspective of score difference. GANs obtain optimization signals from a learned discriminator, whereas the drifting model constructs a kernelized vector field directly from data and model samples. Despite this difference, both mechanisms can be interpreted as providing distribution guidance related to the difference between the data score and the model score Cao et al. (2026); Turan et al. (2026); Lai et al. (2026).

Let $p(x)$ denote the data distribution and $q(x)$ denote the model distribution induced by the generator. Under the standard GAN objective, for a fixed generator, the optimal discriminator is given by

$$D^*(x) = \frac{p(x)}{p(x) + q(x)}. \tag{22}$$

Taking the logit of the optimal discriminator gives

$$\operatorname{logit} D^*(x) = \log \frac{D^*(x)}{1 - D^*(x)} = \log \frac{p(x)}{q(x)} = \log p(x) - \log q(x). \tag{23}$$

Thus, the logit of the optimal discriminator corresponds exactly to the log density ratio between the data and model distributions. Differentiating with respect to the input x yields

$$\nabla_x \operatorname{logit} D^*(x) = \nabla_x \log p(x) - \nabla_x \log q(x). \tag{24}$$

Defining the score function of a distribution r as

$$s_r(x) = \nabla_x \log r(x), \tag{25}$$

the above identity becomes

$$\nabla_x \operatorname{logit} D^*(x) = s_p(x) - s_q(x). \tag{26}$$

Therefore, in the ideal discriminator limit, the discriminator provides more than a binary classification signal. The gradient of its logit defines a vector field in sample space that follows the gradient of the log density ratio between the data and model distributions. In this sense, the optimal GAN discriminator induces a transport direction governed by the score difference.

A closely related structure arises in the drifting model. We define the Gaussian kernel as

$$k_\tau(x, y) := C_\tau \exp\left(-\frac{\|x - y\|_2^2}{\tau}\right), \tag{27}$$

Table 5: A summary of the main hyperparameters used in DriftSR for SD2.1-Base and SD3.5-Medium.

SD2.1

Hyperparameter	Value
Generator's Learning Rate	constant 5e-5
Generator's Optimizer	AdamW, betas=(0.9, 0.999), decay=1e-2, eps=1e-8
Batch Size	16 (gradient accumulation = 1)
LoRA Rank	8
Training Resources	4*NVIDIA A800 GPUs
Noise Strength t in Eq.(9)	0.1
α in Eq.(16)	1.0
m_{min} in Eq.(16)	0.5
λ_1 in Eq.(20)	0.15
λ_2 in Eq.(21)	2
Temperature τ	0.02, 0.05, 0.2
Teacher Drifting Layers $\mathcal{S}$	D_3, M, U_3

SD3.5

Hyperparameter	Value
Generator's Learning Rate	constant 5e-6
Generator's Optimizer	AdamW, betas=(0.9, 0.999), decay=1e-4, eps=1e-8
Batch Size	32 (gradient accumulation = 1)
LoRA Rank	64
Training Resources	8*NVIDIA A800 GPUs
Noise Strength t in Eq.(9)	0.1
α in Eq.(16)	1.0
m_{min} in Eq.(16)	0.5
λ_1 in Eq.(20)	0.2
λ_2 in Eq.(21)	2
Temperature τ	0.02, 0.05, 0.2
Teacher Drifting Layers $\mathcal{S}$	Blocks 6,12,21

Table 6: Ablation study on different combinations of drifting, TSD, and GAN on the RealSR dataset.

Components			Metrics					
Drifting	TSD	GAN	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	QALIGN↑	TOPIQ↑
×	×	×	0.3382	0.6568	67.2920	0.5844	3.5542	0.6524
✓	×	×	0.3166	0.6942	69.4425	0.6316	3.6049	0.6740
✓	✓	×	0.3183	0.6803	69.9424	0.6403	3.6123	0.6642
✓	✓	✓	0.3129	0.6990	68.3028	0.6404	3.6094	0.6829

where C_τ is a normalization constant independent of x and y. The corresponding smoothed densities are defined as

$$p_\tau(x) := \mathbb{E}_{Y^+\sim p}\left[k_\tau(x, Y^+)\right],\ q_\tau(x) := \mathbb{E}_{Y^-\sim q}\left[k_\tau(x, Y^-)\right]. \tag{28}$$

The attractive drift induced by the data distribution is defined as

$$V_p^+(x) := \frac{\mathbb{E}_{Y^+\sim p}\left[k_\tau(x, Y^+)(Y^+ - x)\right]}{\mathbb{E}_{Y^+\sim p}\left[k_\tau(x, Y^+)\right]}. \tag{29}$$

Differentiating equation 27 with respect to x yields

$$\nabla_x k_\tau(x, y) = \frac{2}{\tau} k_\tau(x, y)(y - x). \tag{30}$$

Equivalently,

$$k_\tau(x, y)(y - x) = \frac{\tau}{2}\nabla_x k_\tau(x, y). \tag{31}$$

Substituting equation 31 into the numerator of equation 29 gives

$$\begin{aligned}\mathbb{E}_{Y^+\sim p}\left[k_\tau(x, Y^+)(Y^+ - x)\right] &= \frac{\tau}{2}\mathbb{E}_{Y^+\sim p}\left[\nabla_x k_\tau(x, Y^+)\right] \\ &= \frac{\tau}{2}\nabla_x \mathbb{E}_{Y^+\sim p}\left[k_\tau(x, Y^+)\right] = \frac{\tau}{2}\nabla_x p_\tau(x).\end{aligned} \tag{32}$$

The exchange of the gradient and expectation is valid under the standard regularity conditions satisfied by the Gaussian kernel. Dividing equation 32 by the denominator in equation 29 gives

$$V_p^+(x) = \frac{\tau}{2}\frac{\nabla_x p_\tau(x)}{p_\tau(x)} = \frac{\tau}{2}\nabla_x \log p_\tau(x). \tag{33}$$

Similarly, the corresponding drift term induced by the model distribution is

$$V_q^-(x) := \frac{\mathbb{E}_{Y^-\sim q}\left[k_\tau(x, Y^-)(Y^- - x)\right]}{\mathbb{E}_{Y^-\sim q}\left[k_\tau(x, Y^-)\right]} = \frac{\tau}{2}\nabla_x \log q_\tau(x). \tag{34}$$

Combining the data and model terms with opposite signs, the resulting Gaussian drifting field is proportional to the score difference between the two smoothed distributions:

$$V_{p,q}(x) = \frac{\tau}{2}\left(\nabla_x \log p_\tau(x) - \nabla_x \log q_\tau(x)\right) = \frac{\tau}{2}\left(s_{p_\tau}(x) - s_{q_\tau}(x)\right). \tag{35}$$

GAN estimates the density ratio through a learned discriminator, whereas drifting constructs the corresponding direction directly from samples through kernel smoothing. This shared structure provides a possible explanation for the limited additional gains observed when GAN or score based objectives such as TSD are combined with drifting. These objectives may provide partially overlapping distributional guidance rather than independent supervision, and their effects are therefore not necessarily additive.

Table 7: A comprehensive evaluation against state-of-the-art GAN-based methods on real-world datasets. The top-performing results under each metric are marked in **red**.

Datasets	Methods	PSNR ↑	SSIM ↑	LPIPS ↓	CLIPIQA ↑	MUSIQ ↑	MANIQA ↑	TOPIQ ↑	QALIGN ↑
DRealSR	BSRGAN	**28.70**	0.8028	0.2858	0.5092	57.17	0.4845	0.5060	2.96
	RealESRGAN	28.62	0.8051	0.2819	0.4525	54.27	0.4903	0.4623	2.86
	LDL	28.19	**0.8124**	**0.2792**	0.4475	53.95	0.4894	0.4518	2.86
	FeMASR	26.87	0.7576	0.3156	0.5633	53.70	0.4413	0.4670	2.86
	DriftSR	27.39	0.7584	0.3379	**0.7442**	**67.49**	**0.6226**	**0.6970**	**3.64**
RealSR	BSRGAN	**26.38**	**0.7651**	**0.2656**	0.5116	63.29	0.5420	0.5505	3.18
	RealESRGAN	25.69	0.7614	0.2710	0.4494	60.37	0.5505	0.5148	3.11
	LDL	25.28	0.7565	0.2750	0.4555	60.93	0.5495	0.5125	3.09
	FeMASR	25.06	0.7356	0.2936	0.5405	59.05	0.4872	0.5024	3.01
	DriftSR	24.35	0.7016	0.3287	**0.7334**	**71.28**	**0.6659**	**0.7275**	**3.85**

C Periodic Peak Score

Periodic checkerboard artifacts are observed in the reconstructions generated by the MMDiT architecture based on SD3.5-Medium. Following OP4KSR Deng et al. (2026a), we quantitatively characterize these artifacts according to their spatial periodicity. We first examine the autocorrelation responses across different spatial lags. The strongest periodic response occurs at $P_0 = 16$ pixels, with additional prominent responses at its multiples of 32 and 48 pixels. Accordingly, we define $\mathcal{P} = \{16, 32, 48\}$ as the set of target periods used to compute the Periodic Peak Score (PPS). Given the luminance channel Y of an input image, we first suppress low-frequency image structures using Gaussian high-pass filtering:

$$Y_{\text{hp}} = Y - G_\sigma * Y, \qquad \sigma = 2.0, \tag{36}$$

where G_σ denotes a Gaussian filter. We then compute the normalized autocorrelation at spatial offset l along the horizontal and vertical directions:

$$\rho(l) = \frac{\mathbb{E}\left[(A - \mu_A)(B - \mu_B)\right]}{\sigma_A \sigma_B + \epsilon}, \tag{37}$$

where A and B denote two overlapping regions separated by l pixels along the corresponding directions, μ_A and μ_B denote their means, and σ_A and σ_B denote their standard deviations. For each target period P, we compare its autocorrelation response with the mean response at neighboring spatial lags and define the periodic peak response as

$$S(P) = \max\left(|\rho(P)| - \frac{1}{4}\sum_{\delta\in\{-4,-2,2,4\}} |\rho(P+\delta)|, 0\right). \tag{38}$$

This comparison quantifies the extent to which the response at the target period exceeds the local autocorrelation background. Finally, we compute the periodic peak responses at 16, 32, and 48 pixels over a 2×2 grid of local regions and along both spatial directions. Let $q \in \{1, 2, 3, 4\}$ index the four local regions, and let $S_{q,h}(P)$ and $S_{q,v}(P)$ denote the periodic peak responses in the horizontal and vertical directions, respectively. The PPS of an image is then defined as

$$\text{PPS} = \frac{1}{24}\sum_{P\in\{16,32,48\}}\sum_{q=1}^{4}\left(S_{q,h}(P) + S_{q,v}(P)\right). \tag{39}$$

A larger PPS indicates more pronounced autocorrelation peaks at the target periods and thus stronger periodic artifacts with a checkerboard pattern, whereas a lower PPS indicates weaker periodic structures.

Table 8: Ablation study of noise level t.

Dataset	t	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
RealSR	0	0.3254	0.7217	70.85	0.6587	0.7169
	0.05	0.3249	0.7263	71.32	0.6616	0.7263
	0.1	0.3287	0.7334	71.28	0.6659	0.7275
	0.5	0.3371	0.7372	71.09	0.6642	0.7209
	0.8	0.3445	0.7414	70.86	0.6575	0.7214
DrealSR	0	0.3432	0.7395	66.80	0.6177	0.6930
	0.05	0.3416	0.7474	67.18	0.6294	0.6992
	0.1	0.3379	0.7442	67.49	0.6226	0.6970
	0.5	0.3506	0.7391	67.41	0.6203	0.6958
	0.8	0.3650	0.7412	67.47	0.6110	0.6988

Table 9: Ablation study of drifting temperature τ.

Dataset	τ	LPIPS↓	CLIPIQA↑	MUSIQ↑	MANIQA↑	TOPIQ↑
RealSR	$\tau \times 10$	0.3430	0.7419	71.66	0.6681	0.7352
	$\tau \times 5$	0.3365	0.7407	71.53	0.6660	0.7324
	DriftSR	0.3287	0.7334	71.28	0.6659	0.7275
	$\tau/5$	0.3180	0.7184	70.75	0.6584	0.7087
	$\tau/10$	0.3127	0.7016	70.39	0.6575	0.6994
DrealSR	$\tau \times 10$	0.3523	0.7542	68.36	0.6259	0.7143
	$\tau \times 5$	0.3434	0.7482	68.07	0.6226	0.7062
	DriftSR	0.3379	0.7442	67.49	0.6226	0.6970
	$\tau/5$	0.3285	0.7284	66.96	0.6147	0.6850
	$\tau/10$	0.3263	0.7111	66.63	0.6162	0.6713

D Feature Layer Configurations for Different Backbones

In the feature space ablation study, we replace the pretrained diffusion teacher representation with VGG16, DINOv2, and DINOv3-ConvNeXt to investigate the influence of different representation spaces on spatial feature drifting. For a consistent comparison, we select three representative feature levels from each backbone to construct the drifting objective, while keeping the remaining training settings unchanged. All feature extractors are pretrained and kept frozen throughout training.

VGG16. We adopt an ImageNet-pretrained VGG16 as the frozen feature extractor and extract features from the `second, third, and fourth` convolutional blocks. Specifically, the features are taken at the end of each block, before max pooling. These hierarchical representations at different spatial resolutions are used to construct the spatial drifting objective.

DINOv2. We adopt a pretrained DINOv2 ViT-B/14 as the frozen feature extractor. Intermediate representations are extracted from `Blocks 3, 7, and 11`, corresponding to three feature levels distributed across the transformer backbone.

DINOv3-ConvNeXt. We adopt a pretrained DINOv3-ConvNeXt Large as the frozen feature extractor and use the outputs of the first three hierarchical stages, denoted as `Stages 1, 2, and 3`. For each stage, the feature is extracted after the corresponding downsampling layer and ConvNeXt stage. The resulting multi-scale feature maps are then used to construct the spatial drifting objective.

E Comparison with GAN-based Methods

We additionally compare DriftSR with representative GAN-based Real-ISR methods, including BSRGAN Zhang et al. (2021), Real-ESRGAN Wang et al. (2021), LDL Liang et al. (2022), and FeMaSR Chen et al. (2022), as reported in Table 7. The quantitative results reveal distinct performance characteristics across full-reference and no-reference metrics. Several GAN-based methods remain competitive on full-reference metrics such as PSNR and SSIM, whereas DriftSR achieves stronger performance on multiple no-reference perceptual metrics. This advantage may partly benefit from the strong generative prior of the pretrained diffusion model, which facilitates the recovery of realistic textures and details. Since full-reference metrics emphasize consistency with the reference image, they may penalize perceptually plausible details that differ from the reference at the pixel level. These results further reflect the trade-off between reconstruction fidelity and perceptual quality in Real-ISR.

F Ablation Study of Noise Level t

We investigate the effect of the noise level t on the drifting process, as reported in Table 8. Compared with using $t = 0$, a moderate noise level generally improves perceptual quality on both RealSR and DRealSR, as reflected by the gains in multiple perceptual metrics. However, a larger t results in higher LPIPS and lower scores on several perceptual metrics, indicating that excessive noise may degrade the drifting process. Overall, a moderate noise level achieves a better balance between perceptual quality and reference similarity. Therefore, we set $t = 0.1$ as the default value.

G Ablation Study of Temperature τ

We investigate the influence of the drifting temperature τ by uniformly scaling its default values, which are set to 0.02, 0.05, and 0.2 in DriftSR, as reported in Table 9. A consistent tendency can be

observed on both RealSR and DRealSR. Increasing τ shifts the restoration toward stronger perceptual quality while sacrificing reference fidelity, whereas decreasing τ produces the opposite behavior and gradually weakens the perceptual characteristics of the restored images. This result suggests that the drifting temperature mainly controls the trade-off between fidelity-oriented reconstruction and perceptual restoration. We therefore adopt the default temperatures without additional scaling, which provide a balanced operating point between these two objectives.

H User Study

To further assess the perceptual quality of our method from a human perspective, we conduct a user study on 30 LQ images. For each image, participants are presented with the LR input together with the SR results produced by all compared methods, including PiSA-SR, OSEDiff, TSDSR, StrSR, and our DriftSR. Participants are asked to select their preferred result by jointly considering (1) overall perceptual quality and (2) consistency with the LR input, particularly in terms of structural fidelity and texture plausibility. We invite ten volunteers to participate in the study, and each participant evaluates all 30 samples, resulting in a total of 300 votes.

As shown in Figure 7, DriftSR receives the highest number of votes, with 89 out of 300 votes (29.7%), followed by StrSR with 72 votes (24.0%) and TSDSR with 58 votes (19.3%). PiSA-SR and OSEDiff receive 42 votes (14.0%) and 39 votes (13.0%), respectively. These results indicate a stronger human preference for DriftSR among the compared methods, supporting its ability to achieve a favorable balance between perceptual quality and fidelity to the LR input.

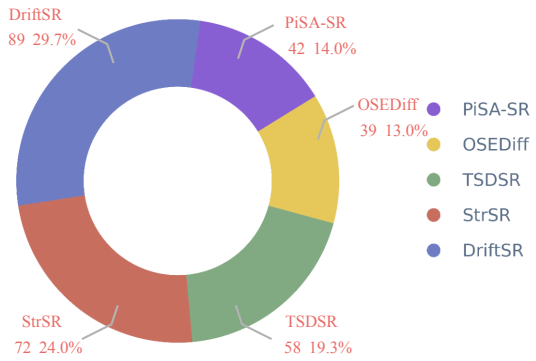


Figure 7: User study results on 30 LR images. DriftSR receives the highest number of votes among all compared methods, suggesting stronger human preference in terms of perceptual quality and consistency with the LR input.

I More Visual Comparisons

In Figures 8 and 9, we present additional visual comparisons between DriftSR and other diffusion-based methods. These examples further demonstrate the effectiveness of DriftSR in recovering realistic details while preserving structural consistency in real-world image super-resolution.

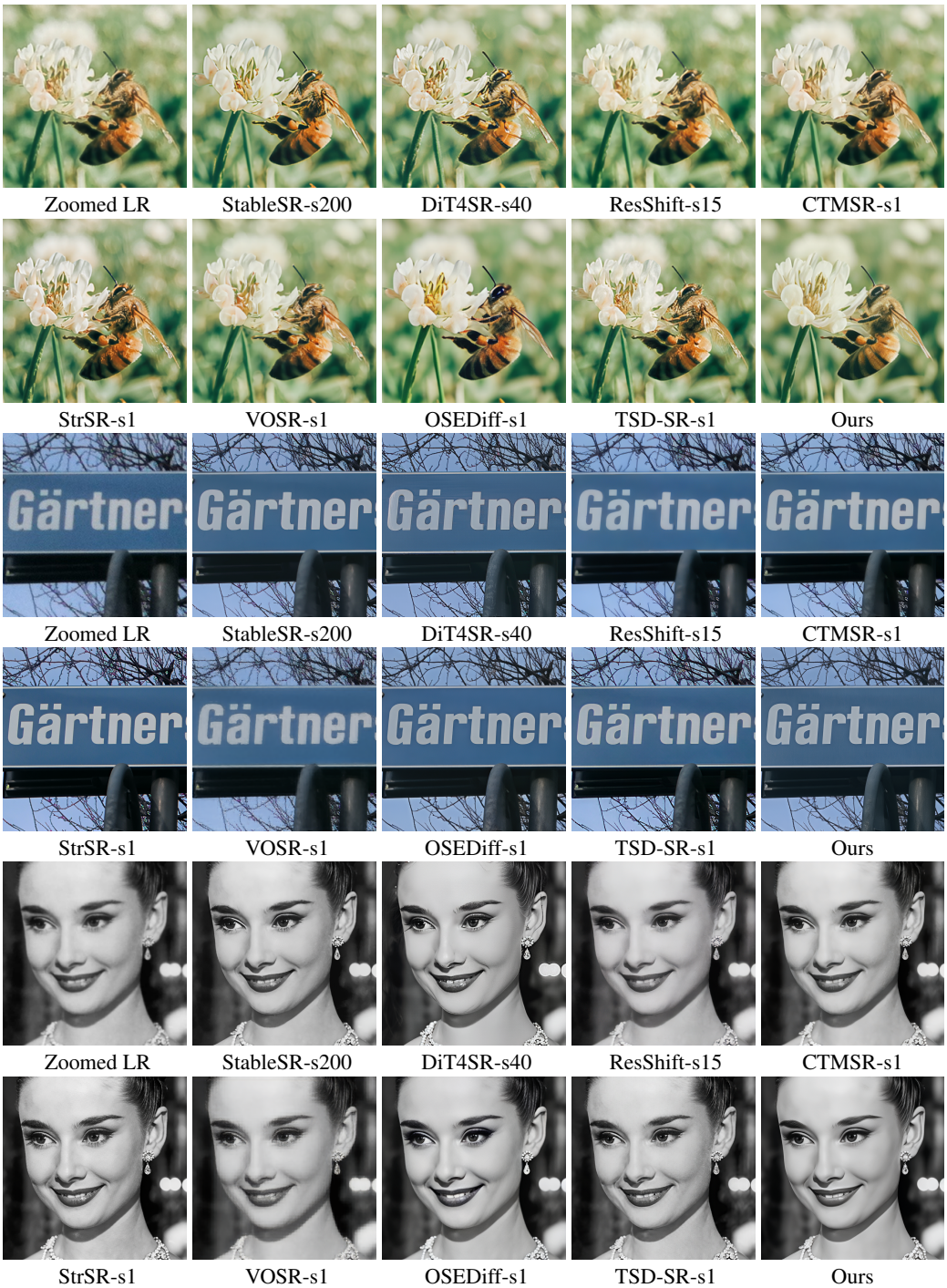


Figure 8: Visual comparisons between DriftSR and other diffusion-based Real-ISR methods. Please zoom in for a better view.



Figure 9: Visual comparisons between DriftSR and other diffusion-based Real-ISR methods. Please zoom in for a better view.