Continuous Degradation Modeling via Latent Flow Matching for Real-World Super-Resolution

Hyeonjae Kim[1*], **Dongjin Kim**[1*], **Eugene Jin**[2], **Tae Hyun Kim**[1†]

[1]Dept. of Computer Science, Hanyang University, Seoul, South Korea
[2]Dept. of Artificial Intelligence Application, Hanyang University, Seoul, South Korea
{khj1112, dongjinkim, eugenebori, taehyunkim}@hanyang.ac.kr

Abstract

While deep learning-based super-resolution (SR) methods have shown impressive outcomes with synthetic degradation scenarios such as bicubic downsampling, they frequently struggle to perform well on real-world images that feature complex, nonlinear degradations like noise, blur, and compression artifacts. Recent efforts to address this issue have involved the painstaking compilation of real low-resolution (LR) and high-resolution (HR) image pairs, usually limited to several specific downscaling factors. To address these challenges, our work introduces a novel framework capable of synthesizing authentic LR images from a single HR image by leveraging the latent degradation space with flow matching. Our approach generates LR images with realistic artifacts at unseen degradation levels, which facilitates the creation of large-scale, real-world SR training datasets. Comprehensive quantitative and qualitative assessments verify that our synthetic LR images accurately replicate real-world degradations. Furthermore, both traditional and arbitrary-scale SR models trained using our datasets consistently yield much better HR outcomes.

Code — https://github.com/present091/DegFlow

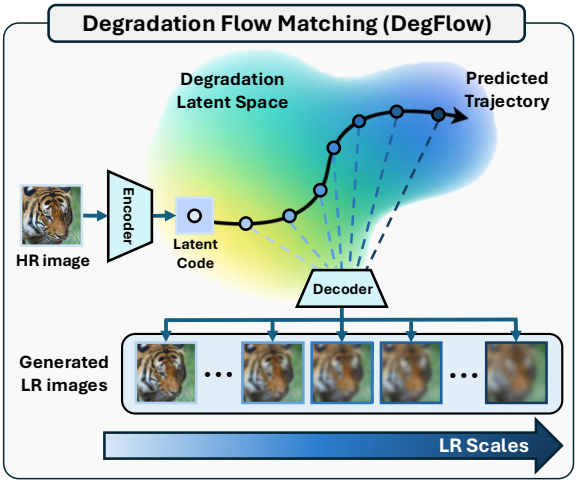

Figure 1: DegFlow generates real-world LR images across continuous scales by modeling degradation trajectories in a learned latent space. The generated LR images are used to train arbitrary SR models for high-quality restoration.

Introduction

Single image super-resolution (SR) aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) observation. Recent deep learning-based methods (Kim, Lee, and Lee 2016; Zhang et al. 2018b; Liang et al. 2021; Chen et al. 2023; Guo et al. 2024, 2025) achieve strong performance in supervised settings by learning an end-to-end mapping from synthetic LR inputs to HR outputs. However, most SR models are trained and evaluated on LR–HR pairs generated with simple operators such as bicubic downsampling. As a result, they often perform poorly on real photographs, where degradations combine unknown blur, noise, and compression artifacts that are not captured by such synthetic pipelines.

One approach to reducing the distribution gap is to augment training data with handcrafted degradation pipelines that consist of blur kernels, noise, downsampling, and compression artifacts (Wang et al. 2021; Zhang et al. 2021). Although these pipelines improve robustness, they still cannot represent the complex properties of real-world degradations (Park et al. 2023; Peng et al. 2025). Another line of research acquires paired HR-LR images with physical devices using DSLR cameras with zoom lenses (Cai et al. 2019; Wei et al. 2020; Fu et al. 2024; Li et al. 2024). While such datasets provide plausible real-world degradations, the collection process is labor-intensive and limits both scale diversity and scene variety. To alleviate these limitations, several recent promising studies learn a degradation model from a small set of real HR-LR images and then synthesize additional LR images to boost realistic SR performance (Wolf et al. 2021; Park et al. 2023; Peng et al. 2025).

Motivated by these approaches, we introduce **DegFlow**, a novel degradation modeling framework, as shown in Fig. 1, that learns real-world degradations from a small set of discrete scale factors (*e.g.,* $\times 2$, $\times 4$) with the corresponding HR image and synthesizes LR images at *unseen* continuous scales (*e.g.,* $\times 2.55$, $\times 3.78$) during inference.

In Tab. 1, we compare the proposed DegFlow with rep-

*These authors contributed equally.
†Corresponding Author.

Generation Method	Realistic LR?	Arbitrary-Scale Generation?	Require only HR for Generation?
Real-ESRGAN, BSRGAN	✗	✗	✓
DeFlow, RealDGen	✓	✗	✓
InterFlow	✓	✓	✗
DegFlow (Ours)	✓	✓	✓

Table 1: Comparison between SR dataset generation methods.

resentative SR dataset generation methods. Unlike Real-ESRGAN (Wang et al. 2021) and BSRGAN (Zhang et al. 2021), which rely on handcrafted operators (*e.g.,* Gaussian noise, blur, bicubic down/up-sampling), DegFlow produces more realistic degradations by modeling degradation in real-world datasets. Compared with DeFlow (Wolf et al. 2021) and RealDGen (Peng et al. 2025), DegFlow offers explicit, scale-specific control, which is essential for training arbitrary-scale SR networks. In contrast to InterFlow, which requires paired LR images at two distinct scales, a setting that may not be applicable in real-world scenarios, DegFlow generates LR outputs from a single HR input at inference time.

DegFlow consists of two modules: Residual Autoencoder (RAE) and Latent Flow Matching (LFM), which are trained sequentially in a two-stage pipeline inspired by latent diffusion models (Rombach et al. 2022; Podell et al. 2024). Specifically, the RAE maps an input image to a compact latent code, reducing computational cost and enabling direct manipulation in latent space. Training on paired HR-LR images embeds degradation cues directly in the latent representation, which benefits the subsequent modeling stage.

LFM learns a continuous degradation trajectory in latent space by training a flow-matching network to a *natural cubic spline* that interpolates the sparse degradation levels available in the training data. In contrast to simple piece-wise linear interpolation in the latent space, the spline model better captures nonlinear geometry in latent space while ensuring the trajectory's first derivative remains continuous, as required by ODE. To further improve perceptual quality, we incorporate an LPIPS loss. Specifically, we project a predicted latent at an intermediate scale (*e.g.,* $\times 3.34$) to its nearest available scale in the training set (*e.g.,* $\times 4.0$), allowing perceptual supervision to be applied even when direct ground truth at the exact target scale is unavailable.

Given a single HR image, the trained LFM samples latent codes at arbitrary points along the predicted latent trajectory path. Subsequently, we utilize the RAE decoder to produce LR images that demonstrate realistic degradations, including those not shown during training. Extensive experiments show that DegFlow produces more realistic degradations than prior methods, while requiring only an HR input at test time, unlike InterFlow, which depends on paired LR examples as well as an HR image. The synthetic datasets generated by DegFlow allow both fixed-scale and arbitrary-scale SR networks to achieve state-of-the-art (SOTA) performance on numerous real-world benchmark datasets.

Related Work

Image Super-Resolution. Single-image super-resolution (SR) remains a fundamental problem in computer vision. Recent deep learning approaches have achieved substantial performance gains. RCAN (Zhang et al. 2018b), a CNN-based SR network, introduces multi-scale skip connections that bypass low-frequency content, allowing the network to concentrate on high-frequency detail. SwinIR (Liang et al. 2021) incorporates window-based self-attention, which enlarges the receptive field while reducing the quadratic complexity of standard attention. MambaIR (Guo et al. 2024) applies a selective structured state-space model to model long-range dependencies with linear computational cost.

Arbitrary-Scale Super-Resolution. The conventional SR models can handle only a discrete set of scale factors, which limits their applicability in scenarios that require continuous zoom. MetaSR (Hu et al. 2019) is the first method to handle arbitrary continuous scales through a Meta-upscale module that predicts scale-conditioned convolution weights. LIIF (Chen, Liu, and Wang 2021) reformulates SR as an implicit neural representation, modeling the image as a continuous function of spatial coordinates and scale.

Real-World SR Dataset. SR models trained on bicubic downsampled synthetic LR images perform poorly on real photographs due to the inability of these operators to capture the complexity of real-world degradations. To reduce the discrepancy between training and real-world test scenarios, BSRGAN (Zhang et al. 2021) and Real-ESRGAN (Wang et al. 2021) propose handcrafted degradation pipelines. These strategies improve robustness, but they still fall short of modeling complex real degradations.

To address this problem, several works manually capture real LR-HR pairs of images using physical equipment (*e.g.,* DSLR cameras). RealSR (Cai et al. 2019), DRealSR (Wei et al. 2020), and RealArbiSR (Li et al. 2024) capture the dataset with multiple focal lengths and then align the images across scales. Although these datasets accurately reflect real degradations, their collection is costly and labor-intensive.

Real-World Degradation Modeling. Recent works reduce data-collection costs by learning the degradation process. DeFlow (Wolf et al. 2021) and NAFlow (Kim et al. 2024) model degradations in latent space using a conditional normalizing flow, while RealDGen (Peng et al. 2025) employs a diffusion model with contrastive disentanglement for unpaired settings. A key limitation of these methods is the absence of explicit scale control, which is essential for training implicit SR networks. InterFlow (Park et al. 2023) addresses controllability by interpolating LR latents with a normalizing flow, thereby synthesizing images at unseen intermediate scales. However, InterFlow requires paired LR observations at two distinct scales, limiting its practicality in real-world scenarios.

Proposed Method

Preliminaries

Real-World Degradation Acquisition. In this paper, we focus on modeling the real-world degradations that occur

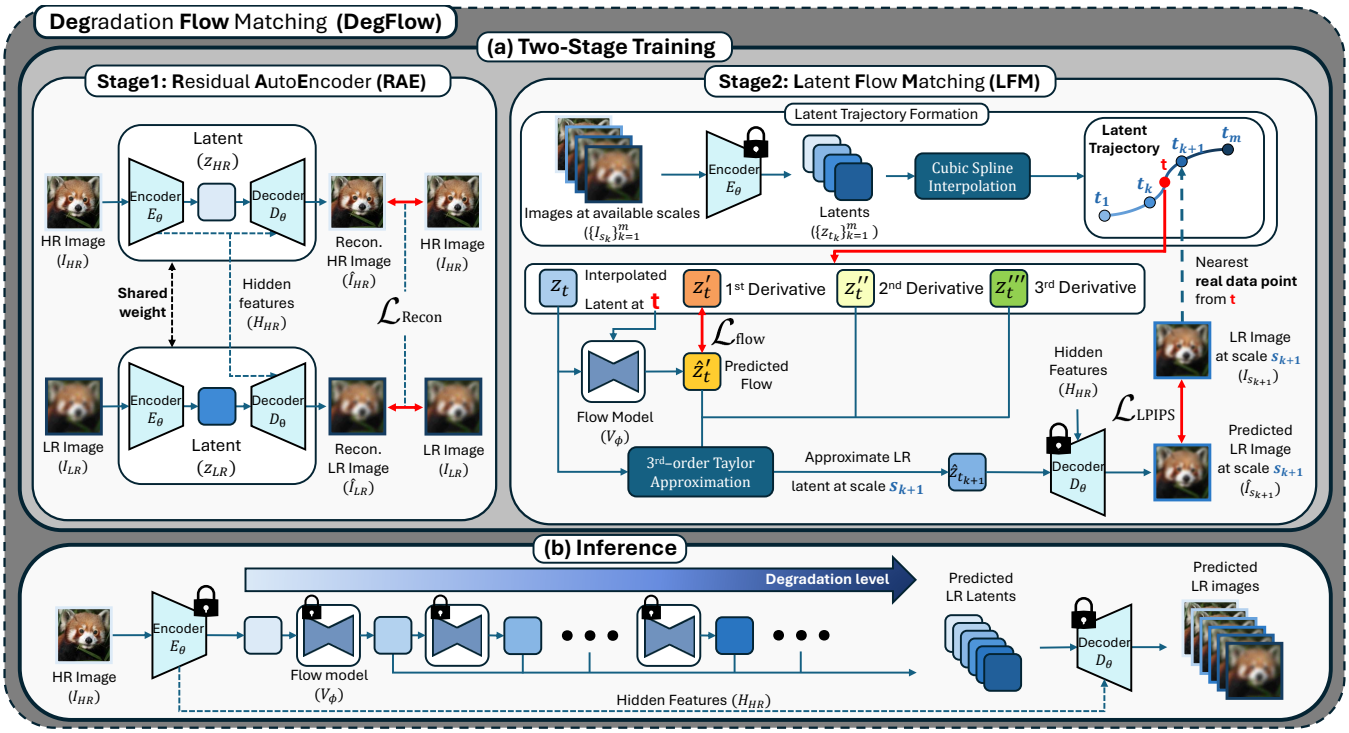

Figure 2: Overview of the proposed method. (a) Two-stage training phase. (b) Inference phase.

in real photographs acquired by DSLR cameras. We follow prior works (Cai et al. 2019; Wei et al. 2020; Fu et al. 2024; Li et al. 2024) to define the image scale s by the ratio of focal lengths. We let $\mathcal{S} = \{s_k\}_{k=1}^m$ denote the discrete set of scales in the dataset. We represent HR images as I_{HR} (or equivalently I_{s_1}) and their corresponding LR counterparts as I_{LR} (or equivalently $\{I_{s_k}\}_{k>1}$). Lower scale values correspond to HR images that preserve finer details, while higher scale values correspond to LR images that exhibit stronger degradations, including blur, noise, and compression artifacts.

Among several real-world SR datasets, we adopt the RealSR benchmark (Cai et al. 2019) for training. RealSR provides carefully aligned LR-HR pairs at multiple scale factors. Specifically, each LR image is aligned to its HR counterpart by correcting misalignment caused by lens distortion and exposure variations through affine registration, followed by luminance compensation. The resulting image pairs $\{I_{\text{HR}}, I_{s_2}, I_{s_3}, I_{s_4}\}$ share the same scene at the same resolution and are both geometrically and photometrically consistent, facilitating accurate modeling of real-world degradations.

Notably, we use the term *degradation level* interchangeably with the scale factor s, following the convention in InterFlow (Park et al. 2023) throughout the paper.

Flow Matching. Flow Matching (FM) (Lipman et al. 2023; Tong et al. 2024; Liu, Gong, and Liu 2023) is a family of generative models that learns a neural velocity field to approximate the true probability flow along a user-specified transport path. Below, we revisit its four essential components: ordinary differential equations (ODEs), vector fields, the FM loss, and probability paths.

First, an ODE specifies how a state x evolves with respect to time $t \in [0, 1]$ as:

$$dx = u(x, t)dt, \tag{1}$$

where u is a vector field and we will use $u(x, t)$ and $u_t(x)$ interchangeably. Given a vector field u and an initial state x_0, the flow (or trajectory) is given as follows:

$$x_t = x_0 + \int_0^t u(x, w)dw. \tag{2}$$

where $u(x, w)$ denotes the exact velocity that transports a target probability path $\{p_t\}_{t\in[0,1]}$. Pushing forward an initial probability distribution p_0 through u_t produces p_t.

FM trains a neural field v_ϕ to regress onto u so that its estimated distribution matches the target distribution p_t at every timestep through the FM loss as:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],x\sim p_t} \left\| v_\phi(x, t) - u(x, t) \right\|_2^2, \tag{3}$$

where time t is sampled from uniform distribution ($\mathcal{U}[0, 1]$). In practice, both $u(x, t)$ and $p_t(x)$ are not given in closed form, making Eq. (3) typically unsolvable. Therefore, Conditional flow matching (CFM) (Tong et al. 2024) is proposed to alleviate this by conditioning on an auxiliary variable ϵ. Specifically, the CFM loss is given as follows:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],x\sim p_t(x|\epsilon),\epsilon\sim q(\epsilon)} \left\| v_\phi(x, t) - u(x, t|\epsilon) \right\|_2^2, \tag{4}$$

where ϵ can be treated as a pair of samples (*e.g.,* HR and LR images) drawn from the joint distribution of initial (source) and target $q(\epsilon) = \pi(x_0, x_1)$. The conditional vector field $u(x, t|\epsilon)$ is defined according to a form of probability path $p_t(x|\epsilon)$. A probability path is a smooth family of distributions $p_{t\in[0,1]}$ connecting p_0 and p_1. Typical choices include stochastic (*e.g.,* Gaussian) bridge (Lipman et al. 2023) and deterministic (*e.g.,* Dirac) bridge (Liu, Gong, and Liu 2023). In this work, we adopt the deterministic bridge as follows:

$$p_t(x \mid \epsilon) = \delta\big(x - \mu_t(\epsilon)\big), \tag{5}$$

where $\delta(\cdot)$ indicates the Dirac delta function and $\mu_t(\epsilon)$ is the interpolant (*e.g.,* linear interpolation) that satisfies the boundary conditions: $\mu_0(\epsilon) = x_0$ and $\mu_1(\epsilon) = x_1$. As setting the intermediate variance to zero removes stochasticity (*i.e.,* $\sigma_t(\epsilon) = 0$), conditional vector field $u(x, t \mid \epsilon)$ in Eq. (4) can be derived as the first derivative of the interpolant $\mu_t'(\epsilon)$ (Liu, Gong, and Liu 2023). Consequently, the CFM loss in Eq. (4) can be reformulated as

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim\mathcal{U}[0,1],x\sim p_t(x|\epsilon),\epsilon\sim q(\epsilon)} \big\| v_\phi(x, t) - \mu_t'(\epsilon)\|_2^2, \tag{6}$$

which remains fully tractable.

Overall Flow

We propose DegFlow, which synthesizes realistic LR images exhibiting real-world degradations while requiring only a single HR input at test time. As illustrated in Fig. 2, DegFlow comprises two components sequentially trained in a two-stage pipeline: a residual autoencoder followed by a latent flow matching model.

Stage 1: Residual Autoencoder (RAE). Motivated by latent diffusion studies (Rombach et al. 2022; Luo et al. 2023), we first train the RAE to map each image to a compact latent code with an $\mathcal{L}_2$ image reconstruction loss. To preserve fine details despite the high compression ratio in the latent space, we incorporate multi-scale skip connections between the encoder and decoder that propagate hidden features from the HR images. In this design, both HR and LR images are used for latent embedding, while only HR features are injected into the decoder through skip connections.

Stage 2: Latent Flow Matching (LFM). Once the RAE is trained and frozen, we embed paired HR-LR images in the latent space and construct trajectories that connect them. The FM network (Song et al. 2021) learns a continuous degradation flow along each trajectory, parameterized by a time variable $t \in [0, 1]$.

Inference. Given a single HR image, the encoder first embeds it into a latent representation. The FM model then evolves this latent over continuous timesteps, and the decoder transforms the evolved latents at arbitrary timesteps back into the image domain. By varying the timestep, DegFlow can synthesize LR outputs corresponding to intermediate scales including previously unseen degradation levels.

Residual Autoencoder (RAE)

As illustrated on the left side of Fig. 2 (a), the RAE consists of an encoder E_θ and a decoder D_θ. Given an input image $I \in \mathbb{R}^{C\times H\times W}$, which can be an HR image I_{HR} or an LR image I_{LR}, the encoder produces a compact latent representation: $z = E_\theta(I) \in \mathbb{R}^{Cr^2\times\frac{H}{r}\times\frac{W}{r}}$, where C, H, and W indicate the channel dimensions, height, and width of the HR image, respectively, and r is the spatial compression factor. A larger r reduces computational cost, but also removes high-frequency details that are essential for high-fidelity restoration (Rombach et al. 2022; Podell et al. 2024; Luo et al. 2023). To mitigate this loss, we propagate multi-scale encoder features to the decoder through residual skip connections as:

$$\hat{I} = D_\theta\big(z;\, H_{\text{HR}}\big), \tag{7}$$

where $H_{\text{HR}} = \{h_{\text{HR}}^{(l)}\}_{l=1}^{L}$ is the set of hidden features on multiple scales, and $h_{\text{HR}}^{(l)}$ denotes the hidden feature at scale level l among L scales. Notably, multi-scale hidden features H_{HR} are extracted from the HR image only, and thus these skip connections enable the decoder to recover fine-grained spatial details while leveraging the compact latent for efficiency.

Reconstruction loss. Inspired by ReFusion (Luo et al. 2023), the RAE is trained with a reconstruction loss $\mathcal{L}_{\text{Recon}}$ applied to HR and LR inputs.

$$\begin{aligned}\mathcal{L}_{\text{Recon}} = \big\|D_\theta(E_\theta(I_{s_1}); H_{\text{HR}}) - I_{s_1}\big\|_2^2 + \\ \big\|D_\theta(E_\theta(I_{s_k}); H_{\text{HR}}) - I_{s_k}\big\|_2^2,\end{aligned} \tag{8}$$

where the scale s_k is drawn uniformly from the available degradation levels in the training dataset, excluding s_1 (*i.e.,* HR scale). This objective ensures that the decoder can faithfully reconstruct the input images while HR features are consistently injected through skip connections, regardless of the input degradation level. As a result, the latent space encodes only the residual information between LR and HR features (*e.g.,* degradation-specific information), providing an informative representation for the subsequent FM model.

Latent Flow Matching (LFM)

This subsection first formalizes the probability path that links latents at different degradation levels and then introduces an auxiliary perceptual loss that further improves visual fidelity.

Probability Path. As illustrated on the right side of Fig. 2 (a), each input image is first embedded into the latent space using the RAE:

$$z_{t_k} = E(I_{s_k}), \quad t_k = \frac{s_k - s_1}{s_m - s_1}, \tag{9}$$

where t_k denotes the min-max normalized timestamp, and s_1 and s_m are the minimum and maximum degradation levels in the scale set $\mathcal{S}$. This normalization linearly maps the degradation level s_k to the timestamp within the range $[0, 1]$. For instance, consider $\mathcal{S} = \{1, 2, 4\}$, where each $s_k \in \mathcal{S}$ is distinctly matched with t_k, and we have $t_1 = 0$, $t_2 = \frac{1}{3}$, and $t_4 = 1$.

We then construct a continuous trajectory that bridges these embedded latents across degradation levels. Several strategies can be adopted for trajectory construction. The

simplest way to connect these marginals is a piecewise linear trajectory, but it is suboptimal because the latent manifold is highly nonlinear. Thus, linear interpolation deviates from natural-image geometry, and the derivative $\mu'(t)$ in Eq. (6) is discontinuous, thus violating the smoothness assumptions required by the ODE formalism (Lipman et al. 2023).

We therefore adopt a *natural cubic spline*, which produces a trajectory with continuous first and second derivatives and a piecewise constant third derivative, whose regularity satisfies the flow matching objective. The deterministic mean trajectory in Eq. (5) is thus defined as a natural cubic spline that interpolates over each sub-interval $[t_k,\, t_{k+1}]$ as:

$$\mu_t(\epsilon) = a_k(\epsilon)(t-t_k)^3 + b_k(\epsilon)(t-t_k)^2 + c_k(\epsilon)(t-t_k) + d_k(\epsilon), \quad (10)$$

where $t \in [t_k,\, t_{k+1}]$, $\epsilon = \{z_{t_k}\}_{k=1}^m$ denotes the set of latent representations at the available degradation levels, and the coefficients $\{a_k(\epsilon), b_k(\epsilon), c_k(\epsilon), d_k(\epsilon)\}$ are obtained by solving a tridiagonal system that enforces the continuity of $\mu_t(\epsilon)$ and its first two derivatives, together with the natural boundary conditions $\mu''_{t_1}(\epsilon) = \mu''_{t_m}(\epsilon) = 0$ (De Boor 1978).

Given this trajectory, the LFM network predicts the velocity $\hat{z}'_t$, which is regressed to the ground-truth velocity using the CFM loss $\mathcal{L}_{\text{CFM}}$ in Eq. (6).

Perceptual Loss. Along with the CFM loss, we introduce an additional perceptual loss to improve visual fidelity at unseen, intermediate degradation scales. Direct supervision at these scales (*e.g.,* $\times 1.532$, $\times 3.361$) is infeasible because ground-truth LR images do not exist. Instead, we approximate the predicted latent $\hat{z}_t$ at an intermediate timestep t satisfying $t_k < t < t_{k+1}$ by extrapolating toward the next degradation level s_{k+1} in the set of training degradation levels $\mathcal{S} = \{s_k\}_{k=1}^m$. Specifically, we employ a third-order Taylor expansion for the extrapolation as follows:

$$\hat{z}_{t_{k+1}} = z_t + \hat{z}'_t\,\Delta t + \tfrac{1}{2} z''_t\,\Delta t^2 + \tfrac{1}{6} z'''_t\,\Delta t^3, \quad (11)$$

where $\hat{z}'_t$ is the predicted velocity from the LFM network, $\Delta t = t_{k+1} - t$, while z_t, z''_t, and z'''_t are computed from $\mu_t(\epsilon)$. Note that $z_t = \mu_t(\epsilon)$ due to the spline defining the deterministic mean path.

The extrapolated latent $\hat{z}_{t_{k+1}}$ corresponds to the coarser degradation level s_{k+1}, for which a ground-truth LR image $I_{s_{k+1}}$ is available. We can therefore compute a perceptual loss between the decoded $\hat{z}_{t_{k+1}}$ and $I_{s_{k+1}}$ using the LPIPS metric (Zhang et al. 2018a) as:

$$\mathcal{L}_{\text{LPIPS}} = \text{LPIPS}\big(I_{s_{k+1}},\, D_\theta\big(\hat{z}_{t_{k+1}}\big)\big). \quad (12)$$

The LPIPS loss back-propagates through Eq. (11) into $\hat{z}'_t$ and subsequently updates the LFM network parameters ϕ in Eq. (6). This enables perceptual supervision to be applied even for intermediate scales without direct ground-truth supervision.

Final Training Objective. The LFM network is optimized with the combined loss as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \lambda\,\mathcal{L}_{\text{LPIPS}}, \quad (13)$$

where λ balances perceptual quality and trajectory fidelity (we set $\lambda = 0.1$ in all experiments). This joint objective encourages LFM to reproduce natural textures while faithfully modeling the underlying degradation trajectory.

Experiments

Experimental Setup

RAE. The RAE is trained using the Adam optimizer to minimize the reconstruction loss in Eq. (8). Training continues for 200k iterations with a cosine-annealed learning rate schedule, decaying from 1×10^{-4} to 1×10^{-7}. Each mini-batch contains 16 randomly cropped 256×256 patches with random horizontal and vertical flips for data augmentation.

LFM. The LFM network uses the Adam optimizer to minimize the CFM and LPIPS losses in Eq. (13) over 400k iterations. A cosine-annealed learning rate schedule decays from 2×10^{-4} to 1×10^{-7}, with mini-batches of 32 randomly cropped 256×256 patches and random flips.

Training. In all experiments, DegFlow is trained on the RealSR-V2 dataset, which contains paired images at degradation levels $\times 1$, $\times 2$, and $\times 4$ from two DSLR camera models: Canon and Nikon. Following InterFlow, we train on the Canon-train dataset and generate LR images from HR images of Nikon-train dataset to test the robustness of our method.

Evaluation. SR performance is evaluated on two real-world benchmarks: RealSR (Cai et al. 2019) and RealArbiSR (Li et al. 2024). These datasets collectively cover a wide range of camera, scene, and degradation characteristics, offering a comprehensive evaluation of generalization.

LR Image Generation Results

Continuous Degradation Modeling Visualization. We first demonstrate the capability of our DegFlow to model continuous real-world degradations in latent space. In Fig. 3 (a), we display RealSR dataset images at discrete scales (HR, $\times 2$, $\times 3$, $\times 4$). Fig. 3 (b) shows LR images generated by DegFlow at uniformly spaced timesteps $0 \le t \le 1$, using the model trained on degradation levels $\mathcal{S} = \{1, 2, 4\}$. Our approach achieves smooth and physically consistent transitions between scales, and the synthesized images exhibit gradual variations in blur and detail loss. These transitions closely match the characteristics of both seen levels ($\times 2$, $\times 4$) and unseen levels ($\times 3$), indicating that DegFlow successfully learns a scale-continuous degradation manifold.

Timestep-Specific Degradation Analysis. To verify whether DegFlow accurately models degradation characteristics across the continuous trajectory, we evaluate its synthesized degradation transition at different timesteps. Fig. 4 shows the normalized PSNR, CLIP (Radford et al. 2021), and FID (Heusel et al. 2017) scores across different timesteps t on the RealSR $\times 3$ test set. For visualization, each metric is normalized to its respective maximum value to facilitate direct comparison, and the FID scores are inverted so that higher values indicate better performance. We observe that PSNR and FID values peak around $t \approx 0.73$ (corresponding to a degradation level of $s \approx 3.2$) and CLIP score peaks around $t \approx 0.70$ (corresponding to $s \approx 3.1$), where the synthesized degradations most closely match the real-world $\times 3$ characteristics. This result demonstrates that DegFlow successfully learns a continuous degradation manifold and captures timestep-specific degradation, both of which are

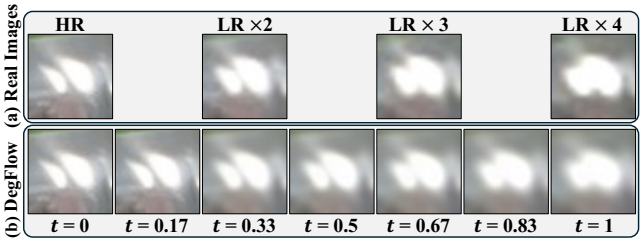

Figure 3: Visualization of continuous degradation. (a) Real images from the RealSR dataset at discrete scales (HR, ×2, ×3, ×4). (b) DegFlow-generated intermediate degradations at evenly spaced timesteps $0 \leq t \leq 1$.

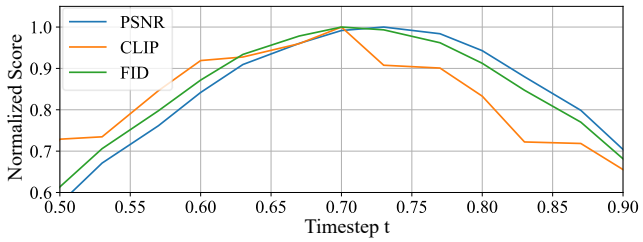

Figure 4: Normalized PSNR, CLIP, and FID scores across different timesteps on the RealSR ×3 test set.

essential for training arbitrary-scale SR models that require scale-specific degradations in the dataset.

Real-world SR Performance

Fixed-Scale SR. Tab. 2 demonstrates quantitative results on the RealSR ×3 test set, comparing five models: RCAN, HAN, SwinIR, HAT, and MambaIR. First, the oracle setting is established by training on RealSR ×3, which directly matches the target degradation level. Next, SR models are trained on RealSR ×2 and ×4, without using the target degradation level. Finally, SR models undergo training using synthetic LR images generated by InterFlow and our model (Ours), ranging from ×2 to ×4. Notably, InterFlow and our model are trained on only RealSR ×2 and ×4 datasets, synthesizing intermediate scales including the target scale (×3).

In the results, the SR models trained with our synthesized dataset consistently outperform those trained on RealSR ×2, ×4 and InterFlow ×2∼×4, achieving higher PSNR and SSIM values while maintaining comparable or better LPIPS. These results validate the effectiveness of our continuous degradation modeling in generating realistic and scale-continuous LR images, enabling SR networks to generalize more effectively to unseen target scales.

Arbitrary-Scale SR. Tab. 3 demonstrates quantitative results for arbitrary-scale SR, where we evaluate three SR models (MetaSR, LIIF, and CiaoSR). RealSR ×3 refers to the ground-truth target scale dataset (oracle setting), without requiring an LR generation method. RealSR ×1 represents the HR images, and is used with LR generation methods

Model	SR Train Set	RealSR ×3 Test Set PSNR↑	SSIM↑	LPIPS↓
RCAN	RealSR ×3	30.68	0.8641	0.3243
	RealSR ×2, ×4	30.30	0.8596	0.3281
	InterFlow ×2∼×4	30.57	0.8631	**0.3155**
	Ours ×2∼×4	**30.72**	**0.8650**	0.3221
HAN	RealSR ×3	30.76	0.8659	0.3216
	RealSR ×2, ×4	30.43	0.8616	0.3261
	InterFlow ×2∼×4	30.68	0.8644	**0.3167**
	Ours ×2∼×4	**30.82**	**0.8660**	0.3212
SwinIR	RealSR ×3	30.69	0.8647	0.3217
	RealSR ×2, ×4	30.23	0.8597	0.3255
	InterFlow ×2∼×4	30.56	0.8634	**0.3166**
	Ours ×2∼×4	**30.78**	**0.8658**	0.3193
HAT	RealSR ×3	30.71	0.8645	0.3221
	RealSR ×2, ×4	30.39	0.8607	0.3248
	InterFlow ×2∼×4	30.65	0.8645	**0.3135**
	Ours ×2∼×4	**30.86**	**0.8668**	0.3186
MambaIR	RealSR ×3	30.62	0.8636	0.3208
	RealSR ×2, ×4	30.29	0.8660	0.3240
	InterFlow ×2∼×4	30.51	0.8625	**0.3138**
	Ours ×2∼×4	**30.73**	**0.8686**	0.3152

Table 2: Fixed-scale SR results on RealSR ×3 test set. Best and second-best are highlighted in **red** and blue.

(Bicubic, BSRGAN, and Real-ESRGAN). Bicubic denotes the conventional arbitrary-scale SR training strategy, where LR images are generated via bicubic downsampling of HR images, and Real-ESRGAN and BSRGAN synthesize LR images for training through a hand-crafted pipeline from the HR images. InterFlow and our model (Ours) are trained on only RealSR ×2 and ×4 datasets, synthesizing intermediate-scale datasets in the range ×2∼×4 to train SR models. Unlike InterFlow, which needs both LR and HR images, we use only HR images.

Compared with the oracle setting, our method achieves higher PSNR and lower LPIPS with comparable SSIM, indicating that it can closely approximate the upper bound without requiring ground-truth LR images at the target scale. Moreover, our method consistently matches or surpasses InterFlow in PSNR and LPIPS across all SR models, with significant perceptual quality enhancements. These results demonstrate that our continuous degradation modeling produces realistic and scale-consistent LR images, enabling arbitrary-scale SR networks to generalize more effectively to unseen target scales.

Ablation Study

For the ablation study, we measure the performance of the HAT SR network on the RealSR ×3 test set.

Effect of Proposed Components. In Tab. 4, we show the effectiveness of each component in our framework. We begin with a baseline variant of DegFlow that adopts a piecewise linear trajectory to model the latent degradation path. Next, using the natural cubic spline for a nonlinear trajectory model improves PSNR and LPIPS, demonstrating the benefits of smooth path modeling for real-world degradations. Then, incorporating the 3rd-order Taylor approximation for the LPIPS-based supervision further improves perceptual fidelity. In addition, introducing skip connections from high-resolution features in the RAE further enhances the reconstruction quality, showing the best performance across all

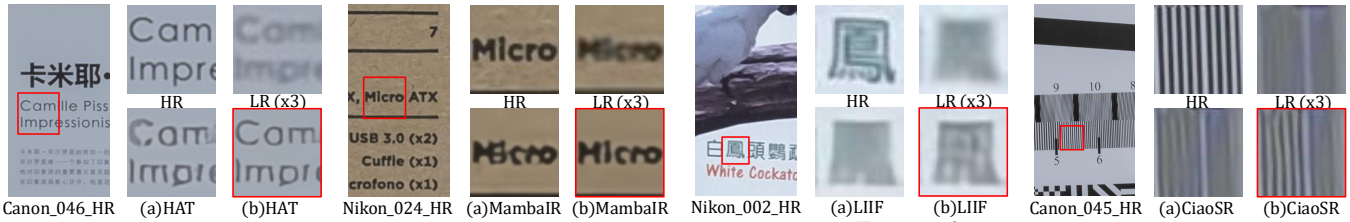

Figure 5: Qualitative comparisons on the RealSR $\times 3$ dataset. Fixed-scale SR results (HAT, MambaIR) and arbitrary-scale SR results (LIIF, CiaoSR) trained with either InterFlow (IF) generated LR (a) or our synthesized LR (b) are compared.

Model	Generation Train set	LR Generation Method	RealSR Test Set ×3 PSNR↑	RealSR Test Set ×3 SSIM↑	RealSR Test Set ×3 LPIPS↓	RealArbiSR Test Set ×2.5 PSNR↑	×2.5 SSIM↑	×2.5 LPIPS↓	RealArbiSR Test Set ×3 PSNR↑	×3 SSIM↑	×3 LPIPS↓	RealArbiSR Test Set ×3.5 PSNR↑	×3.5 SSIM↑	×3.5 LPIPS↓
Meta-SR	RealSR ×3	None (Oracle)	30.43	0.8572	0.3311	29.65	0.8679	0.3330	29.58	0.8338	0.3557	28.15	0.7974	0.3996
	RealSR ×1	Bicubic (Baseline)	28.99	0.8165	0.3488	30.05	0.8473	0.3087	28.77	0.8042	0.3544	27.86	0.7711	0.3886
	RealSR ×1	BSRGAN	28.15	0.8114	0.3867	28.41	0.8326	0.3679	27.41	0.7932	0.3971	26.76	0.7625	0.4199
	RealSR ×1	Real-ESRGAN	26.90	0.8077	0.3813	27.32	0.8177	0.3738	26.50	0.7821	0.4014	25.94	0.7305	0.4380
	RealSR ×2, ×4	InterFlow	30.42	**0.8569**	0.3222	30.71	0.8703	0.3099	29.40	0.8302	0.3504	28.50	0.7983	0.3828
	RealSR ×2, ×4	Ours	**30.58**	0.8565	**0.3190**	**30.88**	**0.8713**	**0.2995**	**29.63**	**0.8321**	**0.3429**	**28.71**	**0.8008**	**0.3780**
LIIF	RealSR ×3	None (Oracle)	30.43	0.8578	0.3324	30.71	0.8718	0.3222	29.56	0.8336	0.3579	28.66	0.8028	0.3861
	RealSR ×1	Bicubic (Baseline)	29.00	0.8167	0.3290	30.04	0.8472	0.3096	28.76	0.8042	0.3550	27.86	0.7711	0.3892
	RealSR ×1	BSRGAN	28.23	0.8133	0.3875	28.28	0.8303	0.3719	27.32	0.7912	0.4009	26.75	0.7617	0.4242
	RealSR ×1	Real-ESRGAN	27.07	0.8090	0.3817	27.24	0.8135	0.3755	26.36	0.7777	0.4025	25.73	0.7492	0.4243
	RealSR ×2, ×4	InterFlow	30.44	**0.8581**	0.3263	30.70	0.8705	0.3144	29.38	0.8307	0.3547	28.44	0.7985	0.3860
	RealSR ×2, ×4	Ours	**30.61**	0.8577	**0.3251**	**30.99**	**0.8729**	**0.3105**	**29.74**	**0.8341**	**0.3517**	**28.78**	**0.8027**	**0.3845**
CiaoSR	RealSR ×3	None (Oracle)	30.65	0.8609	0.3251	30.61	0.8705	0.3105	29.59	0.8339	0.3487	28.54	0.8011	0.3810
	RealSR ×1	Bicubic (Baseline)	28.98	0.8160	0.3496	30.04	0.8472	0.3079	28.76	0.8037	0.3545	27.85	0.7708	0.3881
	RealSR ×1	BSRGAN	28.55	0.8288	0.3638	28.88	0.8443	0.3490	27.90	0.8046	0.3797	27.29	0.7755	0.4069
	RealSR ×1	Real-ESRGAN	27.48	0.8200	0.3696	27.86	0.8341	0.3539	26.95	0.7953	0.3823	26.31	0.7664	0.4091
	RealSR ×2, ×4	InterFlow	30.52	0.8590	0.3162	30.69	0.8702	**0.3057**	29.36	0.8298	**0.3408**	28.52	0.8000	0.3795
	RealSR ×2, ×4	Ours	**30.70**	**0.8590**	**0.3153**	**31.03**	**0.8739**	0.3059	**29.58**	**0.8318**	0.3439	**28.77**	**0.8032**	**0.3793**

Table 3: Arbitrary-scale SR results on RealSR $\times 3$ test set. Best and second-best are highlighted in **red** and blue.

Methods	RealSR ×3 Test Set PSNR↑	SSIM↑	LPIPS↓
Baseline (Piecewise Linear Trajectory)	30.58	0.8640	0.3214
(+) Nonlinear Trajectory (Natural Cubic Spline)	30.68	0.8652	0.3209
(+) LPIPS 3rd-order Taylor Approx.	30.81	0.8662	0.3200
(+) RAE's HR Features Skip Connection	**30.86**	**0.8668**	**0.3186**

Table 4: Impact of each component on RealSR $\times 3$ test set.

Train Set	RealSR ×3 Test Set PSNR↑	SSIM↑	LPIPS↓
DegFlow $\times 2 \sim \times 4$	30.86	0.8668	0.3186
(+) Additional Synthetic Dataset (DIV2K)	**31.00**	**0.8673**	**0.3180**

Table 5: Impact of using external HR images.

metrics. These results validate the importance of each proposed component and highlight their complementary contributions to both perceptual and distortion-based performance.

Effect of External HR Dataset. We further investigate whether synthesizing the SR training set with external HR images can enhance SR performance. In particular, we leverage the DIV2K (Agustsson and Timofte 2017) dataset, which contains high-quality HR images, to generate synthetic LR images for scales in the range $\times 2$ to $\times 4$ using our framework. Notably, our approach can synthesize these intermediate degradations directly from HR images, in contrast to InterFlow, which requires paired LR images captured at multiple degradation levels. As shown in Tab. 5, training the HAT network with the augmented dataset consistently improves PSNR, SSIM, and perceptual quality on the RealSR $\times 3$ test set. Generating high-quality degradations from only HR data offers a practical benefit when paired LR data is unavailable.

Conclusion

We introduce DegFlow, a novel continuous degradation modeling framework for real-world super-resolution. Unlike previous methods that rely on handcrafted degradation pipelines or require paired low-resolution inputs for generation, DegFlow learns a degradation manifold in latent space from only discrete real-world HR-LR pairs and synthesizes realistic degradations at arbitrary, unseen scales using only high-resolution images. By combining a residual autoencoder with latent flow matching, DegFlow effectively captures the nonlinear geometry of real-world degradations while maintaining explicit degradation level control. Experiments show that SR networks trained on our synthetic datasets consistently outperform those trained with existing generation methods in both fidelity and perceptual quality.

Acknowledgments

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00222776), and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022- 0-00156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)) and the research fund of Hanyang University(HY-2025).

References

Agustsson, E.; and Timofte, R. 2017. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*.

Cai, J.; Zeng, H.; Yong, H.; Cao, Z.; and Zhang, L. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. Activating more pixels in image super-resolution transformer. In *CVPR*.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *CVPR*.

De Boor, C. 1978. *A practical guide to splines*, volume 27. Springer.

Fu, H.; Peng, F.; Li, X.; Li, Y.; Wang, X.; and Ma, H. 2024. Continuous optical zooming: A benchmark for arbitrary-scale image super-resolution in real world. In *CVPR*.

Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.-T.; and Li, Y. 2025. MambaIRv2: Attentive state space restoration. In *CVPR*.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. MambaIR: A simple baseline for image restoration with state-space model. In *ECCV*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Hu, X.; Mu, H.; Zhang, X.; Wang, Z.; Tan, T.; and Sun, J. 2019. Meta-SR: A magnification-arbitrary network for super-resolution. In *CVPR*.

Kim, D.; Jung, D.; Baik, S.; and Kim, T. H. 2024. sRGB Real Noise Modeling via Noise-Aware Sampling with Normalizing Flows. In *ICLR*.

Kim, J.; Lee, J. K.; and Lee, K. M. 2016. Accurate image super-resolution using very deep convolutional networks. In *CVPR*.

Li, Z.; Li, M.; Fan, J.; Chen, L.; Tang, Y.; Lu, J.; and Zhou, J. 2024. Learning Dual-Level Deformable Implicit Representation for Real-World Scale Arbitrary Super-Resolution. In *ECCV*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *ICCV*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow matching for generative modeling. In *ICLR*.

Liu, X.; Gong, C.; and Liu, Q. 2023. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2023. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *CVPR*.

Park, S.; Kim, D.; Baik, S.; and Kim, T. H. 2023. Learning controllable degradation for real-world super-resolution via constrained flows. In *ICML*.

Peng, L.; Li, W.; Pei, R.; Ren, J.; Xu, J.; Wang, Y.; Cao, Y.; and Zha, Z.-J. 2025. Towards realistic data generation for real-world super-resolution. In *ICLR*.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-based generative modeling through stochastic differential equations. In *ICLR*.

Tong, A.; FATRAS, K.; Malkin, N.; Huguet, G.; Zhang, Y.; Rector-Brooks, J.; Wolf, G.; and Bengio, Y. 2024. Improving and generalizing flow-based generative models with minibatch optimal transport. In *TMLR*.

Wang, X.; Xie, L.; Dong, C.; and Shan, Y. 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV*.

Wei, P.; Xie, Z.; Lu, H.; Zhan, Z.; Ye, Q.; Zuo, W.; and Lin, L. 2020. Component divide-and-conquer for real-world image super-resolution. In *ECCV*.

Wolf, V.; Lugmayr, A.; Danelljan, M.; Van Gool, L.; and Timofte, R. 2021. Deflow: Learning complex image degradations from unpaired data with conditional flows. In *CVPR*.

Zhang, K.; Liang, J.; Van Gool, L.; and Timofte, R. 2021. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018a. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image super-resolution using very deep residual channel attention networks. In *ECCV*.