NeurOp-Diff: Continuous Remote Sensing Image Super-Resolution via Neural Operator Diffusion

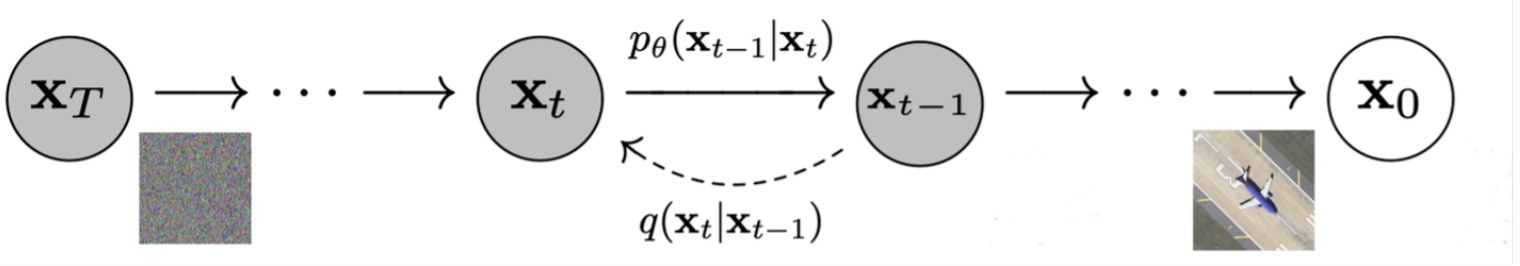
Author
Zihao Xu[1,2], Yuzhi Tang[1,2*],Bowen Xu[1,2], Qingquan Li[1,2]

[1]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), China
[2] Shenzhen University, China

** Corresponding Author*

- Remote sensing image super-resolution is a key technology in remote sensing image processing. Its goal is to restore low-resolution remote sensing images into high-resolution ones, thereby enhancing the spatial details of remote sensing imagery. Although the spatial resolution of modern satellites has improved significantly, many challenges still remain.
- On the one hand, in the process of acquiring high-resolution images, factors such as cloud cover and high acquisition costs may affect the results, making it not always easy to obtain high-quality data.
- On the other hand, different application scenarios–such as urban planning, agricultural monitoring, and environmental protection–require varying levels of image detail and coverage. Continuous super-resolution can dynamically adjust the magnification according to specific needs, providing greater adaptability.

- In recent years, denoising diffusion models have demonstrated outstanding performance in high-fidelity image reconstruction. They possess strong denoising capabilities and produce high-quality generation results. However, their input priors usually rely on simple upsampling operations (such as bicubic interpolation), which are relatively limited in expressiveness, and they are constrained by fixed magnification factors.

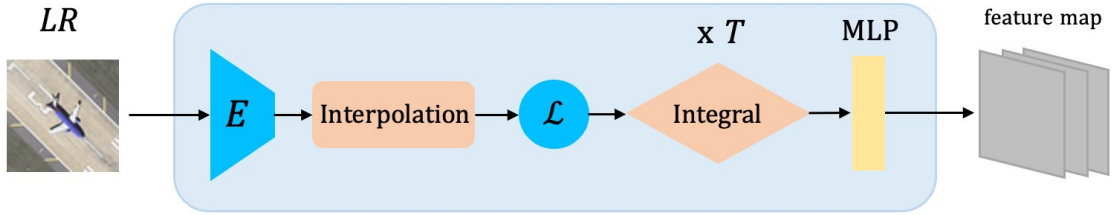


Core idea: Starting from pure noise, a neural network learns a step-by-step denoising process to eventually generate a clear image.

Two stages:

- **Forward process:** Gradually add Gaussian noise to the image until it becomes pure noise.
- **Reverse process:** A neural network learns a gradual denoising process to restore the original image from noise.

- Neural operators can learn mappings between infinite-dimensional function spaces, extract latent continuous high-frequency information, and effectively overcome the limitations of insufficient expressiveness in input priors and fixed resolution. Therefore, we introduce neural operators into the conditional denoising diffusion model to enhance the expressive power of input priors, thereby further improving the quality of generated images and enabling more flexible continuous super-resolution.



Core idea: Model the super-resolution problem as a function space mapping problem. By learning a neural operator, the low-resolution image function is directly mapped to the high-resolution image function.

- **Feature extraction**

Use EDSR to extract low-resolution image features, and combine them with intra-pixel positional offsets to construct input features.

- **Kernel integral layer**

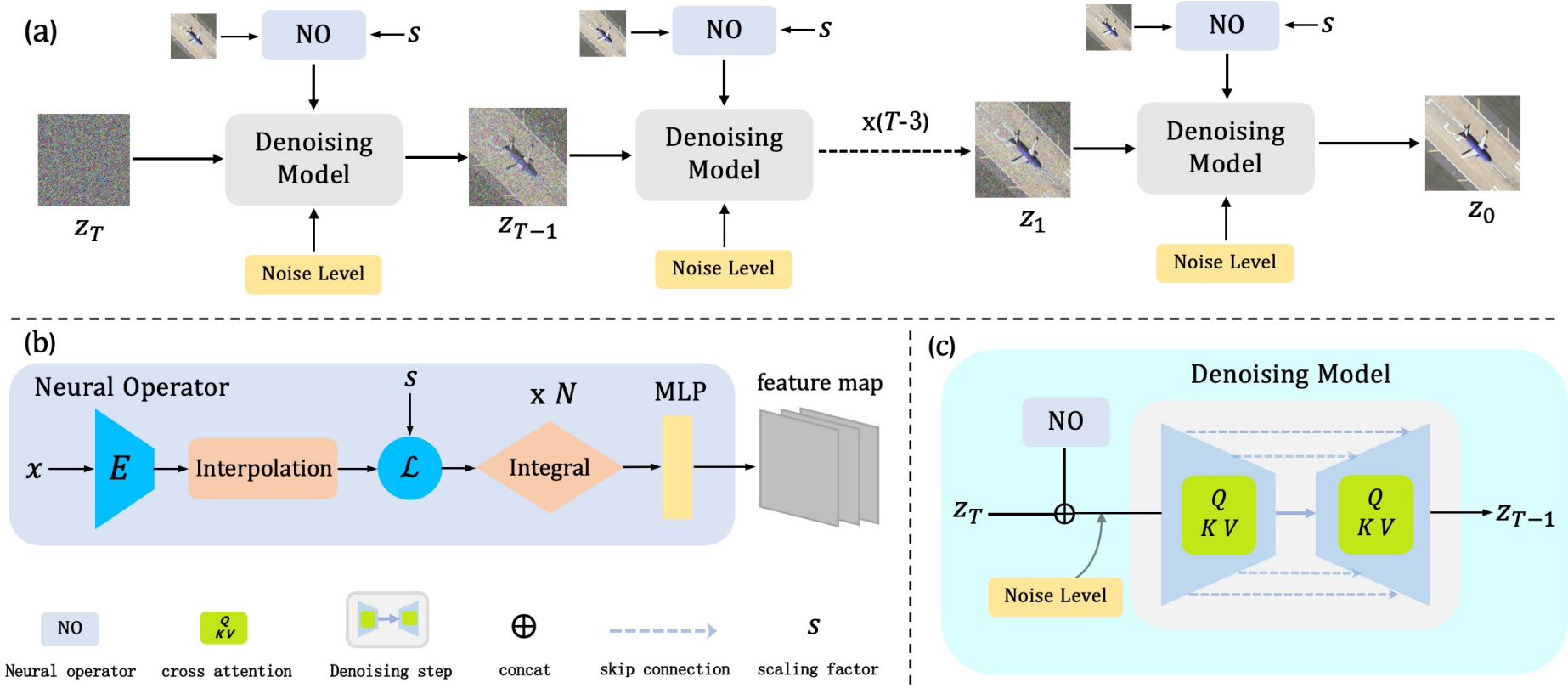
Introduce a Galerkin-type attention mechanism to simulate kernel integral operations, achieving global information aggregation while maintaining linear complexity.

- **Basis function update**

Employ multi-layer attention + feed-forward networks to dynamically adjust hidden representations, enhancing the ability to express high-frequency information.

(a) Overall model architecture. The algorithm's rough flow is that a noisy image undergoes T iterative denoising steps and ultimately generates a clear image.

(c) Integration of the neural operator with the diffusion model. Here Q, K, V denote the query, key and value components in the attention mechanism.



(b) Neural-operator architecture. The low-resolution image is first encoded into high-dimensional features by an encoder (E), an interpolation function, and a lifting operator ($\mathcal{L}$). These features are then passed through a kernel-integral module composed of Galerkin-type attention to produce output features. Finally, an MLP is used for channel transformation.

1. Dataset introduction

- The **UCMerced dataset** contains 21 categories, with 100 images per category, for a total of 2,100 images. The images are uniformly sized at 256 × 256 pixels and are sourced from aerial imagery of different regions in the United States. The dataset mainly includes common land-cover types such as airports, residential areas, commercial areas, and forests.
- The **AID (Aerial Image Dataset)** is built from Google Earth imagery and contains 30 different land-use scene categories, with a total of 10,000 images. The number of images per category ranges from 220 to 420. The image size is 600 × 600 pixels. This dataset covers a broader range of scene categories, including schools, squares, industrial areas, parking lots, and more.
- The **RSSCN7 (Remote Sensing Scene Classification Network 7) dataset** contains 7 categories, with 400 images per category, for a total of 2,800 images. Each image has a size of 400 × 400 pixels. The images are captured under diverse imaging conditions, including different seasons, angles, times, and weather.

2. Model training details

- We first pre-trained the neural operator component for 500 epochs with a batch size of 64. Then, we froze the neural operator's weights and trained the entire NeurOp-Diff model. In this stage, the batch size was set to 10, and training proceeded for about 1 million iterations. We used the Adam optimizer with an initial learning rate of 1×10^{-4} and a minimum learning rate of 2×10^{-6}. For the first 0.1M iterations, the learning rate was kept constant, after which a decay strategy was applied. During training, dropout was set to 0.2.
- For training, the scaling factor s~U(1,M] was sampled from a uniform distribution to cover multiple magnification levels. For the super-resolution task, larger T values usually yield better results. Therefore, in this experiment, we set T=2000. The L1 norm was chosen as the loss function.
- Our experiments were conducted on a single RTX 4090 GPU. During inference, we adopted a uniform sampling strategy to speed up the process. Results showed that setting the number of diffusion steps to 50 produced good inference performance.

3. Generative Model Comparison

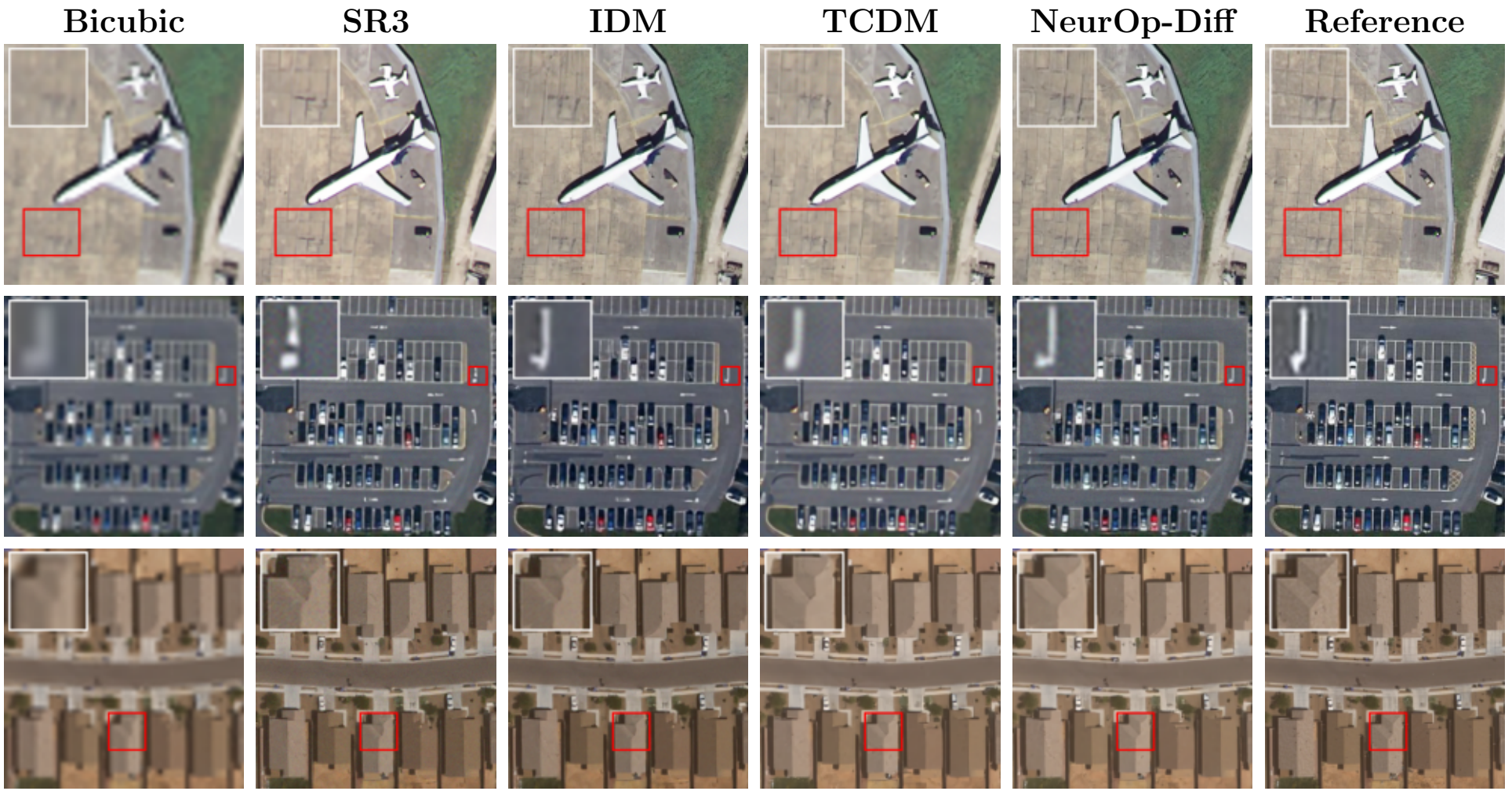
- Our method (**NeurOp-Diff**) was compared with other generative diffusion models on the UCMerced, AID, and RSSCN7 datasets for 4× and 8× super-resolution.
- In terms of **Peak Signal-to-Noise Ratio (PSNR)** and **Structural Similarity Index (SSIM)**, our approach achieved higher scores than the other models (as shown in the figure).

Dataset	Methods	x4		x8	
		PSNR	SSIM	PSNR	SSIM
UCMerced	SR3 [34]	27.15	0.7587	23.12	0.5846
	IDM [12]	27.66	0.7606	23.68	0.6131
	TCDM [52]	27.83	0.7679	24.17	0.6422
	NeurOp-Diff	**28.14**	**0.7761**	**24.51**	**0.6502**
AID	SR3 [34]	26.24	0.6709	22.54	0.5231
	IDM [12]	27.03	0.6778	22.90	0.5487
	TCDM [52]	27.53	0.6840	23.37	0.5584
	NeurOp-Diff	**27.57**	**0.6845**	**23.45**	**0.5599**
RSSCN7	SR3 [34]	26.22	0.5835	22.96	0.5329
	IDM [12]	27.18	0.6235	23.37	0.5603
	TCDM [52]	27.71	0.6489	23.69	0.5873
	NeurOp-Diff	**27.74**	**0.6509**	**23.73**	**0.5889**

ICCV HONOLULU HAWAII
OCT 19-23, 2025

深圳大學 SHENZHEN UNIVERSITY

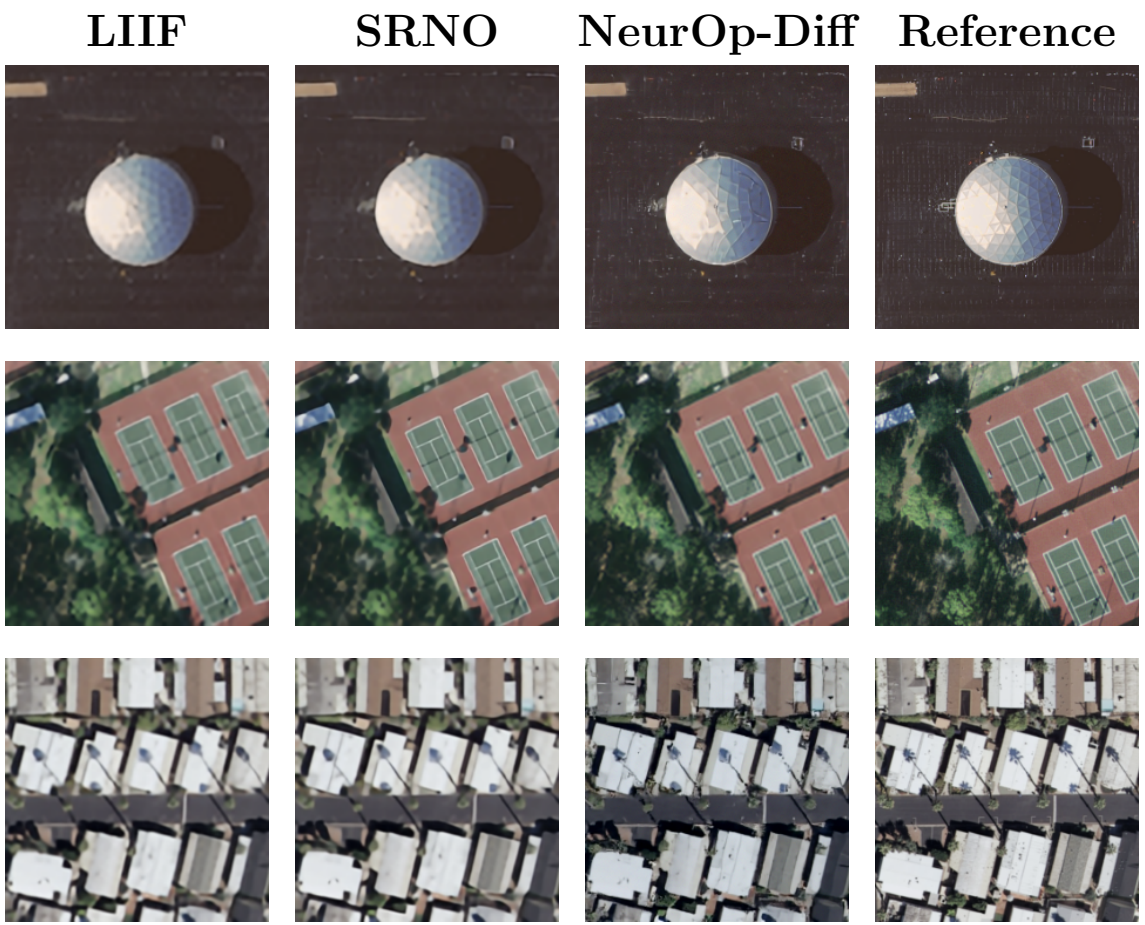
GUANGMING LABORATORY
光明实验室
GUANGMING LABORATORY



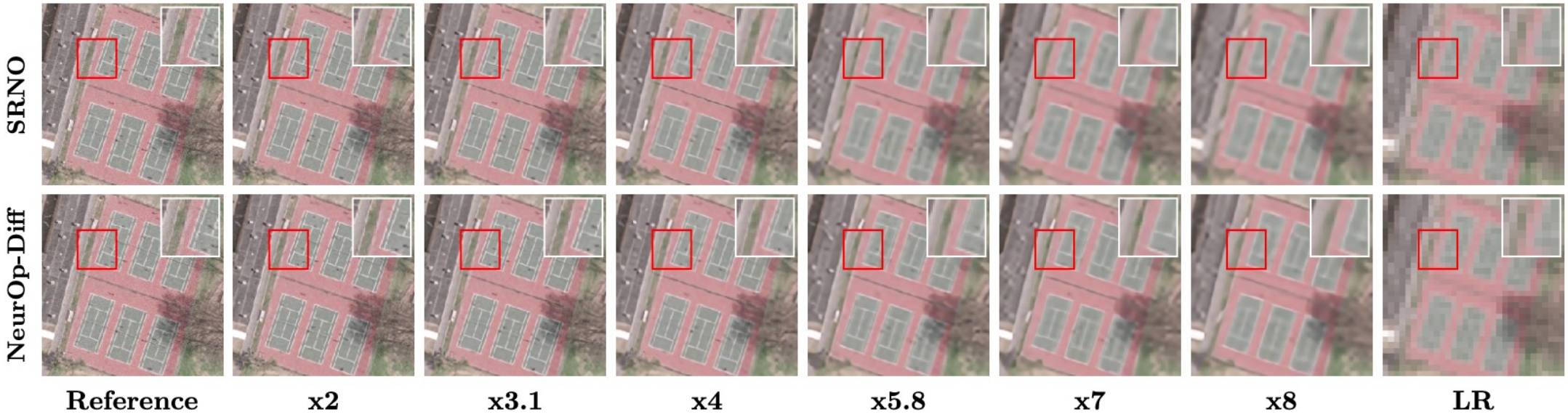
4. Comparison of Regression-Based Continuous Super-Resolution Models

- We conducted a comparison with two regression-based continuous super-resolution models (LIIF and SRNO) under 4× super-resolution.
- As shown in the figure, from the quantitative results, although LIIF and SRNO achieved higher values in **Peak Signal-to-Noise Ratio (PSNR)**, our model demonstrated better performance in **Structural Similarity Index (SSIM)**. In addition, regarding **Learned Perceptual Image Patch Similarity (LPIPS)**, our model generated images that showed smaller differences from the ground-truth images.

Method	PSNR↑	SSIM↑	LPIPS↓
LIIF [5]	28.96	0.69	0.182
SRNO [44]	**29.21**	0.72	0.164
NeurOp-Diff	28.14	**0.78**	**0.136**



5. Comparison of Continuous Super-Resolution Models



- We randomly selected multiple scaling factors within the range of (1, 8] and visualized the results under these magnifications using SRNO and our method.
- As shown in the figure, visually, our model demonstrates better performance in the clarity of court boundary lines and the natural texture transitions of the surrounding environment. These advantages become more pronounced at higher magnification levels. In contrast, SRNO often struggles to fully recover complex textures at larger scaling factors, resulting in generated images that lack detail.

As shown in the figure, from a quantitative perspective, although SRNO achieves better **Peak Signal-to-Noise Ratio (PSNR)** at lower magnification factors (e.g., 2× to 3.1×), our model delivers superior performance in terms of **Structural Similarity Index (SSIM)** and **Learned Perceptual Image Patch Similarity (LPIPS)**, aligning more closely with human visual perception. As the magnification factor increases, the advantages of our model become more pronounced, especially in challenging scenarios with higher magnifications (e.g., 7× to 10×).

Method	Metric	in-distribution					out-of-distribution	
		2×	3.1×	5.8×	7×	8×	9×	10×
SRNO [44]	PSNR	**34.32**	**31.05**	**27.19**	**25.73**	**24.67**	**24.32**	**23.92**
	SSIM	0.786	0.746	0.681	0.649	0.623	0.607	0.592
	LPIPS	0.118	0.142	0.194	0.207	0.215	0.221	0.228
IDM [12]	PSNR	33.54	30.32	26.37	25.02	23.68	23.34	23.05
	SSIM	0.811	0.781	0.692	0.646	0.613	0.601	0.587
	LPIPS	0.103	0.119	0.187	0.208	0.219	0.225	0.230
NeurOp-Diff	PSNR	33.93	30.76	26.89	25.50	24.51	24.13	23.88
	SSIM	**0.824**	**0.792**	**0.717**	**0.673**	**0.650**	**0.641**	**0.633**
	LPIPS	**0.098**	**0.112**	**0.159**	**0.188**	**0.195**	**0.206**	**0.212**

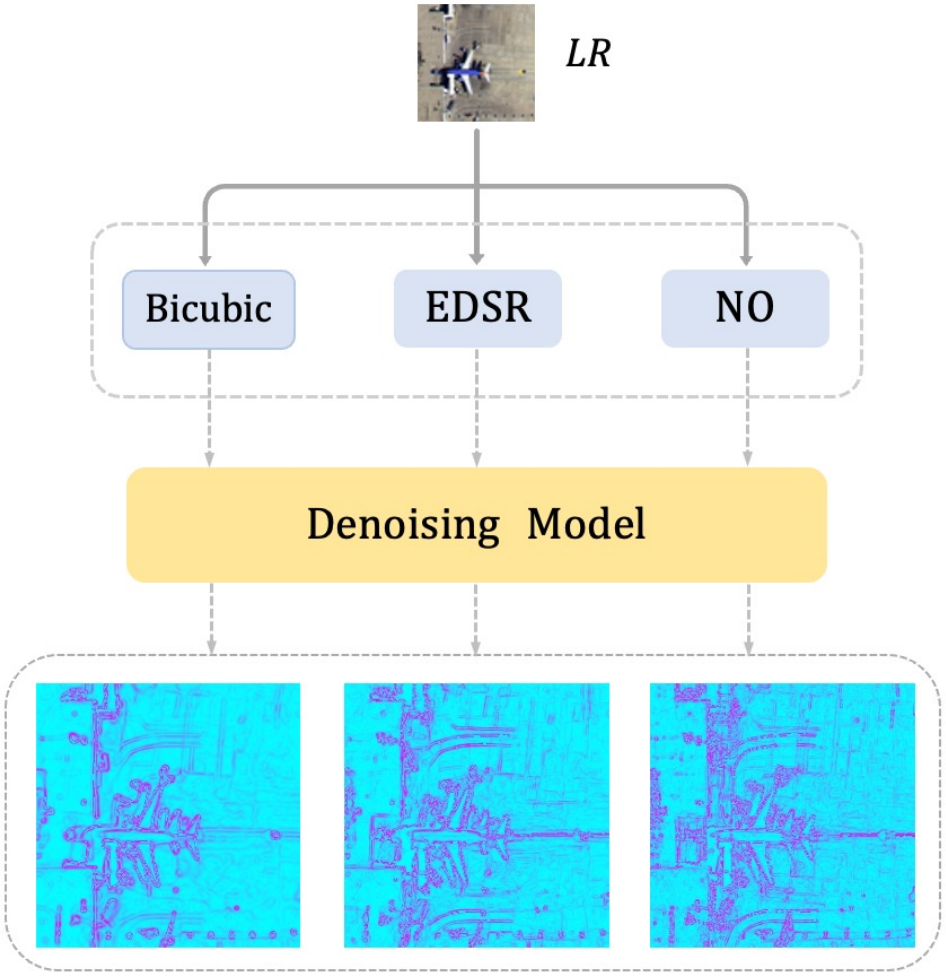
6. Ablation experiment

We replaced the scale-adaptive conditional network in our model with three different conditioning mechanisms to construct comparison models:

(1) Directly concatenating the upsampled low-resolution image with the ground-truth image.
(2) Concatenating the low-resolution image features encoded by EDSR with the ground-truth image.
(3) Concatenating the low-resolution features encoded by the neural operator (NO) with the ground-truth image.

From the quantitative results, our adaptive conditional network provides richer prior knowledge.

	Bicubic	EDSR	Neural Operator
PSNR↑	27.15	27.42	**28.14**
SSIM↑	0.7587	0.7602	**0.7761**



ICCV OCT 19-23, 2025 HONOLULU HAWAII

深圳大学 SHENZHEN UNIVERSITY

GUANGMING LABORATORY 光明实验室 GUANGMING LABORATORY

Thank you for watching !