Fidelity- and Perception-Aware Local Implicit Attention for Arbitrary-Scale Image Super-Resolution

Yu-Syuan Xu[1,2] Hao-Lun Sun[2] Hao-Wei Chen[1]
Hsien-Kai Kuo[2] Chun-Yi Lee[1]
{Yu-Syuan.Xu, Hsienkai.Kuo}@mediatek.com
{d13922023,cylee}@csie.ntu.edu.tw

[1]National Taiwan University, Taipei City 100233, Taiwan
[2]MediaTek Inc., Hsinchu City 300, Taiwan

Abstract. Arbitrary-scale image super-resolution (ASISR) aims to reconstruct high-resolution images from low-resolution inputs over a continuous range of upscaling factors. While traditional pixel-regression approaches often produce overly smooth results that lack realistic details, recent diffusion methods can produce sharper and more realistic textures. However, these diffusion techniques frequently introduce the risk of structural hallucinations. To address these issues, we propose Fidelity- and Perception-Aware Local Implicit Attention (FPLIA), a framework that effectively integrates fidelity-oriented features into a diffusion pipeline to produce realistic and faithful reconstructions for ASISR. We introduce a Fidelity and Perception Attention Module (FPAM), which applies both self-attention and cross-attention to fidelity-oriented and perceptual features to enhance representational capacity. To further exploit their complements, we design a Fidelity and Perception Select Module (FPSM) that adaptively selects the most representative features for RGB values prediction. We conduct extensive experiments to validate the effectiveness of these components. Both qualitative and quantitative results show that FPLIA delivers superior perceptual realism while maintaining reconstruction accuracy on standard ASISR benchmarks. The source code is accessible at the following repository: https://github.com/XUSean0118/FPLIA.

Keywords: Arbitrary-scale image super-resolution · Local implicit function · Diffusion Model

1 Introduction

Arbitrary-scale image super-resolution (ASISR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) inputs at any continuous upscaling factor within a single model [3–5, 7, 9, 12–14, 17, 22, 25–27]. This formulation provides a practical and flexible alternative to fixed-scale super-resolution methods [6, 15, 16, 20, 24, 30, 31], which require training and maintaining a separate network for each target resolution. A prevailing paradigm in ASISR combines deep feature extractors with local implicit image functions, *i.e.*, coordinate-conditioned

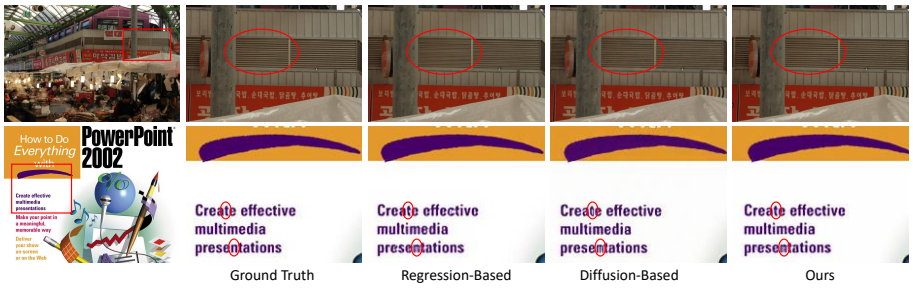

Fig. 1: Regression-based and diffusion-based approaches for arbitrary-scale image super-resolution (ASISR) exhibit complementary failure modes. Regression methods (*e.g.*, LIT [4] with SwinIR [15]) reconstruct structurally faithful images that suppress fine textures, such as the blinds in the first row. Diffusion methods (*e.g.*, Kim *et al.* [13] with LDM [19]) recover perceptually realistic textures, yet may hallucinate content absent from the ground truth, such as the spurious characters in the second row. FPLIA harnesses both feature types through cross-feature attention and adaptive per-pixel selection, producing reconstructions that are both faithful and perceptually sharp.

networks that map 2D query positions and their associated latent features to RGB pixel values [3–5, 7, 12–14]. The reconstruction quality of such methods depends critically on the nature of the underlying latent representations. As Fig. 1 illustrates, features extracted by pixel-regression backbones [15, 16, 31] trained with ℓ_1 or ℓ_2 objectives yield images with high pixel-wise fidelity, yet the reconstructions tend to suppress fine-grained textures and high-frequency details. Features generated by diffusion models [19, 20], by contrast, recover perceptually convincing textures, yet they risk introducing hallucinated structures that deviate from the ground-truth content. This fundamental tension between reconstruction fidelity and perceptual realism poses two intertwined challenges for ASISR. First, fidelity-oriented and perceptual features encode substantially different information: one prioritizes low-level pixel accuracy, while the other captures distributional statistics that favor visual realism. Facilitating productive interaction between these heterogeneous representations without mutual degradation remains an open problem. Second, the relative utility of each feature type is not spatially uniform: structurally simple regions benefit primarily from fidelity features, whereas richly textured areas demand stronger perceptual cues. A framework that fuses the two in a content-agnostic manner inevitably compromises one objective for the other across different image regions.

Existing ASISR approaches have addressed the fidelity-perception divide primarily by committing to one paradigm. Regression-based methods such as LIIF [5], LTE [14], CiaoSR [3], LIT [4], and HIIF [12] pair local implicit functions with pixel-regression backbones and achieve strong fidelity-oriented performance, yet consistently underperform on perceptual metrics, as the regression objective inherently suppresses high-frequency variations. Diffusion-based approaches such as IDM [7] and Kim *et al.* [13] replace the regression backbone with a generative model, improving perceptual quality at the expense of reconstruction fidelity; hallucinated textures and structural distortions arise when the generative process

lacks sufficient structural constraint from the input, particularly in regions with ambiguous or repetitive content. Regarding cross-feature interaction, Kim *et al*. [13], the closest effort toward bridging the two paradigms, conditions its diffusion pipeline on regression-derived features, yet the two streams remain loosely coupled, without mechanisms for bidirectional information exchange. The ASISR landscape thus remains divided along the fidelity-perception axis.

Beyond the surface-level fidelity-perception dichotomy identified above, the complementarity between the two feature types exhibits a finer structure governed by two axes of variation. Along the spatial axis, the relative informativeness of each feature type depends on local image content: structurally simple regions are already well-served by fidelity features, while richly textured areas require perceptual cues to recover plausible high-frequency details. Along the scale axis, fidelity features grow progressively less informative at higher upscaling factors, as the LR input retains less high-frequency content and perceptual features become increasingly necessary to synthesize fine details that the fidelity stream can no longer provide. This dual variability implies that no fixed combination strategy can serve all spatial locations and scale factors simultaneously. Effective integration instead requires two capabilities: cross-feature interaction that produces diverse candidate representations spanning different fidelity-perception profiles, and per-pixel selection that identifies the most reliable candidate at each queried coordinate. The interaction between the two feature streams is itself asymmetric in nature. Fidelity features provide a structurally anchored scaffold capable of guiding perceptual features toward more faithful textures, while perceptual features inject high-frequency diversity that enriches otherwise smooth fidelity representations. This asymmetry suggests that the two directions of information exchange between the streams should be modeled as separate pathways rather than through a single symmetric operation such as feature concatenation.

Building on these observations, this paper presents **Fidelity- and Perception-aware Local Implicit Attention (FPLIA)**, a framework that integrates fidelity-oriented and perceptual features for ASISR through explicit cross-feature interaction and adaptive per-pixel feature selection. FPLIA addresses the two challenges above with two complementary modules. The Fidelity and Perception Attention Module (FPAM) models asymmetric bidirectional interaction between the two feature streams, producing four candidate representations at every queried coordinate, each capturing a different balance between pixel-level accuracy and perceptual realism. The Fidelity and Perception Select Module (FPSM) estimates the reliability of each candidate by combining content-dependent attention cues with scale-dependent signals, and commits to the single highest-confidence candidate at every position. The framework is designed to be backbone-agnostic: the fidelity-oriented extractor and the perceptual generator can be instantiated with different architectures, as demonstrated with several representative choices in the supplementary material. Extensive experiments across integer and non-integer upscaling factors on DIV2K [1], Set5 [2], Set14 [28], B100 [18], and Urban100 [10] confirm that FPLIA achieves a Pareto-optimal balance between fidelity and perceptual quality, simultaneously improving both PSNR and LPIPS over the leading

diffusion-based methods across the majority of settings. The key contributions of this work are as follows:

- Interactions between fidelity and perceptual feature streams produce more diverse latent representations than either stream in isolation. This finding motivates the design of FPAM, which employs self-attention within each branch and cross-attention across branches to generate four candidate features with distinct fidelity-perception profiles at every queried coordinate.
- Adaptive per-pixel feature selection, guided by confidence scores derived from attention maps and scale-dependent cues, enables content-aware reconstruction. The proposed FPSM estimates the reliability of each candidate at every queried coordinate, allowing the framework to favor fidelity features in structurally simple regions and perceptual features in texture-rich areas.
- FPLIA achieves Pareto-optimal fidelity-perception balance on standard ASISR benchmarks, improving LPIPS by up to 29.6% over the best regression-based methods while maintaining a PSNR trade-off within 2.6%, and simultaneously improving both metrics over the leading diffusion-based methods, with less than 5% latency and 2% memory overhead.

2 Preliminary

2.1 Local Implicit Image Functions

ASISR seeks to reconstruct a high-resolution image from a low-resolution input at any upsampling factor within a single model. A widely adopted paradigm for this task represents images as continuous functions defined over a 2D coordinate space $\mathcal{X}$ [3–5, 12, 14]. Given a low-resolution image, a deep feature extractor first produces a feature map. For each queried high-resolution coordinate $x_q \in \mathcal{X}$, the framework identifies the nearest feature positions on the low-resolution grid and retrieves the corresponding latent vectors. These local features are concatenated with the cell size c that encodes the spatial extent of a pixel at the target resolution. A coordinate-conditioned decoder, typically implemented as a multi-layer perceptron, then maps this concatenated representation to an RGB value at x_q. Because the decoder operates on continuous coordinates rather than fixed pixel locations, a single trained model can serve arbitrary upsampling factors.

2.2 Cross-Scale Local Attention

To move beyond uniform feature aggregation, recent extensions of the local implicit paradigm replace the simple concatenation scheme described above with an attention-based mechanism that computes content-dependent weightings over the local neighborhood of each queried coordinate [3, 4]. The feature map is first projected into query, key, and value embeddings through convolutional layers. The query embedding $q \in \mathbb{R}^{1\times C}$ at a target coordinate x_q is obtained via bilinear interpolation, while key and value embeddings $k, v \in \mathbb{R}^{G_h G_w \times C}$ are sampled over a local grid $x^{G_h \times G_w}$ of size $G_h \times G_w$ centered at x_q. To encode

the geometric relationship between the query and each grid position, the relative offset $\delta x_{ij} = x_q - x_{ij}$ is passed through a positional encoding function γ followed by a fully connected layer to yield a positional bias $B = \mathrm{FC}(\gamma(\delta x))$. An attention map $a = \mathrm{Softmax}(qk^\top/\sqrt{C} + B)$ is then formed, and the attended feature is computed as $\tilde{f} = \mathrm{FC}(\mathrm{Concat}(a \times v,\ c))$, where the cell size c is appended before the final projection so that the aggregation adapts to the target resolution. This mechanism allows the decoder to focus on the most informative neighbors for each query and to modulate its response according to the upscaling factor.

3 Methodology

3.1 Problem Formulation

Given a low-resolution image $I^{LR} \in \mathbb{R}^{H \times W \times 3}$, the goal is to reconstruct a high-resolution image $I^{HR} \in \mathbb{R}^{r_h H \times r_w W \times 3}$ under an arbitrary upsampling factor $r = \{r_h, r_w\}$. Following the local implicit paradigm reviewed in Section 2.1, reconstruction proceeds by predicting the RGB value at each queried coordinate x_q in the continuous space $\mathcal{X}$. The framework has access to two complementary feature sources: fidelity-oriented features $\mathcal{F}_f \in \mathbb{R}^{H \times W \times C}$, extracted by a pixel-regression backbone, and perceptual features $\mathcal{F}_p \in \mathbb{R}^{H \times W \times C}$, produced by a diffusion-based generator. The two streams encode fundamentally different information: $\mathcal{F}_f$ preserves pixel-level accuracy at the cost of high-frequency detail, while $\mathcal{F}_p$ recovers perceptually convincing textures at the risk of structural hallucination. A direct combination of the two through fixed blending or simple concatenation treats all spatial locations and scale factors uniformly, an assumption that conflicts with the observation that the relative utility of each feature type varies with both local image content and the upscaling factor. Addressing this challenge is the central motivation for the architectural choices developed in the following subsections.

3.2 Design Analysis

As established in Section 1, the interaction between fidelity-oriented and perceptual features is inherently asymmetric. The concrete consequence for attention-based interaction is directional: when fidelity features attend to perceptual keys, the resulting representation inherits fine-grained detail while remaining spatially grounded; when perceptual features attend to fidelity keys, the result trades some textural richness for structural consistency. These two information flows address different failure modes and should therefore be modeled as separate, directional pathways rather than collapsed into a single symmetric operation.

Remark 1 (Asymmetric cross-feature interaction). Let $\tilde{f}_{fp}$ denote the feature obtained when a fidelity-derived query attends to perceptual keys and values, and let $\tilde{f}_{pf}$ denote the converse. The two cross-attention outputs occupy distinct regions of the fidelity-perception spectrum: $\tilde{f}_{fp}$ supplements a fidelity-anchored representation with perceptual detail, whereas $\tilde{f}_{pf}$ constrains a perception-driven representation with structural fidelity. Neither subsumes the other, and both

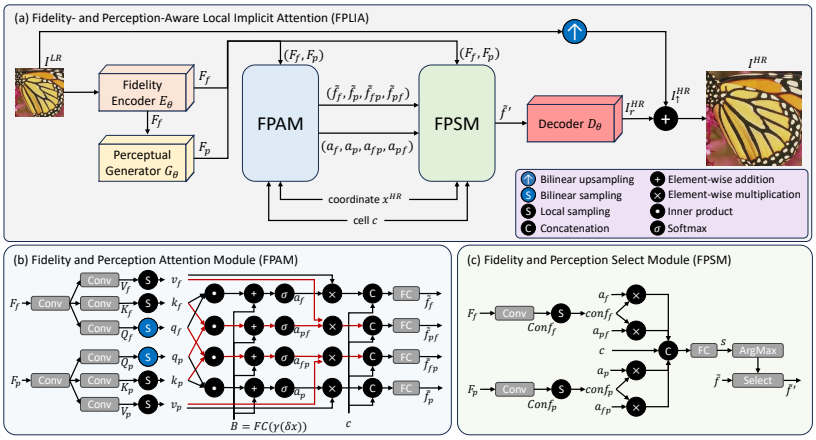

Fig. 2: Overview of the proposed FPLIA framework. (a) FPLIA jointly exploits fidelity-oriented and perceptual features for advancing ASISR. (b) FPAM employs self-attention and bi-directional cross-attention mechanisms to produce enhanced representative latent features. (c) FPSM estimates confidence scores and adaptively selects higher-confidence features for predicting RGB values at each queried coordinate.

complement the self-attention-refined features $\tilde{f}_f$ and $\tilde{f}_p$, which remain within their respective original manifolds.

The four candidate features produced by this architecture each encode a different fidelity-perception balance, and their diversity is valuable only if the framework can commit to the most suitable one at each coordinate rather than averaging across them. Soft blending is insufficient for this purpose: features from different statistical manifolds can produce degenerate outputs when averaged, and a globally learned weighting forfeits the spatial adaptivity the architecture is designed to provide. The optimal candidate varies with both local content and upscaling factor, as confirmed by the analysis in Section 4.4. A discrete, per-pixel selection conditioned on content-derived attention cues and scale-encoding cell size c is therefore the natural design choice.

3.3 Framework Overview

Guided by the analysis above, FPLIA instantiates these principles in a five-component pipeline illustrated in Fig. 2: a fidelity-oriented feature extractor E_θ, a perceptual feature generator G_θ conditioned on $\mathcal{F}_f$, FPAM, FPSM, and a decoder D_θ that maps the selected feature to a residual RGB prediction added to a bilinearly upsampled I^{LR}. Both E_θ and G_θ are modular plug-ins; Sections 3.4 and 3.5 detail FPAM and FPSM respectively.

3.4 Fidelity and Perception Attention Module

To realize the asymmetric bidirectional interaction established in Remark 1, FPAM extends the single-stream cross-scale local attention reviewed in Section 2.2 to a dual-stream architecture that operates jointly over $\mathcal{F}_f$ and $\mathcal{F}_p$. Both feature

maps are projected through convolutional layers into query, key, and value embeddings (Q_f, K_f, V_f) and (Q_p, K_p, V_p). For a queried coordinate $x_q \in x^{HR}$, the query embeddings $q_f, q_p \in \mathbb{R}^{1\times C}$ are bilinearly interpolated from Q_f and Q_p at x_q, while the local key and value embeddings (k_f, v_f) and $(k_p, v_p) \in \mathbb{R}^{G_h G_w \times C}$ are sampled from their respective maps over a local grid $x^{G_h \times G_w}$ centered at x_q. The relative positional offset $\delta x_{ij} = x_q - x_{ij}$ for each grid position x_{ij} and the cell size $c = (2/r_h, 2/r_w)$ are shared across all four attention pathways. Applying the CSLA$(\cdot)$ function defined in Section 2.2 in four configurations yields:

$$\begin{aligned} \tilde{f}_f,\ a_f &= \mathrm{CSLA}(q_f, k_f, v_f, \delta x, c), & \tilde{f}_p,\ a_p &= \mathrm{CSLA}(q_p, k_p, v_p, \delta x, c), \\ \tilde{f}_{pf},\ a_{pf} &= \mathrm{CSLA}(q_p, k_f, v_f, \delta x, c), & \tilde{f}_{fp},\ a_{fp} &= \mathrm{CSLA}(q_f, k_p, v_p, \delta x, c), \end{aligned} \tag{1}$$

where $\tilde{f}_f, \tilde{f}_p \in \mathbb{R}^{1\times C}$ are the self-attention outputs that refine each stream within its own statistical manifold, and $\tilde{f}_{pf}, \tilde{f}_{fp} \in \mathbb{R}^{1\times C}$ are the cross-attention outputs that produce the asymmetric hybrid features described in Remark 1. The corresponding attention maps $a_{(\cdot)} \in \mathbb{R}^{1\times G_h G_w}$ are retained for use in the downstream selection stage.

3.5 Fidelity and Perception Select Module

Following the content- and scale-dependent selection rationale developed in Section 3.2, FPSM identifies the most reliable candidate feature at each queried coordinate x_q by estimating a confidence score for each of the four outputs produced by FPAM. To capture the reliability of each feature stream at a global spatial level, FPSM first predicts dense confidence maps $Conf_f$ and $Conf_p$ from the enhanced feature maps $\mathcal{F}_f$ and $\mathcal{F}_p$ through convolutional layers, and samples local confidence values $conf_f$ and $conf_p$ on the grid around x_q. These confidence values are combined with the attention maps a_f, a_p, a_{pf}, a_{fp} from FPAM, which encode the local content structure around the query, and with the cell size c, which encodes the target scale. The per-candidate confidence scores $s = [s_f, s_p, s_{pf}, s_{fp}]$ are computed as:

$$s = \mathrm{FC}(\mathrm{Concat}(a_f \times conf_f,\ a_p \times conf_p,\ a_{pf} \times conf_f,\ a_{fp} \times conf_p,\ c)), \tag{2}$$

where the element-wise products $a_{(\cdot)} \times conf_{(\cdot)}$ fuse content-level attention cues with stream-level reliability estimates. Conditioning on both signals enables FPSM to adapt its selection along the spatial and scale axes analyzed in Section 3.2.

Given the confidence scores s, FPSM selects the highest-confidence candidate through an arg max operation. Because discrete selection is non-differentiable, the Gumbel-Softmax relaxation [11] is adopted during training to enable gradient propagation. A soft one-hot vector is formed as $s' = \mathrm{Softmax}((s + \Delta)/\tau)$, where $\Delta \sim \mathrm{Gumbel}(0, 1)$ is stochastic noise and $\tau > 0$ is a temperature that is initialized to 1 and annealed to 0.001 during training. The selected feature $\tilde{f}'$ at x_q is then obtained as a weighted sum during training and as a hard selection at inference:

$$\tilde{f}' = \sum_{i\in\{f,p,pf,fp\}} s'_i \times \tilde{f}_i \quad \text{(training)}, \qquad \tilde{f}' = \tilde{f}_{j^\star},\ j^\star = \arg\max_j s_j \quad \text{(inference)}. \tag{3}$$

Temperature annealing gradually sharpens the soft distribution toward this hard regime, ensuring a smooth transition between training and inference behavior.

3.6 Training Objectives

The decoder D_θ is applied not only to the selected feature $\tilde{f}'$ but also independently to each of the four candidate features, producing five predictions at every sampled coordinate: the selected output I_k^{HR}, the fidelity output $I_k^{HR_f}$, the perceptual output $I_k^{HR_p}$, the perception-to-fidelity output $I_k^{HR_{pf}}$, and the fidelity-to-perception output $I_k^{HR_{fp}}$. Supervising all five branches ensures that each candidate is trained to be independently informative, which in turn provides FPSM with a meaningful selection landscape in which confidence scores reflect genuine differences in reconstruction quality rather than artifacts of unbalanced training. For each sampled pixel $k \in \{1, \ldots, n\}$, the per-branch losses are:

$$\begin{aligned} L_{select} &= \tfrac{1}{n}\sum\nolimits_k (I_k^{HR} - I_k^{GT})^2, \qquad L_f = \tfrac{1}{n}\sum\nolimits_k (I_k^{HR_f} - I_k^{GT})^2, \\ L_i &= \tfrac{1}{n}\sum\nolimits_k \left(\tfrac{\mu}{\sigma}(I_k^{HR_i} - I_k^{GT})\right)^2, \quad i \in \{p,\, pf,\, fp\}, \end{aligned} \tag{4}$$

where the μ/σ weighting on the perceptual-stream losses follows the one-step diffusion approximation of Kim *et al.* [13], accounting for the signal-to-noise ratio of the diffusion process. The overall objective is the unweighted sum $L = L_{select} + L_f + L_p + L_{pf} + L_{fp}$. Training details including optimizer, learning rate schedule, and data augmentation are provided in the supplementary material.

4 Experiments

4.1 Experimental Setup

The framework is trained on DIV2K [1] and Flickr2K [21], which together provide 3,450 images at 2K resolution. Evaluation is conducted on the DIV2K validation set [1], Set5 [2], Set14 [28], B100 [18], and Urban100 [10] across integer upscaling factors from $\times 4$ to $\times 16$; non-integer factors are additionally examined to validate continuous-scale generalization. Reconstruction quality is measured by PSNR for pixel-level fidelity and LPIPS [29] for perceptual similarity, directly reflecting the two axes of the trade-off that motivates FPLIA. Throughout all tables, the best and second-best results are highlighted in **bold** and underline, respectively.

FPLIA is compared against seven representative ASISR methods spanning both paradigms. The regression-based group includes LIIF [5], LTE [14], LIT [4], CiaoSR [3], and HIIF [12], all equipped with SwinIR [15] backbones, representing the state of the art in fidelity-oriented ASISR. The diffusion-based group includes IDM [7] and Kim *et al.* [13]; the latter is the most recent diffusion-based ASISR method and serves as the primary competitor. This selection covers the full fidelity-perception spectrum and enables direct assessment of how FPLIA bridges the two paradigms. For LIIF, LTE, CiaoSR, and HIIF, results are obtained using official pre-trained checkpoints. For LIT, IDM, and Kim *et al.*, models are retrained using official source code, as pre-trained checkpoints are not available.

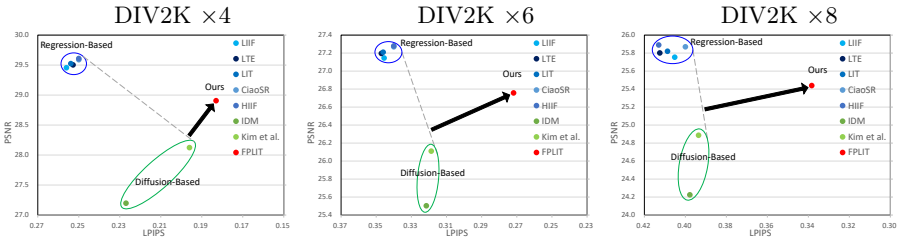

Fig. 3: Fidelity-oriented metric PSNR versus perceptual metric LPIPS across ASISR methods on DIV2K validation set [1] and Set5 [2] at wide-range upscaling factors. Our method (red circle) achieve the **Pareto frontier** between PSNR and LPIPS compared to regression-based methods (blue circle) and diffusion methods (green circle).

Implementation details. The fidelity-oriented feature extractor E_θ is instantiated with SwinIR [15], and the perceptual feature generator G_θ with Stable Diffusion v1.5 [19]. The decoder D_θ is a five-layer MLP with GELU activation [8]. Both E_θ and G_θ are pretrained following their respective protocols [5, 19] and kept frozen during finetuning; only FPAM, FPSM, and D_θ are updated. The framework is backbone-agnostic, and results with alternative extractor and generator choices are reported in the supplementary material.

4.2 Quantitative Results

The ASISR landscape has been characterized by a persistent dichotomy: regression-based methods cluster in the high-PSNR, low-perceptual-quality region, while diffusion-based methods occupy the complementary quadrant. Fig. 3 visualizes this trade-off on the DIV2K validation set and Set5 across upscaling factors from $\times 4$ to $\times 12$, plotting PSNR against LPIPS for all compared methods. FPLIA consistently occupies the Pareto frontier, achieving a fidelity-perception balance that no existing method reaches from either paradigm alone.

Table 1 reports detailed results on Set14 [28], B100 [18], and Urban100 [10] across upscaling factors from $\times 4$ to $\times 16$. Two complementary findings emerge from this comparison. Against regression-based methods, FPLIA achieves consistently lower LPIPS with only a modest PSNR concession. Among the regression baselines, CiaoSR [3] and HIIF [12] alternate as the strongest competitors, with CiaoSR leading on LPIPS and HIIF on PSNR, yet neither excels on both metrics simultaneously because their architectures rely exclusively on pixel-regression features and lack the perceptual stream necessary for high-frequency synthesis. FPLIA addresses this gap through the cross-attention pathways in FPAM, which enrich fidelity-anchored representations with perceptual diversity. FPLIA improves LPIPS by 29.6% over the best regression-based method on Set14 at $\times 4$, while the overall PSNR trade-off remains within 2.6%.

Against diffusion-based methods, the result is more definitive: FPLIA improves both PSNR and LPIPS simultaneously across the majority of benchmark-scale combinations. Kim *et al.* [13], the strongest diffusion baseline, obtains competitive LPIPS through generative texture synthesis, but this comes at the cost of structural fidelity, with PSNR consistently falling below the leading regression methods. The root cause is the uniform application of generative processing to all

Table 1: Comparison of regression-based and diffusion-based approaches with FPLIA. Average PSNR and LPIPS on Set5 [2], Set14 [28], B100 [18], and Urban100 [10].

Set5		×4		×6		×8		×12		×16	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Regression Methods	LIIF-SwinIR [5]	32.71	0.167	29.44	0.233	27.33	0.288	24.90	0.386	23.26	0.474
	LTE-SwinrIR [14]	32.79	0.169	29.51	0.236	27.32	0.297	25.00	0.413	23.29	0.505
	LIT-SwinrIR [4]	32.82	0.168	29.49	0.235	27.38	0.292	24.88	0.399	23.35	0.483
	CiaoSR-SwinrIR [3]	32.83	0.167	29.57	**0.231**	27.41	**0.280**	25.06	**0.378**	23.43	**0.462**
	HIIF-SwinIR [12]	**32.87**	**0.166**	**29.63**	0.235	**27.48**	0.296	**25.13**	0.412	**23.47**	0.498
Diffusion Methods	IDM [7]	30.11	**0.110**	26.86	**0.170**	24.67	**0.246**	20.01	0.396	18.77	0.506
	Kim *et al.* [13]	**31.07**	0.128	**27.74**	0.196	**25.81**	0.284	**22.57**	**0.362**	**21.50**	**0.434**
FPLIA		**31.95**	**0.116**	**28.73**	**0.171**	**26.76**	**0.238**	**24.24**	**0.322**	**22.83**	**0.404**
vs. Best Regression Method		-2.80%	+30.12%	-3.04%	+25.97%	-2.62%	+15.00%	-3.54%	+14.81%	-2.73%	+12.55%
vs. Best Diffusion Method		+2.83%	-5.45%	+3.57%	-0.59%	+3.68%	+3.25%	+7.40%	+11.05%	+6.19%	+6.91%

Set14		×4		×6		×8		×12		×16	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Regression Methods	LIIF-SwinIR [5]	28.95	0.272	26.82	0.384	25.33	0.431	23.35	0.530	22.20	0.585
	LTE-SwinrIR [14]	29.04	0.269	26.87	0.376	25.41	0.440	**23.45**	0.546	**22.30**	0.607
	LIT-SwinrIR [4]	29.00	0.269	26.84	0.375	25.38	0.437	23.39	0.537	22.27	0.595
	CiaoSR-SwinrIR [3]	29.08	0.269	26.87	**0.371**	25.43	**0.430**	23.44	**0.521**	22.28	**0.576**
	HIIF-SwinIR [12]	**29.10**	**0.267**	**26.98**	0.374	**25.49**	0.440	23.43	0.542	22.29	0.605
Diffusion Methods	IDM [7]	27.13	0.221	25.11	**0.299**	23.60	**0.370**	20.74	0.485	19.03	0.573
	Kim*et al.* [13]	**27.77**	**0.201**	**25.91**	0.325	**24.43**	0.389	**22.29**	**0.460**	**20.89**	**0.531**
FPLIA		**28.48**	**0.188**	**26.50**	**0.283**	**25.01**	**0.350**	**23.01**	**0.439**	**21.74**	**0.493**
vs. Best Regression Method		-2.13%	+29.59%	-1.78%	+23.72%	-1.88%	+18.60%	-1.88%	+15.74%	-2.51%	+14.41%
vs. Best Diffusion Method		+2.56%	+6.47%	+2.28%	+5.35%	+2.37%	+5.41%	+3.23%	+4.57%	+4.07%	+7.16%

B100		×4		×6		×8		×12		×16	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Regression Methods	LIIF-SwinIR [5]	27.82	0.360	26.05	0.466	25.00	0.534	23.63	0.627	22.73	0.683
	LTE-SwinrIR [14]	27.85	0.355	26.08	0.469	25.02	0.545	23.64	0.648	22.77	0.710
	LIT-SwinrIR [4]	27.87	0.358	26.09	0.469	25.04	0.540	23.65	0.635	22.77	0.695
	CiaoSR-SwinrIR [3]	27.90	0.355	26.13	**0.461**	25.06	**0.532**	23.68	**0.623**	22.77	**0.677**
	HIIF-SwinIR [12]	**27.92**	**0.351**	**26.15**	0.467	**25.09**	0.543	**23.69**	0.644	**22.80**	0.701
Diffusion Methods	IDM [7]	26.36	0.279	24.70	**0.352**	23.68	**0.430**	21.85	**0.530**	18.46	0.668
	Kim*et al.* [13]	**26.64**	**0.262**	**25.16**	0.397	**24.18**	0.466	**22.68**	0.549	**21.22**	**0.608**
FPLIA		**27.33**	**0.251**	**25.71**	**0.354**	**24.70**	**0.431**	**23.30**	**0.520**	**22.23**	**0.569**
vs. Best Regression Method		-2.11%	+28.49%	-1.68%	+23.21%	-1.55%	+18.98%	-1.65%	+16.53%	-2.50%	+15.95%
vs. Best Diffusion Method		+2.59%	+4.20%	+2.19%	-0.57%	+2.15%	-0.23%	+2.73%	+1.89%	+4.76%	+6.41%

Urban100		×4		×6		×8		×12		×16	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Regression Methods	LIIF-SwinIR [5]	27.14	0.200	24.59	0.298	23.14	0.375	21.42	0.501	20.41	0.595
	LTE-SwinrIR [14]	27.23	0.194	24.66	0.296	23.18	0.381	21.46	0.520	20.41	0.620
	LIT-SwinrIR [4]	27.21	0.195	24.68	0.295	23.20	0.375	21.51	0.502	20.47	0.600
	CiaoSR-SwinrIR [3]	27.42	**0.187**	**24.86**	**0.281**	**23.34**	**0.357**	21.61	**0.477**	20.54	**0.567**
	HIIF-SwinIR [12]	**27.44**	0.188	24.81	0.302	23.32	0.385	**21.62**	0.516	**20.56**	0.612
Diffusion Methods	IDM [7]	25.13	0.208	23.15	0.324	22.01	0.409	20.53	0.523	19.57	0.610
	Kim*et al.* [13]	**26.42**	**0.170**	**24.07**	**0.284**	**22.74**	**0.356**	**21.05**	**0.466**	**20.03**	**0.549**
FPLIA		**26.87**	**0.160**	**24.44**	**0.260**	**23.06**	**0.338**	**21.34**	**0.453**	**20.31**	**0.534**
vs. Best Regression Method		-2.08%	+14.44%	-1.69%	+7.47%	-1.20%	+5.32%	-1.30%	+5.03%	-1.22%	+5.82%
vs. Best Diffusion Method		+1.70%	+5.88%	+1.54%	+8.45%	+1.41%	+5.06%	+1.38%	+2.79%	+1.40%	+2.73%

spatial locations, including regions where pixel-regression features already suffice. FPSM's content-aware selection avoids this failure mode by restricting perceptual candidates to regions where high-frequency synthesis is genuinely needed.

4.3 Qualitative Results

Fig. 4 presents visual comparisons against all seven baselines. Two capabilities of FPLIA stand out when examined alongside the failure modes of existing methods. For content with repetitive structural patterns, such as the striped texture in the first example, regression-based methods produce incorrect orientations and blurred reconstructions because their pixel-regression features lack the high-frequency diversity needed to resolve ambiguous periodic structures. The diffusion-based method of Kim *et al.* [13] introduces noisy artifacts in the same region,

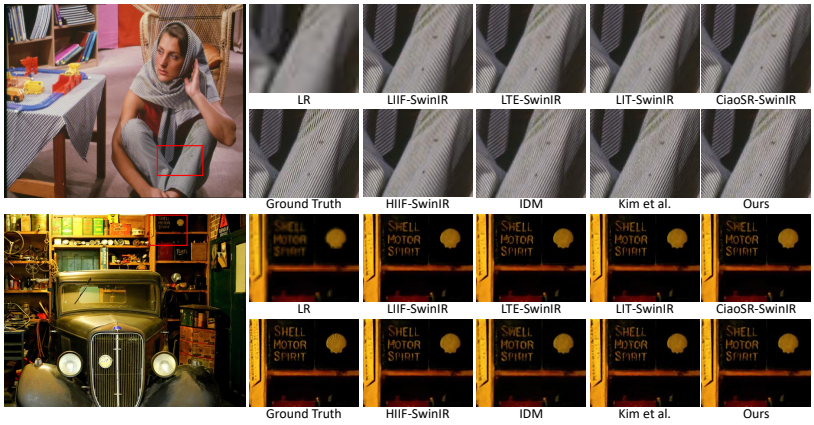

Fig. 4: Comparison of the qualitative results of LIIF [5], LTE [14], LIT [4], CiaoSR [3], HIIF [12], IDM [7], Kim *et al.* [13], and our proposed FPLIA.

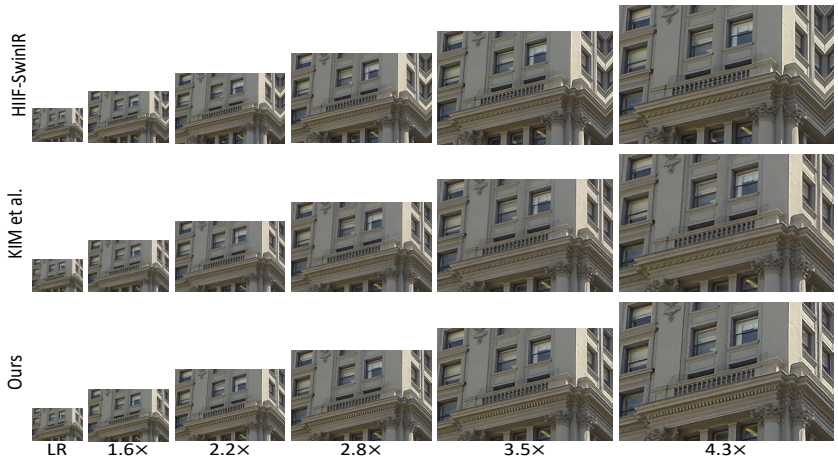

Fig. 5: Comparison of the qualitative results of HIIF [12], Kim *et al.* [13], and our proposed FPLIA with non-integer upscaling factors.

as its generative process hallucinates textures without structural anchoring. FPLIA recovers the correct pattern with sharp detail, a result attributable to the bidirectional cross-attention in FPAM, which enables the fidelity stream to anchor the structural layout while the perceptual stream supplies high-frequency sharpness. For content requiring semantic consistency, such as the text in the second example, diffusion methods exhibit a different failure mode: IDM [7] hallucinates spurious characters, and Kim *et al.* produce discontinuous letter forms. FPLIA avoids these artifacts through FPSM's content-aware selection, which favors fidelity-anchored candidates in semantically sensitive regions.

Fig. 5 further validates the continuous-scale capability central to ASISR. On a building image upscaled through non-integer factors from $\times 1.6$ to $\times 4.3$, HIIF [12] and Kim *et al.* [13] fail to faithfully reconstruct the architectural patterns, whereas FPLIA maintains consistent sharpness and structural correctness across the entire

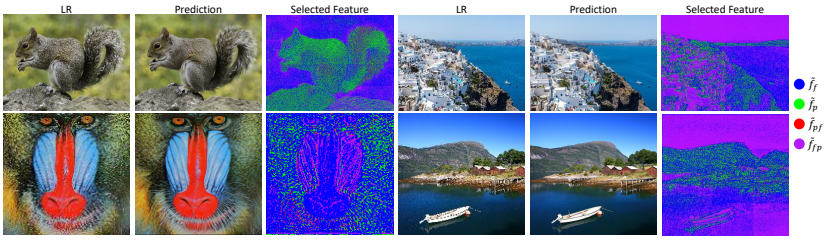

Fig. 6: Visualization of the distribution of various selected features on a per-pixel basis, including fidelity feature $\tilde{f}_f$ (●), perceptual feature $\tilde{f}_p$ (●), perception-to-fidelity feature $\tilde{f}_{pf}$ (●), and fidelity-to-perception feature $\tilde{f}_{fp}$ (●).

Table 2: Percentage of various selected features on DIV2K validation set [1], including the fidelity feature $\tilde{f}_f$, perceptual feature $\tilde{f}_p$, perception-to-fidelity feature $\tilde{f}_{pf}$, and fidelity-to-perception feature $\tilde{f}_{fp}$.

Feature Branch	×2	×4	×6	×8	×12	×16
$\tilde{f}_f$	47.35%	37.34%	34.02%	32.27%	32.60%	33.49%
$\tilde{f}_p$	28.92%	19.09%	16.96%	16.97%	17.05%	17.53%
$\tilde{f}_{pf}$	1.21%	2.42%	2.91%	3.29%	3.64%	3.87%
$\tilde{f}_{fp}$	22.52%	41.15%	46.11%	47.48%	46.71%	45.11%

range. The coordinate-based local implicit formulation naturally accommodates arbitrary scales, and the adaptive selection performed by FPSM ensures that the fidelity-perception balance remains appropriate at each factor.

4.4 Analysis of Adaptive Feature Selection

This subsection examines the internal selection behavior of FPSM. Fig. 6 visualizes the per-pixel feature selections on representative images. Along the spatial axis, FPSM consistently assigns the fidelity feature $\tilde{f}_f$ and the fidelity-to-perception feature $\tilde{f}_{fp}$ to smooth and background regions, while the perceptual feature $\tilde{f}_p$ is preferentially selected in areas with rich high-frequency textures. The selections remain spatially coherent across similar textures, avoiding the discontinuity artifacts that would arise from noisy or inconsistent confidence estimates.

Along the scale axis, Table 2 reveals a complementary trend in the global selection statistics. The proportion of coordinates assigned to $\tilde{f}_f$ decreases as the upscaling factor grows, while the proportion assigned to $\tilde{f}_{fp}$ increases. This shift demonstrates that as fidelity features become less informative at higher scales, FPSM increasingly relies on candidates that incorporate perceptual diversity. The perceptual feature $\tilde{f}_p$ follows a similar increasing trend, consistent with the greater reliance on generative modeling at large magnifications. The perception-to-fidelity feature $\tilde{f}_{pf}$ is selected at consistently low rates across all scales, aligning with the weaker independent performance of that branch observed in Table 4.

Table 3: Average PSNR and LPIPS on DIV2K validation set [1] of FPLIA with various component configurations.

FPAM	FPSM	× 4		× 6		× 8	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
-	-	28.40	0.231	26.45	0.320	25.15	0.383
✓	-	28.53	0.188	26.55	0.304	25.28	0.371
-	✓	28.60	0.211	26.53	0.302	25.24	0.366
✓	✓	**28.91**	**0.183**	**26.76**	**0.272**	**25.44**	**0.338**

Table 4: Average PSNR and LPIPS on DIV2K validation set [1] of FPLIA using various combinations of the fidelity feature $\tilde{f}_f$, perceptual feature $\tilde{f}_p$, perception-to-fidelity feature $\tilde{f}_{pf}$, and fidelity-to-perception feature $\tilde{f}_{fp}$ in FPAM.

Feature Branch	× 4		× 6		× 8	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
$\tilde{f}_f$	**29.51**	0.258	**27.22**	0.351	**25.84**	0.414
$\tilde{f}_p$	28.12	0.194	26.11	0.310	24.89	0.380
$\tilde{f}_{pf}$	27.52	0.215	25.89	0.338	24.68	0.401
$\tilde{f}_{fp}$	28.28	<u>0.186</u>	26.31	<u>0.289</u>	25.03	<u>0.352</u>
$\tilde{f}_f + \tilde{f}_p$	28.71	0.206	26.61	0.299	25.31	0.361
$\tilde{f}_f + \tilde{f}_p + \tilde{f}_{pf} + \tilde{f}_{fp}$	<u>28.91</u>	**0.183**	<u>26.76</u>	**0.272**	<u>25.44</u>	**0.338**

4.5 Ablation Studies

All ablation experiments are conducted on the DIV2K validation set [1] at upscaling factors ×4, ×6, and ×8. Table 3 isolates the contributions of FPAM and FPSM by replacing each with its baseline counterpart: FPAM is substituted with the LIIF [5] feature extraction pipeline, and FPSM with standard feature concatenation. Removing FPAM while retaining FPSM yields improved PSNR relative to the baseline without either module, yet LPIPS remains noticeably worse than the full model, confirming that the cross-feature interaction provided by FPAM is essential for generating candidates with sufficient perceptual diversity. Conversely, removing FPSM while retaining FPAM improves LPIPS over the no-module baseline but falls short of the full configuration, particularly at higher scales. This asymmetry is consistent with the design rationale in Section 3.2. The full model outperforms both partial configurations on every metric and scale, and the combined gain exceeds the sum of the individual improvements, indicating that the two modules operate synergistically rather than independently.

Table 4 further examines the role of each feature branch within FPAM. Using only the self-attention fidelity feature $\tilde{f}_f$ produces the highest PSNR across all scales but also the worst LPIPS, confirming that pixel-regression features alone carry no perceptual information. Adding the self-attention perceptual feature $\tilde{f}_p$ alongside $\tilde{f}_f$ yields only marginal LPIPS improvement, because the two self-attention outputs reside on separate manifolds and their naive combination lacks the cross-stream interaction needed to bridge the fidelity-perception gap. The decisive improvement arrives when the cross-attention features $\tilde{f}_{fp}$ and $\tilde{f}_{pf}$ are included: LPIPS drops markedly while PSNR remains competitive. This result validates the asymmetric interaction principle from Section 1: the directional information flow modeled by cross-attention produces candidates that neither self-attention branch can generate alone.

Table 5: Average PSNR and LPIPS on DIV2K validation set [1] of FPLIA using various selection module in FPSM.

Select Mechanism	× 4		× 6		× 8	
	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Concatnation	28.53	0.188	26.55	0.304	25.28	0.371
Spatial Feature Transform [23]	28.45	**0.183**	26.51	0.307	25.23	0.375
Per-pixel Selection	**28.91**	**0.183**	**26.76**	**0.272**	**25.44**	**0.338**
Per-channel Selection	28.55	0.193	26.51	0.302	25.23	0.366

Table 6: Inference latency and GPU memory footprint comparison.

Method	Latency (s)				GPU Memory
	Fidelity Encoder	Perceptual Generator	Implicit Neural Function	Total	Footprint (GB)
Kim *et al.* [13]	-	91.5 (LDM)	2.5 (IND)	94.0	25.1
FPLIA	2.2 (SwinIR)	92.1 (LDM)	4.1 (FPAM+FPSM+Decoder)	98.4	25.5

Table 5 compares four mechanisms for combining the candidate features before decoding: direct concatenation, Spatial Feature Transform (SFT) [23] modulation, per-pixel selection, and per-channel selection. Concatenation performs worst overall, confirming the analysis in Section 3.2 that linear fusion of features from different statistical manifolds is inherently suboptimal. SFT improves LPIPS at $\times 4$ by introducing spatially varying modulation weights, yet this soft blending still averages across candidates and fails to maintain the advantage at higher scales where the performance gap between candidates widens. Per-channel selection also underperforms per-pixel selection, because the fidelity-perception trade-off varies spatially rather than along the channel dimension. Per-pixel selection achieves the best results at every scale on both metrics, validating the discrete selection rationale developed in Section 3.2.

The performance gains from FPAM and FPSM come at minimal computational cost. Table 6 reports inference latency and GPU memory averaged over 100 runs of a $128 \times 128 \rightarrow 512 \times 512$ upscaling task on an NVIDIA A6000 GPU. Relative to the diffusion baseline of Kim *et al.* [13], FPLIA adds less than 5% latency and less than 2% peak memory. The overhead is small because FPAM operates on local coordinate neighborhoods rather than global feature maps, and FPSM consists of lightweight convolutional and fully connected layers. These figures confirm that the Pareto-frontier performance is achieved without meaningful efficiency penalty, closing the loop with the efficiency claim in Section 1.

5 Conclusion

This paper presents FPLIA, a novel framework that synergistically integrates fidelity-oriented features with a diffusion model to advance ASISR. FPLIA incorporates two architectural innovations including FPAM and FPSM to leverage both fidelity-oriented and perceptual features for realistic and faithful reconstructions. Extensive quantitative and qualitative evaluations across standard ASISR benchmarks with widely integer and non-integer upscaling factors substantiate the efficacy of these innovations. The results show that FPLIA achieves superior perceptual quality and lies on the Pareto frontier of perceptual quality versus reconstruction fidelity compared with state-of-the-art methods. Comprehensive ablation analyses further validate the contribution of each proposed component.

Acknowledgements

The authors gratefully acknowledge the support from the National Science and Technology Council (NSTC) in Taiwan under grant numbers NSTC 114-2221-E-002-069-MY3, NSTC 113-2221-E-002-212-MY3, and NSTC 115-2218-E-A49-010, as well as the support from the Academia Sinica Scholar Award (ASSA) under grant number AS-ASSA-115-02, NTU Artificial Intelligence Center of Research Excellence, and Taiwan Centers of Excellence in Artificial Intelligence. This research was also supported by Mediatek Inc. The authors would also like to express their appreciation for the hardware grant donation of the GPUs from NVIDIA Academic Grant Program and NVIDIA AI Technology Center (NVAITC) used in this work. Furthermore, the authors extend their gratitude to the National Center for High-Performance Computing (NCHC) for providing computational and storage resources.

References

1. Agustsson, E., Timofte, R.: NTIRE 2017 challenge on single image super-resolution: Dataset and study. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW). pp. 1122–1131 (2017)
2. Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.: Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In: Proc. British Machine Vision Conf. (BMVC). pp. 1–10 (2012)
3. Cao, J., Wang, Q., Xian, Y., Li, Y., Ni, B., Pi, Z., Zhang, K., Zhang, Y., Timofte, R., Van Gool, L.: Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1796–1807 (2023)
4. Chen, H.W., Xu, Y.S., Hong, M.F., Tsai, Y.M., Kuo, H.K., Lee, C.Y.: Cascaded local implicit transformer for arbitrary-scale super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 18257–18267 (2023)
5. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 8628–8638 (2021)
6. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)**38**(2), 295–307 (2016)
7. Gao, S., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Gao, B.Z., Liu, X., Zeng, B., Xu, S., Li, Y., Luo, X., Liu, J., Zhen, X., Zhang, B.: Implicit diffusion models for continuous super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10021–10030 (2023)
8. Hendrycks, D., Gimpel, K.: Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR **abs/1606.08415** (2016)
9. Hu, X., Mu, H., Zhang, X., Wang, Z., Tan, T., Sun, J.: Meta-sr: A magnification-arbitrary network for super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1575–1584 (2019)
10. Huang, J., Singh, A., Ahuja, N.: Single image super-resolution from transformed self-exemplars. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 5197–5206 (2015)

11. Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: Proc. Int. Conf. on Learning Representations (ICLR)(2017)
12. Jiang, Y., Kwan, H.M., Peng, T., Gao, G., Zhang, F., Zhu, X., Sole, J., Bull, D.: HIIF: hierarchical encoding based implicit image function for continuous super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2289–2299 (2025)
13. Kim, J., Kim, T.: Arbitrary-scale image generation and upsampling using latent diffusion model and implicit neural decoder. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 9202–9211 (2024)
14. Lee, J., Jin, K.H.: Local texture estimator for implicit representation function. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1929–1938 (2022)
15. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proc. IEEE Int. Conf. on Computer Vision Workshop (ICCVW). pp. 1833–1844 (2021)
16. Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M.: Enhanced deep residual networks for single image super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW). pp. 1132–1140 (2017)
17. Liu, Y., Guo, Y., Zhang, S.: Enhancing multi-scale implicit learning in image super-resolution with integrated positional encoding. CoRR **abs/2112.05756** (2021)
18. Martin, D.R., Fowlkes, C.C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. IEEE Int. Conf. on Computer Vision (ICCV). pp. 416–425 (2001)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685 (2022)
20. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI)**45**(4), 4713–4726 (2023)
21. Timofte, R., Agustsson, E., Gool, L.V., Yang, M., Zhang, L.: NTIRE 2017 challenge on single image super-resolution: Methods and results. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW). pp. 1110–1121 (2017)
22. Wang, X., Chen, X., Ni, B., Wang, H., Tong, Z., Liu, YutianWang, X., Chen, X., Ni, B., Wang, H., Tong, Z., Liu, Y.: Deep arbitrary-scale image super-resolution via scale-equivariance pursuit. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1786–1795 (2023)
23. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 606–615 (2018)
24. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.C.: Esrgan: Enhanced super-resolution generative adversarial networks. In: Proc. European Conf. on Computer Vision Workshop (ECCVW). pp. 63–79 (2018)
25. Xu, X., Wang, Z., Shi, H.: Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution. CoRR **abs/2103.12716** (2021)
26. Yang, J., Shen, S., Yue, H., Li, K.: Implicit transformer network for screen content image continuous super-resolution. In: Proc. Conf. on Neural Information Processing Systems (NeurIPS). pp. 13304–13315 (2021)

27. Yao, J., Tsao, L., Lo, Y., Tseng, R., Chang, C., Lee, C.: Local implicit normalizing flow for arbitrary-scale image super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 1776–1785 (2023)
28. Zeyde, R., Elad, M., Protter, M.: On single image scale-up using sparse-representations. In: Curves and Surfaces. Lecture Notes in Computer Science, vol. 6920, pp. 711–730 (2010)
29. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
30. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proc. European Conf. on Computer Vision (ECCV). pp. 294–310 (2018)
31. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2472–2481 (2018)