BI-IISR: BINARIZED ONE-STEP DIFFUSION FOR INFRARED IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based infrared image super-resolution (IISR) has shown strong capability in recovering fine structures and improving perceptual quality under real-world degradations. However, its heavy computational cost and memory footprint severely limit practical deployment on resource-constrained devices. To address these issues, we propose Bi-IISR, a binarized one-step diffusion framework for robust and structurally faithful infrared super-resolution under stringent efficiency constraints. First, we introduce a Degradation-Responsive Conditioning (DRC) module on the low-resolution conditioning branch to perform degradation rectification and structural compensation, thereby suppressing distortion amplification induced by binarized processing. Second, we incorporate the Shift-Gain Regulation Mechanism (SGRM) into the binarization procedure to prevent sign collapse caused by biased channel distributions and to counteract energy attenuation, stabilizing inter-layer propagation. Finally, we adopt a Multi-level Attention Distillation (MAD) loss that aligns intermediate representations of the binarized student with full-precision teacher, which stabilizes optimization and improves structural fidelity. Extensive experiments demonstrate that Bi-IISR substantially enhances robustness and boundary preservation for binarized IISR under real degradations, while markedly reducing computation and memory consumption.

1 INTRODUCTION

Infrared imaging has shown strong robustness in challenging environments, such as low-light, foggy, and nighttime conditions, making it an important sensing modality for a wide range of vision tasks (Liu et al., 2020a; Wang et al., 2022; 2023). As a result, it has been widely deployed in applications including autonomous driving (Cao et al., 2023; Sun et al., 2024), robotics, and intelligent surveillance. Nevertheless, due to the physical imaging mechanism and high manufacturing cost of infrared sensors, infrared images are often captured at limited spatial resolution and are frequently affected by complex degradations, such as blur, noise contamination, and structured artifacts. These factors weaken thermal target contours and structural boundaries, limiting the practical utility of infrared imagery for downstream visual tasks.

Infrared image super-resolution (IISR) reconstructs high-resolution images from low-resolution infrared inputs to improve structural clarity (Wang et al., 2020; Yang et al., 2019; Yao et al., 2020). Diffusion-based super-resolution methods use strong priors to recover fine details (Cui & Harada, 2024; Yu et al., 2025), as demonstrated by ResShift (Yue et al., 2023) and StableSR (Wang et al., 2024a). This advantage is particularly valuable for infrared image restoration, where preserving target contours and structural boundaries is of central importance in downstream perception. However, their computational and memory costs limit use in resource-constrained scenarios. As illustrated in Figure 1, ordinary infrared diffusion relies on degraded low-resolution guidance and a heavy full-precision core, whereas the proposed framework enhances the conditioning quality and performs efficient reconstruction with a 1-bit computational core.

Aggressive quantization is one way to improve efficiency (Zhang et al., 2020). Among different options, 1-bit quantization (*i.e.*, binarization) offers the largest savings in memory and arithmetic cost (Liu et al., 2025b), since floating-point operations can be replaced by bitwise computations. However, applying binarization to diffusion-based super-resolution is not straightforward. Diffusion reconstruction depends on fine-grained and continuous feature transformations, whereas binarization

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

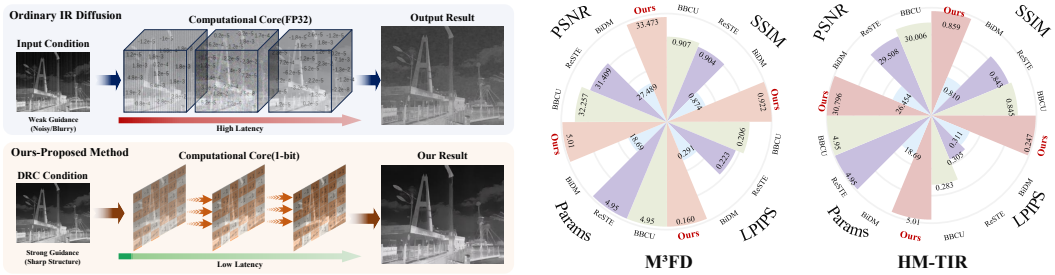


Figure 1: Motivation and quality-efficiency overview. Left: ordinary infrared diffusion processes weak degraded guidance with a full-precision core, whereas Bi-IISR uses DRC-enhanced conditioning and a low-latency 1-bit core. Right: radial plots compare PSNR, SSIM, LPIPS, and parameter counts with selected 1-bit baselines on M^3FD (left plot) and HM-TIR (right plot).

constrains both weights and activations to a binary space. Under such an extreme quantization setting, feature representation and optimization become much more fragile, making diffusion-based restoration especially sensitive to 1-bit constraints.

The problem is even more challenging in infrared image super-resolution. Infrared images are often affected by compound degradations, such as blur, speckle noise, and structured artifacts (Yue et al., 2022). When these degraded inputs are processed by binarized networks, the degradation can be further amplified, which weakens conditioning quality and may blur target boundaries or even damage object contours. In addition, generic binarization is not well matched to infrared feature statistics. Many infrared activations are not naturally centered around zero, so sign-based projection can easily produce degenerated binary responses (Liu et al., 2020c). At the same time, binary propagation removes magnitude information and tends to weaken feature responses, which may lead to layer-wise scale drift and unstable intermediate representations, and eventually impair structural recovery. These issues raise a practical question: can a binarized infrared super-resolution model remain robust to complex degradations, adapt to infrared feature distributions, and maintain stable structural recovery under limited computational budgets?

Motivated by the above observations, we propose **Bi-IISR**, a binarized one-step diffusion framework for infrared image super-resolution. The framework is built around three components. First, we introduce a Degradation-Responsive Conditioning (DRC) module to refine the low-resolution input condition by suppressing directional structured artifacts (stripe noise and fixed-pattern noise) and compensating for blur-related high-frequency loss. Second, we develop a Shift-Gain Regulation Mechanism (SGRM), which adjusts feature distributions before and after binary mapping to make binarized propagation more stable. Third, we use a Multi-level Attention Distillation (MAD) loss to guide the binarized student with a full-precision teacher, so that structurally important responses can be better preserved across different feature levels. The main contributions of this work are summarized as follows:

- We present Bi-IISR, a binarized one-step diffusion framework for infrared image super-resolution. To the best of our knowledge, this is the first attempt to study binarized one-step diffusion in this task.
- We introduce a degradation-aware conditioning strategy that improves the robustness of binarized diffusion reconstruction by suppressing structured interference and compensating for high-frequency loss in low-resolution infrared inputs.
- We propose a shift-gain regulation mechanism for binary propagation, which reduces the mismatch between infrared feature distributions and sign-based quantization, while alleviating response attenuation during propagation.
- We design a multi-level attention distillation objective to stabilize intermediate representations under aggressive binarization and improve structural consistency in reconstructed infrared images, particularly near target boundaries.

2 RELATED WORK

This section reviews infrared and diffusion-based super-resolution methods, followed by binarization methods relevant to Bi-IISR.

2.1 INFRARED AND DIFFUSION-BASED SUPER-RESOLUTION

IISR reconstructs high-resolution infrared images from degraded inputs. Early CNN-based methods (Yang et al., 2018), including ChasNet (Prajapati et al., 2021), improved thermal structure recovery. Adversarial methods (Yan et al., 2021), such as IPRSR (Huang et al., 2021), pursued perceptual quality but could introduce artifacts or unstable structures under complex degradations. CoRPLE (Li et al., 2024) uses cross-modal priors for structural guidance in texture-deficient regions. Real-IISR (Zou et al., 2026) models real-world infrared degradation through autoregressive reconstruction. Diffusion-based IISR has also been explored for complex degradations; DifIISR (Li et al., 2025) incorporates infrared structural priors for relatively stable reconstruction.

Beyond IISR, diffusion-based super-resolution methods use pretrained generative priors (StableSR (Wang et al., 2024a)), semantic guidance (SeeSR (Wu et al., 2024b)), or diffusion priors (DiffIR (Xia et al., 2023)) to improve restoration under complex degradations. ResShift (Yue et al., 2023) reorganizes residual prediction to improve sampling efficiency. Nevertheless, iterative sampling and large generative backbones remain costly (Liu et al., 2021). To reduce sampling cost, SinSR (Wang et al., 2024b) approximates the reverse trajectory, while OSEDiff (Wu et al., 2024a) uses score distillation for one-step reconstruction. S3Diff (Zhang et al., 2026) combines degradation-aware conditioning with lightweight adaptation. PiSA-SR (Sun et al., 2025) separates pixel and semantic branches to balance fidelity and perceptual quality, while TSD-SR (Dong et al., 2025) improves one-step stability through target score distillation. Despite fewer sampling steps, backbones remain typically large and feature modeling complex, leaving efficient structural reconstruction in IISR a challenge for resource-constrained deployment.

2.2 BINARIZATION

Binarization maps weights and activations to ± 1 to enable efficient bitwise convolution (Courbariaux et al., 2016; Rastegari et al., 2016), but the resulting information loss limits representation. Subsequent work improves feature propagation and activation design (Liu et al., 2020b; Qin et al., 2020; Liu et al., 2020c). For low-level vision, BBCU (Xia et al., 2022) and binarized super-resolution methods (Xin et al., 2020) seek to preserve image structure under binary constraints. BiDM (Zheng et al., 2024) and Bi-DiffSR (Chen et al., 2024) extend binarization to diffusion models and diffusion-based super-resolution, respectively. Yet these methods mainly target generic backbones, natural-image restoration, or general diffusion rather than infrared super-resolution. In IISR, binarization can amplify input degradations, infrared feature distributions are less compatible with sign projection, and unstable binary propagation threatens structural recovery. Therefore, a binarized diffusion framework specifically designed for infrared super-resolution is still lacking.

3 METHOD

3.1 OVERALL FRAMEWORK

Given a low-resolution infrared image $\mathbf{I}_{\mathrm{LR}}$, our goal is to reconstruct its high-resolution counterpart $\mathbf{I}_{\mathrm{SR}}$ within an efficient one-step diffusion framework. As shown in Figure 2, Bi-IISR performs reconstruction in the latent space of a pretrained VQ-GAN to reduce inference cost. DRC operates on a separate LR conditioning branch, while the one-step student predicts the teacher's latent output.

The full-precision teacher supervises a student whose main convolutions are binarized. At inference, the teacher is discarded and the student's latent output is decoded by the VQ-GAN.

3.2 DEGRADATION-RESPONSIVE CONDITIONING

Infrared degradations, including stripe noise, fixed-pattern noise (FPN), and blur, can be amplified by binarization and weaken LR conditioning. To address this, **Degradation-Responsive Condi-**

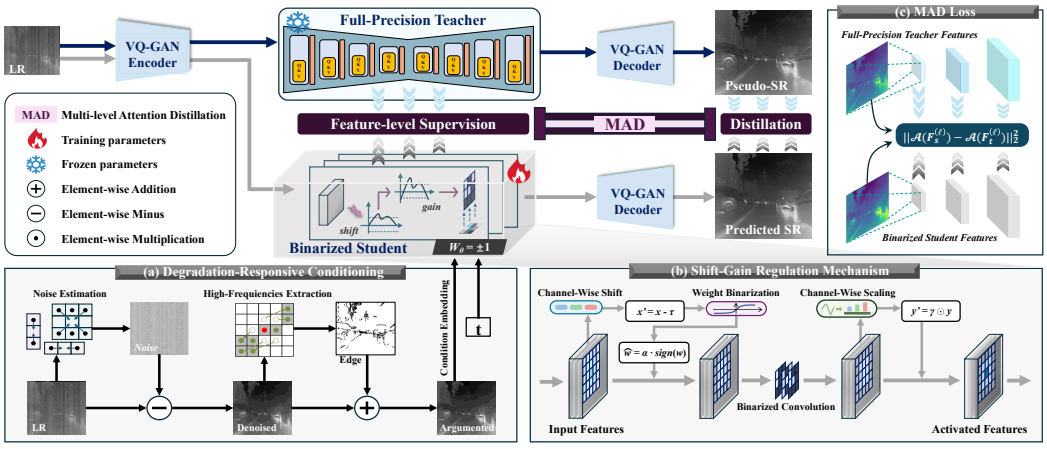


Figure 2: Overview of Bi-IISR. A frozen full-precision teacher guides a trainable 1-bit student for one-step diffusion-based infrared super-resolution in VQ-GAN latent space. (a) DRC refines LR conditioning by suppressing structured artifacts and compensating for high-frequency loss. (b) SGRM shifts activations and rescales binary responses. (c) MAD aligns multi-level teacher and student attention patterns.

tioning (DRC) first suppresses directional artifacts and then compensates for blur-related high-frequency loss to preserve boundary cues.

Let $\mathbf{I}_{\mathrm{LR}} \in \mathbb{R}^{B\times C\times H\times W}$ denote the LR condition and $\tilde{\mathbf{I}}_{\mathrm{LR}} = \mathrm{DRC}(\mathbf{I}_{\mathrm{LR}})$ its refined form. Two depthwise directional operators, $\mathrm{Conv}^{\mathrm{dw}}_{1\times k}(\cdot)$ and $\mathrm{Conv}^{\mathrm{dw}}_{k\times 1}(\cdot)$, extract horizontal and vertical stripe/FPN responses. A depthwise $\mathrm{Conv}^{\mathrm{dw}}_{3\times 3}(\cdot)$ captures local structural cues. Their outputs are concatenated and fused by a 1×1 convolution and sigmoid into a gate map $\mathbf{g} \in (0,1)^{B\times C\times H\times W}$. The gate modulates directional responses to reduce suppression of true boundaries.

Under the gating constraint, the directional responses are summed and element-wise modulated by $\mathbf{g}$ to form a degradation estimate $\widehat{\mathbf{n}}$, which is removed from the input in a residual form:

$$\mathbf{I}_{\mathrm{clean}} = \mathbf{I}_{\mathrm{LR}} - \alpha_{\mathrm{b}}\widehat{\mathbf{n}}, \tag{1}$$

where the learnable rectification strength α_{b} is initialized to zero, so that $\mathbf{I}_{\mathrm{clean}} \approx \mathbf{I}_{\mathrm{LR}}$ initially.

Blur and artifact suppression can weaken high-frequency boundary cues. To avoid amplifying interference, compensation acts on $\mathbf{I}_{\mathrm{clean}}$ rather than the raw input. A depthwise high-pass operator $\mathcal{H}(\cdot)$, implemented as a 3×3 convolution initialized with a Laplacian kernel, extracts responses for residual addition to the cleaned LR condition:

$$\tilde{\mathbf{I}}_{\mathrm{LR}} = \mathbf{I}_{\mathrm{clean}} + \alpha_{\mathrm{g}}\mathcal{H}(\mathbf{I}_{\mathrm{clean}}), \tag{2}$$

where the learnable compensation strength α_{g} is initialized to zero, preserving the near-identity mapping at the start of training.

3.3 Shift-Gain Regulation Mechanism

Infrared activations biased away from zero can collapse under sign projection, while binary convolution attenuates response magnitude. The **Shift-Gain Regulation Mechanism (SGRM)** addresses both failures with a channel-wise shift before activation binarization and a gain after binary convolution to restore response magnitude.

Let $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$ denote an input feature map. Before binary projection, SGRM introduces a learnable channel-wise shift parameter $\boldsymbol{\tau} \in \mathbb{R}^{C}$, which is broadcast along the spatial dimensions to re-center the activation distribution:

$$\mathbf{x}' = \mathbf{x} - \boldsymbol{\tau}. \tag{3}$$

This re-centers activations toward zero without changing their spatial layout, thereby reducing the risk of collapsed binary responses under sign projection.

The adjusted feature $\mathbf{x}'$ is then projected into the binary space using a straight-through estimator (STE)-based sign function:

$$\hat{\mathbf{x}} = \mathrm{sign}(\mathbf{x}'), \qquad \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}'} \approx \mathbf{1}(|\mathbf{x}'| \leq 1), \tag{4}$$

where the indicator function allows gradient propagation within a bounded interval. For weight binarization, we adopt a channel-wise scaled binary form,

$$\hat{\mathbf{W}} = \alpha \, \mathrm{sign}(\mathbf{W}), \qquad \alpha = \mathbb{E}[|\mathbf{W}|], \tag{5}$$

where α is computed per output channel to preserve response scale.

After binary convolution, the output response $\mathbf{y}$ retains sign information but usually suffers from reduced magnitude expressiveness. Such attenuation is harmful in infrared super-resolution, because boundary-related and structurally important responses can gradually weaken as features propagate through multiple layers. To compensate for this effect, SGRM further introduces a learnable channel-wise gain parameter $\boldsymbol{\gamma}$ to rescale the response:

$$\mathbf{y}' = \boldsymbol{\gamma} \odot \mathbf{y}, \tag{6}$$

where $\odot$ denotes channel-wise multiplication. This gain operation restores response strength after binary mapping and helps suppress layer-wise scale drift, thereby improving the stability of structural and boundary-related responses during propagation. Compared with the shift operation, which mainly corrects the activation distribution before projection, the gain operation mainly compensates for the loss of magnitude information after binary mapping. These operations therefore play complementary roles in stabilizing binary feature propagation.

In practice, $\boldsymbol{\tau}$ is initialized near zero and $\boldsymbol{\gamma}$ is initialized near one, so that SGRM starts from an approximately identity transformation and maintains stable optimization at the early stage of training. In this way, SGRM provides a simple yet effective feature regulation mechanism for binarized diffusion reconstruction, reducing activation collapse before binarization and response drift after binary convolution across successive layers.

3.4 Multi-level Attention Distillation Loss

Aggressive binarization can make intermediate representations unstable in infrared super-resolution. In particular, binary propagation may weaken important responses and cause them to drift across layers, which is unfavorable for boundary preservation and stable reconstruction. To improve feature stability in the binarized student, we introduce a **Multi-level Attention Distillation (MAD) loss**, which transfers structural response patterns from a full-precision teacher to the student at multiple feature levels to support boundary preservation.

Let $\{\mathbf{F}_T^{(l)}\}_{l=1}^{L}$ and $\{\mathbf{F}_S^{(l)}\}_{l=1}^{L}$ denote feature tensors from the teacher and the student, respectively, where $\mathbf{F}_{\cdot}^{(l)} \in \mathbb{R}^{B \times C_l \times H_l \times W_l}$. Here, l indexes the selected feature levels and L is their number; these levels span the input, middle, and output blocks. A direct feature matching loss is sensitive to the magnitude discrepancy introduced by binary propagation, and may therefore over-penalize scale differences that are not essential to structural reconstruction. In contrast, our goal is to preserve the structural response layout of the teacher, rather than enforcing exact agreement in absolute feature magnitude. MAD aligns normalized energy-attention forms, which emphasize the channel-spatial distribution of strong responses instead of their raw scale.

Specifically, for each level l, we define

$$\mathcal{A}\left(\mathbf{F}^{(l)}\right) = \frac{\mathbf{F}^{(l)} \odot \mathbf{F}^{(l)}}{\left\|\mathbf{F}^{(l)} \odot \mathbf{F}^{(l)}\right\|_2 + \varepsilon}, \tag{7}$$

where $\odot$ denotes element-wise multiplication, $\|\cdot\|_2$ is computed per sample over the (C_l, H_l, W_l) dimensions, and $\varepsilon > 0$ is a small constant for numerical stability. This normalization suppresses global scale variations while preserving the relative layout of activations, *i.e.*, where the network responds more strongly. Different feature levels capture complementary structural information. Shallow layers are more sensitive to local boundaries, fine contours, and high-frequency responses, while deeper layers encode more global structural organization and context-aware response patterns. Since

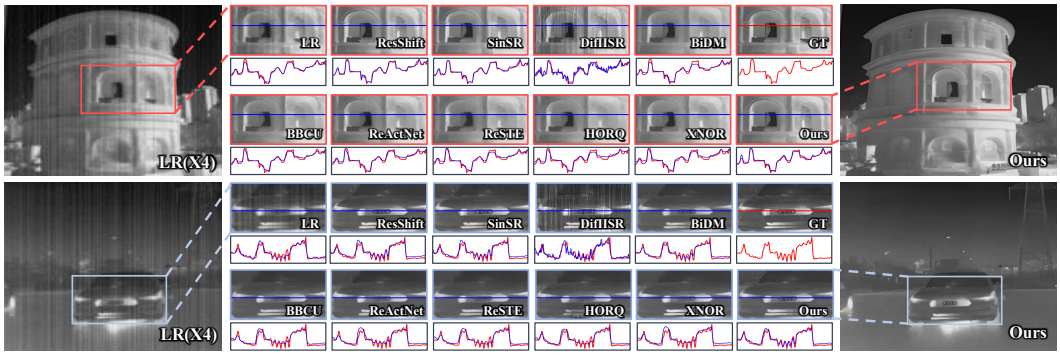


Figure 3: Visual comparison of image super-resolution results. We compare our proposed Bi-IISR with state-of-the-art binarization methods and the full-precision (FP) model. The top row shows results on the **HM-TIR** dataset, while the bottom row shows results on the **M^{3}FD** dataset.

infrared super-resolution requires both local boundary preservation and global structure consistency, aligning teacher and student responses across multiple levels is more suitable than single-level distillation for stabilizing binary diffusion reconstruction.

Based on the above attention form, MAD enforces consistency between the teacher and student across all selected levels using a mean-squared error loss:

$$\mathcal{L}_{\text{MAD}} = \sum_{l=1}^{L} \text{MSE}\left(\mathcal{A}\left(\mathbf{F}_S^{(l)}\right), \mathcal{A}\left(\mathbf{F}_T^{(l)}\right)\right). \tag{8}$$

By aligning normalized energy-attention patterns at multiple levels, MAD helps reduce representation drift caused by binarization and encourages the student to preserve the teacher's response layout, especially around boundaries and other salient regions. As a result, the student can maintain more stable representations for subsequent diffusion reconstruction.

3.5 Loss Function

The training objective of the proposed framework consists of three parts: a latent reconstruction loss, the proposed multi-level attention distillation loss, and a focal frequency loss in the image space. The reconstruction term serves as the main optimization target, while MAD regularizes intermediate structural responses and the focal frequency loss further enhances high-frequency recovery.

Formally, the overall training objective is written as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{mad}}\mathcal{L}_{\text{MAD}} + \lambda_{\text{ffl}}\mathcal{L}_{\text{FFL}}, \tag{9}$$

where $\mathcal{L}_{\text{rec}} = \text{MSE}(\hat{\mathbf{z}}_S, \mathbf{z}_T)$ matches the student's one-step latent prediction $\hat{\mathbf{z}}_S$ to the teacher's latent result $\mathbf{z}_T$, and $\mathcal{L}_{\text{FFL}}$ is computed between the decoded prediction and the ground-truth image. In our experiments, the balancing weights are set to $\lambda_{\text{mad}} = 0.01$ and $\lambda_{\text{ffl}} = 0.1$ for the MAD and FFL terms, respectively.

4 Experiment

4.1 Experimental Settings

Datasets. We conduct infrared super-resolution training and evaluation on two benchmarks, HM-TIR (Liu et al., 2025a) and M^{3}FD (Liu et al., 2022). For both datasets, we adopt an 8:2 split for training and testing. Low-resolution inputs are synthesized using the same HM-TIR degradation pipeline to emulate thermal-image degradations. Specifically, the process includes Gaussian blur, optical non-uniformity, stripe noise, and Gaussian noise, followed by $\times 4$ downsampling.

Evaluation Metrics. We evaluate super-resolution performance from distortion and perceptual perspectives. PSNR and SSIM (Wang et al., 2004) measure pixel-level fidelity and structural similarity.

Table 1: Infrared super-resolution quality and efficiency on M^{3}FD and HM-TIR. Red-highlighted bold values and blue highlights mark the best and second-best 1-bit quality results, respectively.

Datasets		M^{3}FD					HM-TIR					Step	OPs(G)	Params(M)
Methods	Bit (W/A)	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	MANIQA↑	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	MANIQA↑			
ResShift	32/32	34.410	0.941	0.070	0.091	0.471	31.272	0.877	0.107	0.123	0.528	15	753.45	118.59
SinSR	32/32	34.017	0.935	0.078	0.096	0.460	30.990	0.873	0.125	0.141	0.494	1	50.23	118.59
DifIISR	32/32	27.645	0.633	0.343	0.628	0.537	27.292	0.649	0.300	0.494	0.533	1	50.15	118.59
BBCU	1/1	32.257	0.907	0.194	0.206	0.410	30.006	0.845	0.239	0.283	0.411	1	5.83	4.95
BiDM	1/1	27.489	0.874	0.308	0.291	0.416	26.454	0.810	0.387	0.311	0.422	1	11.60	18.69
HORQ	1/1	31.918	0.910	0.182	0.199	0.417	29.955	0.850	0.228	0.272	0.418	1	5.83	4.95
ReActNet	1/1	32.615	0.912	0.178	0.194	0.414	30.241	0.851	0.226	0.271	0.414	1	5.83	4.95
ReSTE	1/1	31.409	0.904	0.200	0.223	0.405	29.508	0.843	0.242	0.305	0.403	1	5.83	4.95
XNOR	1/1	30.777	0.907	0.187	0.217	0.408	29.192	0.847	0.233	0.293	0.410	1	5.83	4.95
Ours	1/1	**33.473**	**0.922**	**0.145**	**0.160**	**0.428**	**30.796**	**0.859**	**0.204**	**0.247**	**0.442**	1	1.83	5.01

DISTS (Ding et al., 2020), LPIPS (Zhang et al., 2018), and MANIQA (Yang et al., 2022) further assess perceptual quality. We also report parameters and overall operations (OPs) to quantify the efficiency of the 1-bit architecture. Following Xia et al. (2022), the 1-bit student's counts are $\text{Params} = \text{Params}^b + \text{Params}^f$ and $\text{OPs} = \text{OPs}^b + \text{OPs}^f$, where $\text{Params}^b = \text{Params}_{\text{bin}}/32$ and $\text{OPs}^b = \text{OPs}_{\text{bin}}/64$. The FP32 and unscaled binary counts come from Conv2d/Linear and BinaryConv2d/BinaryLinear layers, respectively. We report Params in M and OPs in G, measuring OPs on a 64×64 input for fair comparison.

Implementation Details. We adopt the one-step diffusion model SinSR as the reconstruction backbone of the student network. A full-precision ResShift teacher is first pretrained on each dataset and then frozen during subsequent training of our method and the binarized baselines. The student follows the SinSR architecture, with the main convolutional and linear layers in the denoising network replaced by the proposed 1-bit quantized layers. To limit information loss at the input and output stages, the head and tail layers remain in full precision. The student is initialized from the corresponding pretrained SinSR weights.

Training Settings. We train on NVIDIA RTX 4090 GPUs using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 2×10^{-5} and decays at 2K and 10K iterations. Training lasts 200K iterations with a batch size of 8.

4.2 Comparison with State-of-the-Art Methods.

We compare our method with six binarization baselines: BBCU (Xia et al., 2022), BiDM (Zheng et al., 2024), HORQ (Li et al., 2017), ReActNet (Liu et al., 2020c), ReSTE (Wu et al., 2023), and XNOR-Net (Rastegari et al., 2016). All are implemented on SinSR (Wang et al., 2024b) and trained under identical settings. Full-precision diffusion baselines include SinSR, ResShift (Yue et al., 2023), and DifIISR (Li et al., 2025), which is designed for infrared image super-resolution.

Results. Tab. 1 shows that our method outperforms all other 1-bit baselines across the reported quality metrics on both datasets, improving reconstruction fidelity and perceptual quality. Although it remains below the strongest full-precision diffusion models, it is closer to their performance than existing 1-bit baselines. Figure 3 presents $\times 4$ visual comparisons on HM-TIR and M^{3}FD. Relative to the binarized baselines, our method produces cleaner reconstructions with clearer contours and fewer degradation artifacts. On HM-TIR, it restores building boundaries and window structures while suppressing stripe-like interference; on M^{3}FD, it better preserves the vehicle outline and local bright regions. The profile curves follow the ground-truth intensity trend more closely, with more consistent peak-valley transitions around structural boundaries. These observations support improved structural coherence, although the visual comparisons also show that the strongest full-precision models retain a quality advantage in some regions.

Efficiency Comparison. The rightmost columns of Tab. 1 report efficiency. Our method has the lowest OPs among the 1-bit baselines at a comparable parameter scale, using about 69% fewer OPs than most lightweight competitors. Relative to BiDM, it reduces OPs by over 84% and Params by about 73%. Compared with SinSR and DifIISR, it reduces OPs by over 96% and Params by about 96%, yielding a better quality-efficiency trade-off.

Table 2: Quantitative comparison on **M^{3}FD** (Object Detection) and **FMB** (Semantic Segmentation) datasets. For FMB, eight selected class IoUs are shown alongside the overall mIoU. **Boldface** with red background indicates best results, blue indicates second-best.

Datasets	M^{3}FD							FMB								
Methods	People	Car	Bus	Light	Moto	Truck	mAP	Road	Build	Sign	Tree	Sky	Human	Car	Pole	mIoU
BBCU	63.82	59.94	37.18	5.75	26.91	33.03	37.77	83.00	67.20	34.55	69.10	83.58	53.82	64.79	12.72	38.98
BiDM	58.57	47.33	23.15	4.12	22.46	23.16	29.80	55.60	46.80	12.15	43.14	68.59	28.98	29.76	6.92	22.07
HORQ	63.49	61.90	40.10	5.88	31.15	32.61	39.19	82.75	67.49	32.01	70.49	84.91	52.88	64.47	12.97	38.70
ReActNet	63.92	61.97	38.21	5.70	26.53	33.55	38.32	82.99	67.72	35.86	69.70	84.18	53.19	64.72	13.08	39.30
ReSTE	62.49	59.31	34.23	4.83	25.86	32.41	36.52	82.39	64.15	28.98	68.02	82.42	**54.26**	64.65	11.88	37.75
XNOR	62.77	59.34	34.42	4.74	24.10	31.19	36.09	82.36	65.56	31.11	68.55	82.85	53.17	64.21	12.52	37.98
Ours	**65.57**	**65.00**	**46.32**	**6.62**	**35.09**	**33.62**	**42.04**	**83.77**	**71.01**	**41.14**	**73.20**	**87.40**	54.22	**87.40**	**16.70**	**41.24**

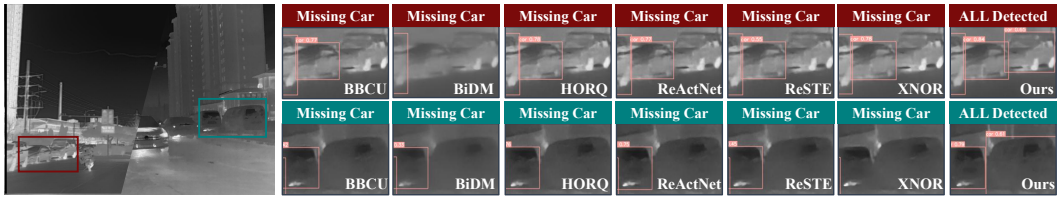


Figure 4: Qualitative comparisons of object detection results with state-of-the-art binarization methods using YOLOv8.

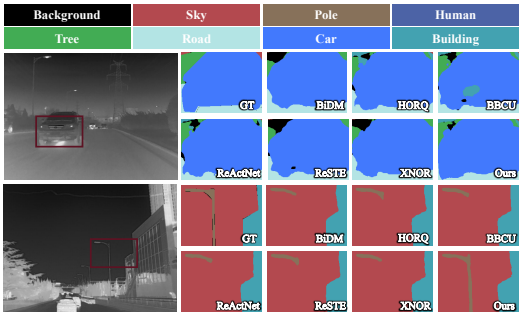


Figure 5: Qualitative semantic segmentation comparisons using SegFormer and state-of-the-art binarization methods.

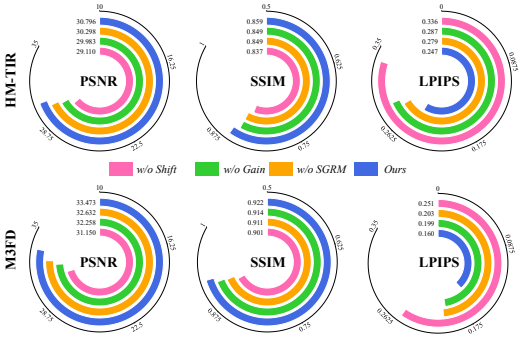


Figure 7: Quantitative ablation results of SGRM on the HM-TIR and M^{3}FD datasets.

4.3 Experiments on Infrared Object Detection

On M^{3}FD, we train YOLOv8-s (Jocher et al., 2023) with SGD (batch size 16, initial learning rate 0.01) and evaluate detection using mAP@.5:.95. As shown in the left part of Tab. 2, our method achieves the highest overall mAP among binarized baselines and ranks first in every reported object category. Gains for people, cars, and motorcycles suggest better preservation of target contours and local contrast under degraded infrared conditions. In Figure 4, several baseline reconstructions cause the detector to miss a car in the highlighted regions of both examples, whereas the car is detected from our reconstruction. Together, the category-wise gains and visual examples suggest that our reconstruction retains target structures useful for downstream detection.

4.4 Experiments on Infrared Semantic Segmentation

For segmentation on FMB (Liu et al., 2023), we train SegFormer-B0 (Xie et al., 2021) with SGD for 40,000 iterations (batch size 2, initial learning rate 0.01) and report mIoU. As shown in the right part of Tab. 2, our method achieves the highest overall mIoU among binarized baselines and leads in most displayed classes. Gains on Sign, Tree, Car, and Pole suggest better preservation of region-level structure and fine details, although Human remains slightly below the best baseline. Figure 5 shows a coherent vehicle region in the first example. In the second, several baselines weaken or

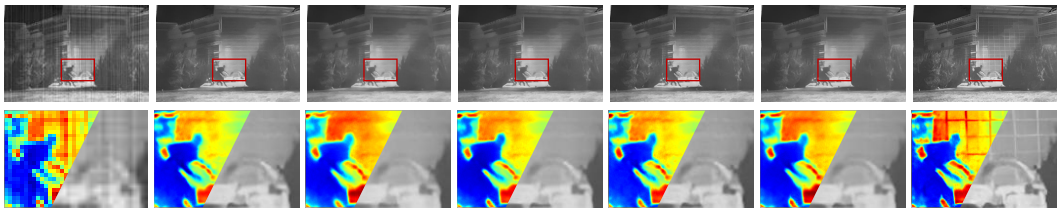


Figure 6: Qualitative comparisons of different ablative variants.

omit the narrow pole, whereas our result preserves its vertical structure more continuously against the background. Together, these results suggest that our reconstruction provides more structurally informative inputs for infrared semantic segmentation.

5 Ablation Study

We conduct two sets of ablations to clarify the roles of the proposed components. The first removes DRC, SGRM, or MAD from the full model, while the second examines Shift and Gain separately within SGRM.

Table 3: Ablation results (best in bold).

Methods	PSNR↑	SSIM↑	DISTS↓	LPIPS↓
w/o DRC	32.152	0.911	0.181	0.202
w/o SGRM	32.632	0.911	0.184	0.203
w/o MAD	32.338	0.916	0.174	0.196
Ours	**33.473**	**0.922**	**0.145**	**0.160**

5.1 Study on model architectures.

Removing DRC, SGRM, or MAD degrades performance in Tab. 3. The w/o DRC and w/o SGRM variants have nearly identical Params and OPs to Ours (Appendix A, Tab. 4), yet both show lower reconstruction quality. Their measured quality declines therefore occur without a substantial change in inference cost.

Without DRC, PSNR drops most, and the highlighted structure appears blurrier in Figure 6. Removing SGRM yields the worst DISTS and LPIPS, suggesting that stable binary propagation helps preserve perceptual quality. Removing MAD worsens all four metrics, supporting the complementary role of teacher-guided multi-level supervision. Among the three ablations, w/o MAD retains the highest SSIM and lowest DISTS and LPIPS, but it still trails the full model on every reported metric. Across all three ablations, fidelity and perceptual measures decline together, so the full model's gain is not a simple trade-off between these criteria. As a training-only loss, MAD adds no inference-time overhead. The full model achieves the best overall reconstruction quality.

5.2 Study on SGRM.

Figure 7 shows that removing Shift, Gain, or SGRM lowers PSNR and SSIM and raises LPIPS on both datasets. Removing Shift generally causes a larger decline, especially in PSNR and LPIPS, highlighting the importance of re-centering infrared activations before sign projection. The full model exceeds w/o Shift in PSNR by 1.686 dB on HM-TIR and 2.323 dB on M^3FD. Removing Gain also degrades quality, consistent with its role in restoring response magnitude after binary convolution. In Figure 6, the variants without either operation show less compact local responses and weaker boundary transitions. Together, these results support the complementary roles of Shift and Gain in stable structural reconstruction.

6 Conclusion

In this paper, we proposed Bi-IISR, a binarized one-step diffusion framework for infrared super-resolution. DRC, SGRM, and MAD jointly improve degraded conditioning, binary propagation, and structural fidelity. Experiments show a better quality-efficiency trade-off than existing binarized baselines and competitive quality against full-precision diffusion models.

AI USE STATEMENT

We used generative AI tools to assist in identifying relevant literature and improving the language clarity and readability of the manuscript. The authors reviewed all AI-assisted content and take full responsibility for the final version of the manuscript.

REFERENCES

Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 23555–23564, 2023.

Zheng Chen, Haotong Qin, Yong Guo, Xiongfei Su, Xin Yuan, Linghe Kong, and Yulun Zhang. Binarized diffusion model for image super-resolution. *Advances in Neural Information Processing Systems*, 37:30651–30669, 2024.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *European Conference on Computer Vision*, pp. 37–56. Springer, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728, 2020. URL https://arxiv.org/abs/2004.07728.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23174–23184, 2025.

Yongsong Huang, Zetao Jiang, Rushi Lan, Shaoqin Zhang, and Kui Pi. Infrared image super-resolution via transfer learning and psrgan. *IEEE Signal Processing Letters*, 28:982–986, 2021.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Xingyuan Li, Jinyuan Liu, Zhixin Chen, Yang Zou, Long Ma, Xin Fan, and Risheng Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pp. 270–288. Springer, 2024.

Xingyuan Li, Zirui Wang, Yang Zou, Zhixin Chen, Jun Ma, Zhiying Jiang, Long Ma, and Jinyuan Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7534–7544, 2025.

Zefan Li, Bingbing Ni, Wenjun Zhang, Xiaokang Yang, and Wen Gao. Performance guaranteed network acceleration via high-order residual quantization. In *Proceedings of the IEEE international conference on computer vision*, pp. 2584–2592, 2017.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5811, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8115–8124, 2023.

Jinyuan Liu, Zihang Chen, Zhu Liu, Zhiying Jiang, Long Ma, Xin Fan, and Risheng Liu. Enhancing infrared vision: Progressive prompt fusion network and benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=yEddfz9SgJ.

Kai Liu, Kaicheng Yang, Zheng Chen, Zhiteng Li, Yong Guo, Wenbo Li, Linghe Kong, and Yulun Zhang. Bimacosr: Binary one-step diffusion model leveraging flexible matrix compression for real super-resolution. *arXiv preprint arXiv:2502.00333*, 2025b.

Risheng Liu, Jinyuan Liu, Zhiying Jiang, Xin Fan, and Zhongxuan Luo. A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion. *IEEE transactions on image processing*, 30:1261–1274, 2020a.

Yuqing Liu, Shiqi Wang, Jian Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Iterative network for image super-resolution. *IEEE Transactions on Multimedia*, 24:2259–2272, 2021.

Zechun Liu, Wenhan Luo, Baoyuan Wu, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Binarizing deep network towards real-network performance. *International Journal of Computer Vision*, 128(1):202–219, 2020b.

Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *European conference on computer vision*, pp. 143–159. Springer, 2020c.

Kalpesh Prajapati, Vishal Chudasama, Heena Patel, Anjali Sarvaiya, Kishor P Upla, Kiran Raja, Raghavendra Ramachandra, and Christoph Busch. Channel split convolutional neural network (chasnet) for thermal image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4368–4377, 2021.

Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2250–2259, 2020.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pp. 525–542. Springer, 2016.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Dynamic brightness adaptation for robust multi-modal image fusion. *arXiv preprint arXiv:2411.04697*, 2024.

Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022.

Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98:101828, 2023.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024a.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25796–25805, 2024b.

Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25456–25467, 2024b.

Xiao-Ming Wu, Dian Zheng, Zuhao Liu, and Wei-Shi Zheng. Estimator meets equilibrium perspective: A rectified straight through estimator for binary neural networks training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17055–17064, 2023.

Bin Xia, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Radu Timofte, and Luc Van Gool. Basic binary convolution unit for binarized image restoration network. *arXiv preprint arXiv:2210.00405*, 2022.

Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13095–13105, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Jingwei Xin, Nannan Wang, Xinrui Jiang, Jie Li, Heng Huang, and Xinbo Gao. Binarized neural network for single image super resolution. In *European conference on computer vision*, pp. 91–107. Springer, 2020.

Yitong Yan, Chuangchuang Liu, Changyou Chen, Xianfang Sun, Longcun Jin, Xinyi Peng, and Xiang Zhou. Fine-grained attention and feature-sharing generative adversarial networks for single image super-resolution. *IEEE Transactions on Multimedia*, 24:1473–1487, 2021.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1191–1200, 2022.

Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21 (12):3106–3121, 2019.

Xin Yang, Haiyang Mei, Jiqing Zhang, Ke Xu, Baocai Yin, Qiang Zhang, and Xiaopeng Wei. Drfn: Deep recurrent fusion network for single-image super-resolution with large factors. *IEEE Transactions on Multimedia*, 21(2):328–337, 2018.

Xunxiang Yao, Qiang Wu, Peng Zhang, and Fangxun Bao. Weighted adaptive image super-resolution scheme based on local fractal feature and image roughness. *IEEE Transactions on Multimedia*, 23:1426–1441, 2020.

Wei Yu, Qinling Liu, Quanling Meng, Chenyang Wang, and Xin Sun. D3bsr: Blind super-resolution via diffusion-based disentangled degradation representation. *IEEE Transactions on Multimedia*, 2025.

Zongsheng Yue, Qian Zhao, Jianwen Xie, Lei Zhang, Deyu Meng, and Kwan-Yee K Wong. Blind image super-resolution with elaborate degradation modeling on noise and kernel. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2128–2138, 2022.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36: 13294–13307, 2023.

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

Xianjie Zhang, Yusen Zhang, Yujie He, and Min Li. S3diff: Semantic fusion and structure-guided global generation from a single image with diffusion models. *Journal of Artificial Intelligence and Soft Computing Research*, 16(1):39–53, 2026.

Xinyan Zhang, Peng Gao, Sunxiangyu Liu, Kongya Zhao, Guitao Li, Liuguo Yin, and Chang Wen Chen. Accurate and efficient image super-resolution via global-local adjusting dense network. *IEEE Transactions on multimedia*, 23:1924–1937, 2020.

Xingyu Zheng, Xianglong Liu, Yichen Bian, Xudong Ma, Yulun Zhang, Jiakai Wang, Jinyang Guo, and Haotong Qin. Bidm: Pushing the limit of quantization for diffusion models. *Advances in Neural Information Processing Systems*, 37:39009–39035, 2024.

Yang Zou, Jun Ma, Zhidong Jiao, Xingyuan Li, Zhiying Jiang, and Jinyuan Liu. Toward real-world infrared image super-resolution: A unified autoregressive framework and benchmark dataset. *arXiv preprint arXiv:2603.04745*, 2026.

A Parameter and Operation Breakdown

A.1 Structural Ablation Efficiency

Table 4: Parameter and operation breakdown for structural ablations.

Methods	Parameters					Operations		
	Conv2d	Linear	BinaryConv2d	BinaryLinear	Params(M)	FP32	Binary	OPs(G)
base	0.82	0.43	105.59	14.57	**5.0050**	1.0471	50.0706	**1.8295**
w/o DRC	0.82	0.43	105.64	14.60	**5.0075**	1.0471	50.0706	**1.8295**
w/o Shift	0.82	0.43	105.58	14.58	**5.0050**	1.0477	50.0706	**1.8301**
w/o Gain	0.82	0.43	105.59	14.57	**5.0050**	1.0477	50.0706	**1.8301**
w/o SGRM	0.82	0.43	105.53	14.55	**5.0025**	1.0477	50.0706	**1.8301**
Ours	0.82	0.43	105.64	14.60	**5.0075**	1.0477	50.0706	**1.8301**

Table 4 separates full-precision and binary costs before reporting their 1-bit-equivalent totals. Conv2d and Linear are FP32 parameter counts, whereas BinaryConv2d and BinaryLinear are raw binary parameter counts, all measured in M. Following Sec. 4.1, Params(M) sums the FP32 counts and one thirty-second of the binary counts. Likewise, OPs(G) sums the FP32 operations and one sixty-fourth of the binary operations for a 64×64 input.

The FP32 Conv2d and Linear counts remain 0.82 M and 0.43 M across all variants, and raw binary operations remain 50.0706 G. Consequently, Params(M) ranges only from 5.0025 to 5.0075, while OPs(G) ranges from 1.8295 to 1.8301. Compared with base, Ours differs by just 0.0025 M and 0.0006 G. Relative to Ours, the w/o DRC variant has the same reported Params and 0.0006 G fewer OPs; w/o SGRM has 0.0050 M fewer Params and the same reported OPs. Removing either Shift or Gain reduces reported Params by 0.0025 M without changing reported OPs. Thus, the quality gains in Tab. 3 involve negligible inference overhead. MAD is absent from this breakdown because its distillation loss is used only during training.