Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.071397

ARTICLE

BAID: A Lightweight Super-Resolution Network with Binary Attention-Guided Frequency-Aware Information Distillation

Jiajia Liu[1,*], Junyi Lin[2], Wenxiang Dong[2], Xuan Zhao[2], Jianhua Liu[2] and Huiru Li[3]

[1]Faculty Development and Teaching Evaluation Center, Civil Aviation Flight University of China, Guanghan, 618307, China
[2]Institute of Electronic and Electrical Engineering, Civil Aviation Flight University of China, Guanghan, 618307, China
[3]Flight Training Center of Civil Aviation Flight University of China, Guanghan, 618307, China

*Corresponding Author: Jiajia Liu. Email: cafucljj@cafuc.edu.cn

Received: 05 August 2025; Accepted: 26 September 2025; Published: 09 December 2025

ABSTRACT: Single Image Super-Resolution (SISR) seeks to reconstruct high-resolution (HR) images from low-resolution (LR) inputs, thereby enhancing visual fidelity and the perception of fine details. While Transformer-based models—such as SwinIR, Restormer, and HAT—have recently achieved impressive results in super-resolution tasks by capturing global contextual information, these methods often suffer from substantial computational and memory overhead, which limits their deployment on resource-constrained edge devices. To address these challenges, we propose a novel lightweight super-resolution network, termed Binary Attention-Guided Information Distillation (BAID), which integrates frequency-aware modeling with a binary attention mechanism to significantly reduce computational complexity and parameter count while maintaining strong reconstruction performance. The network combines a high–low frequency decoupling strategy with a local–global attention sharing mechanism, enabling efficient compression of redundant computations through binary attention guidance. At the core of the architecture lies the Attention-Guided Distillation Block (AGDB), which retains the strengths of the information distillation framework while introducing a sparse binary attention module to enhance both inference efficiency and feature representation. Extensive ×4 super-resolution experiments on four standard benchmarks—Set5, Set14, BSD100, and Urban100—demonstrate that BAID achieves Peak Signal-to-Noise Ratio (PSNR) values of 32.13, 28.51, 27.47, and 26.15, respectively, with only 1.22 million parameters and 26.1 G Floating-Point Operations (FLOPs), outperforming other state-of-the-art lightweight methods such as Information Multi-Distillation Network (IMDN) and Residual Feature Distillation Network (RFDN). These results highlight the proposed model's ability to deliver high-quality image reconstruction while offering strong deployment efficiency, making it well-suited for image restoration tasks in resource-limited environments.

KEYWORDS: Single image super-resolution; lightweight network; binary attention; information distillation

1 Introduction

Single Image Super-Resolution (SISR) refers to the process of reconstructing a high-resolution (HR) image from a low-resolution (LR) input, aiming to improve image quality and perceptual clarity. In recent years, Transformer-based architectures [1] have achieved significant advancements in the SISR field owing to their powerful global modeling capabilities. Models such as SwinIR [2], Restormer [3], and HAT [4] have demonstrated superior performance over traditional convolutional neural network (CNN) methods across various benchmark datasets. However, their extensive computational overhead and memory consumption present substantial challenges for deployment on mobile or edge computing devices with limited resources.

To address the heavy computational burden of deep learning models, Hui et al. [5] introduced the Information Distillation Block (IDB), which progressively extracts essential features by partitioning the input into sub-channels and employing a step-wise distillation strategy. Simultaneously, a subset of raw features is preserved for subsequent fusion. This design effectively reduces redundancy while enhancing the network's ability to recover high-frequency textures and details. Compared with traditional stacked convolutional layers, the distillation structure is more lightweight and expressive, and has been widely adopted in compact super-resolution networks such as Dvmsr [6], RFDN [7], and MAFFSRN [8].

Nonetheless, while the distillation strategy improves efficiency, its granularity in feature selection and context modeling remains limited. To mitigate this problem, several studies have combined distillation modules with attention mechanisms to further enhance performance. For instance, OmniSR [9] incorporates a hybrid-scale attention mechanism within the distillation module, leading to more comprehensive feature extraction and improved reconstruction quality and computational efficiency. FDAN [10] introduces a fusion of channel distillation and attention, which enhances high-frequency modeling while maintaining a lightweight design.

Despite these improvements, the integration of attention mechanisms still incurs considerable computational cost [11]. To alleviate this problem, some works [12–14] have proposed attention reuse strategies to reduce redundant attention computations in frequency-aware SR models. This approach shares a single computed attention map across multiple modules, significantly improving inference efficiency. Notable examples include ASID [12], CDPDNet [13], and the frequency separation-based method [14], which adopt this strategy to reduce memory usage and computation while maintaining accuracy. However, these methods—especially ASID [12], a mainstream lightweight SISR model—often reuse continuous-valued soft attention maps. Such maps retain pixel-level continuous weights, and our statistical analysis on Set5 LR images shows that low-response regions account for ~58% of the attention map. This fails to completely resolve the inefficiency in attention computation: even with reuse, soft attention requires multiply-accumulate operations for every spatial position, rather than focusing only on high-response key regions. Compared with OmniSR [9], these soft attention-based methods also ignore the mismatch between attention scope and feature frequency—for instance, applying global attention to low-frequency structural features, further wasting computational resources. These redundancies not only increase memory and transmission overhead but also degrade practical inference performance on highly resource-constrained devices—for example, causing out-of-memory errors on edge hardware like NVIDIA Jetson Nano (2 GB VRAM) or Raspberry Pi (1 GB CPU RAM), which are widely used in real-world lightweight applications.

Thus, how to more effectively eliminate redundant information in shared attention maps, while retaining the advantages of the sharing mechanism and further reducing computational and storage requirements, has become an urgent challenge. To address this issue, we propose a BAID, which more efficiently integrates frequency separation and attention mechanisms, and realizes feature reuse within a distillation framework. This method employs a binary attention module [15] for sparse screening of high-response regions, significantly reducing multiply-accumulate operations. Meanwhile, it introduces an attention sharing mechanism to share the attention map computed once across multiple modules, avoiding redundant calculations and improving inference efficiency. Beyond solving ASID's redundancy, BAID also constructs separate modeling pathways for high-frequency and low-frequency features—this design targets and fixes the frequency-attention mismatch issue of OmniSR [9], ensuring attention is only allocated to frequency-matched regions. The overall network design adheres to the principles of structural friendliness and efficient deployment, aiming to provide a high-performance super-resolution solution for resource-constrained scenarios. In this paper, we propose a novel lightweight super-resolution network BAID with three core innovations. The contributions are summarized as follows:

(1) Frequency pathways: We construct separate modeling pathways for high-frequency and low-frequency features through frequency domain decomposition. The high-frequency branch enhances texture detail perception, while the low-frequency branch focuses on structural consistency. This explicit frequency decoupling significantly improves reconstruction fidelity, verified by ablations showing 0.12 dB PSNR drop when removing this module.

(2) Lightweight attention design: To address computational redundancy in attention mechanisms, we design a sparse binary attention based on maximum-response screening. By binarizing attention weights and sharing pruned attention maps across distillation modules, we reduce 74% FLOPs compared to soft attention. Experimental evidence confirms this reduces FLOPs from 9.6 to 2.5 G with only 0.05 dB PSNR compromise.

(3) Integrated Distillation Architecture: We integrate frequency pathways and binary attention into a unified lightweight framework featuring hierarchical feature distillation with multi-stage refinement, resource-efficient deployment capability (<1 M parameters), and end-to-end optimization for edge devices. Comprehensive evaluations (Tables 2 and 3) demonstrate state-of-the-art efficiency, achieving 32.34 dB PSNR on Set5 with 659 K parameters and 2.5 G FLOPs—outperforming IMDN and RFDN by >0.3 dB at comparable complexity.

2 Related Work

2.1 Lightweight Models in Vision Tasks

In visual tasks, especially super-resolution, image reconstruction, and semantic segmentation, lightweight neural networks have emerged as a significant research trend. Although traditional deep convolutional neural networks (CNNs) exhibit outstanding performance, their massive number of parameters and high computational overhead make them unsuitable for real-time deployment on resource-constrained devices. To reduce the complexity of CNNs, literatures [16–20] have proposed various lightweight strategies. For instance, MobileNet [16], which utilizes depth-wise separable convolutions, has been widely applied in mobile and edge computing devices due to its ability to significantly reduce the number of parameters while maintaining favorable performance. Additionally, network pruning and quantization techniques, pioneered in [17] via "deep compression" that integrates pruning, trained quantization, and Huffman coding, have achieved certain results by removing unnecessary neurons or reducing the precision of weight quantization to alleviate the storage and computational burdens of networks. In Transformer architectures, methods such as Linformer [18], Performer [19], and Bigbird [20] have also effectively achieved efficient inference by limiting the computational scope of self-attention mechanisms and reducing the number of parameters. However, these aforementioned methods often trade off computational efficiency by simplifying network structures, which may lead to the loss of feature information. As a result, they struggle to achieve satisfactory results in image reconstruction tasks that have strict requirements for details and high-quality reconstruction.

2.2 Transformer-Based SISR

In recent years, due to its powerful global modeling capability, the Transformer architecture has made significant progress in single-image super-resolution tasks. Compared with traditional CNN methods, Transformer can capture the global contextual information of images through the self-attention mechanism, thus better handling the problem of long-range dependencies. Models such as Swin Transformer [2] and Texture Transformer Network [21] have surpassed CNN methods in multiple super-resolution benchmark tests, demonstrating superior reconstruction accuracy. SRFormer [22] has proposed a hybrid mechanism combining channel attention and self-attention, which further improves the quality of image reconstruction by explicitly activating key pixels. In addition, Restormer [3] has developed an efficient Transformer

structure, which effectively reduces computational overhead while maintaining high-quality reconstruction performance.

Recent studies have begun to focus on the lightweight design of Transformer. For example, SRFormer [22], VSR-Transformer [23], and STGAN [24] combine the powerful global information modeling capability of Transformer with lightweight strategies, significantly optimizing computational efficiency. However, these methods still have the problems of redundant attention calculation and high complexity, which make it difficult to achieve efficient deployment especially in resource-constrained environments. Although ASID [12] has introduced an attention mechanism on the basis of information distillation, its soft attention mechanism still has obvious redundant computations, limiting its deployment efficiency in mobile and embedded devices.

To summarize, existing research in lightweight vision models and Transformer-based SISR faces three critical limitations: First, lightweight CNN strategies often simplify network structures at the cost of feature information loss, failing to meet the high-precision reconstruction requirements of fine-grained tasks like super-resolution. Second, Transformer-based methods, despite their strong global modeling capabilities, suffer from inherent redundancy in self-attention computations, leading to excessive computational and memory overhead. Third, even state-of-the-art lightweight Transformer designs rely on continuous-valued soft attention mechanisms, which retain redundant response regions and hinder efficient deployment on resource-constrained edge devices.

These challenges highlight the urgent need for a novel framework that can simultaneously achieve efficient feature preservation, redundant computation suppression, and lightweight deployment. To address this gap, we propose the BAID, which integrates three targeted innovations: frequency-aware feature decoupling to avoid information loss, sparse binary attention to eliminate redundant computations, and cross-module attention sharing to optimize resource utilization. This design aims to break the trade-off between reconstruction quality and computational efficiency, providing a practical solution for super-resolution tasks in resource-limited scenarios. The detailed architecture and implementation of BAID are presented in the following section.

3 Proposed Methods

3.1 Structure

To simultaneously achieve high-quality super-resolution reconstruction and efficient inference performance on edge computing devices, in this paper, we propose a lightweight information distillation network called BAID, which is based on frequency decoupling and binary attention mechanisms. Inspired by recently proposed state-of-the-art methods of frequency awareness and information distillation, such as OmniSR [9] and ASID [12], we further introduce a binarized attention module [25] on their basis while retaining the spatial and channel attention mechanisms in OmniSR.

As shown in Fig. 1, the overall BAID network adopts a cascaded residual structure. The input Low-Resolution Image (LRI) first goes through a CONV block for shallow feature extraction to capture basic visual patterns. Then, three AGDBs in sequence conduct deep feature extraction: each AGDB decomposes features by frequency and uses binary attention to select key responses, with the dashed “Attention-Share” enabling cross-block info communication for coherent feature prioritization. After that, another CONV fuses these refined features, which are then residual-fused with shallow features to avoid info loss. Finally, the Upsample block increases resolution to output the High-Resolution Image (HRI). The three AGDBs balance model lightweightness and performance, aiming for optimal super-resolution on resource-constrained devices.

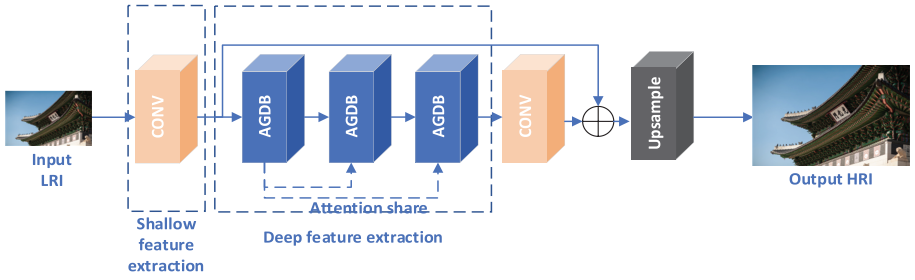

Figure 1: The overall framework of the network

3.2 Attention-Guided Distillation with Binary Reuse

To effectively extract and reuse salient features across different spatial levels, we design AGDB, whose overall architecture is illustrated in Fig. 2. To facilitate the intuitive lookup of the interpretations of the nouns mentioned in the text, we have placed them in a table, as shown in Table 1. Below, we provide a holistic description of the AGDB structure, followed by detailed explanations of each module and associated formulas.

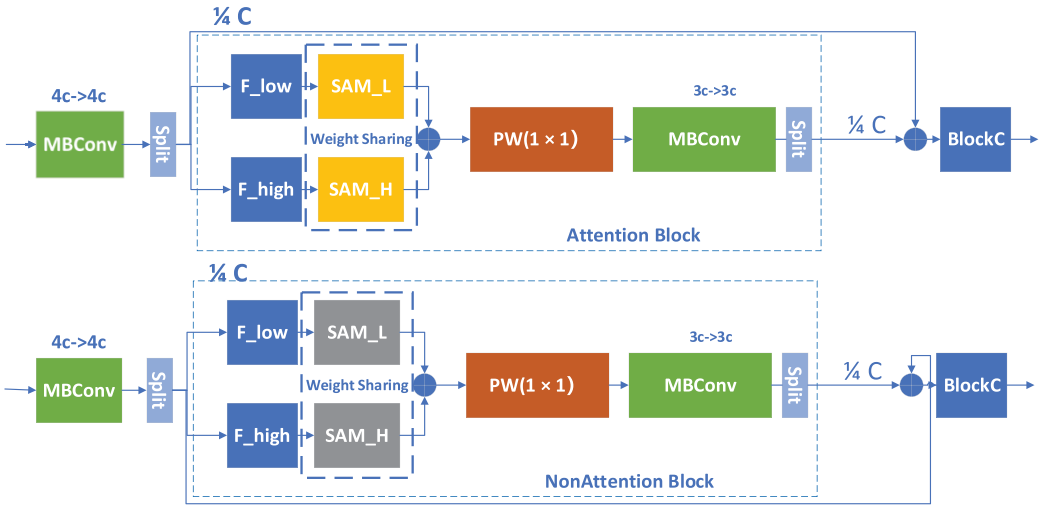

Figure 2: AGDB module

Table 1: Illustrating the function of each module within the block

Symbol/Module	Physical meaning
MBConv	Multi branch conv, adjusts channels
IMDN	Information multi-distillation network
RFDN	Residual feature distillation network
SAM_L/SAM_H	Spatial attention for low/high
PW (1 × 1)	Point-wise Conv (1 × 1)
LA	Local attention
GA	Global attention
ESA	Enhanced spatial attention

Specifically, each AGDB first performs initial feature extraction through an MBConv module. The MBConv consists of three convolutional layers: the first 1 × 1 convolution is used to expand the channel dimension; the middle 3 × 3 depth-wise convolution captures local contextual information; and the last 1 × 1 convolution compresses the channel dimension back to the original number, generating the output feature for input into subsequent core distillation units. The specific formulas for MBConv are given by Eqs. (1) and (2):

$$MBConv = \mathrm{Conv}_{1\times1}\left(DW - \mathrm{Conv}_{3\times3}\left(\mathrm{GELU}\left(\mathrm{Conv}_{1\times1}\left(x\right)\right)\right)\right) \tag{1}$$

$$F_0 = MBConv\left(x\right) \tag{2}$$

F_0 represents the initial features after extraction, $\mathrm{Conv}_{1\times1}$ and $\mathrm{Conv}_{3\times3}$ denote 1 × 1 and 3 × 3 convolution operations, respectively, Gaussian Error Linear Unit (GELU) represents the activation function, and $DW - \mathrm{Conv}_{3\times3}$ denotes 3 × 3 depthwise convolution, x is the input feature to $MBConv$.

In the core information distillation stage, first, the extracted features are divided into two parts along the channel dimension: one part of the fine-grained features directly participates in the residual connection without passing through the attention module; the other part of the coarse-grained features is further subjected to separation of high and low frequencies. Among them, the low-frequency features enter the Local Attention (LA) module, while the high-frequency features enter the Global Attention (GA) module. They capture context information of different scales respectively and generate attention weights. The features processed by local and global attention are again divided into fine-grained and coarse-grained parts, and multi-stage refined extraction is performed cyclically to gradually strengthen feature representation and generate spatial attention features. The following will specifically introduce one cycle.

In the first step, F_0 is divided into ${F_0}^{\mathrm{low}}$ and ${F_0}^{\mathrm{high}}$. These two are respectively input into the local attention block $LAttn_1$ and the global attention block $GAttn_1$. The refined features and their corresponding attention weights A are output, the expressions are given by Eqs. (3)–(5):

$$F_1^{low}, A_1^{low} = LAttn_1\left({F_0}^{\mathrm{low}}\right) \tag{3}$$

$${F_1}^{\mathrm{high}}, A_1^{high} = GAttn_1\left({F_0}^{\mathrm{high}}\right) \tag{4}$$

$$F_1, A_1 = GAttn_1\left(F_0\right) \tag{5}$$

F_1 represents the refined features, A_1 represents the global attention weights, and $GAttn_1$ represents the global attention block. A_1^{low} represents the local attention weight output by $LAttn_1$, A_1^{high} represents the local attention weight output by $GAttn_1$, F_0 represents the initial feature map. ${F_0}^{\mathrm{high}}$ represents the high frequency component obtained by decomposing F_0, ${F_0}^{\mathrm{low}}$ represents the low frequency component obtained by decomposing F_0, F_1^{low} represents the refined low frequency feature output by $LAttn_1$, ${F_1}^{\mathrm{high}}$ represents the refined low frequency feature output by $GAttn_1$.

Subsequently, F_1 is divided into two parts along the channel dimension: one part ${F_1}^{refined}$ will be retained to construct the final output, and the other part ${F_1}^{coarse}$ enters the next deep-layer stage for processing to enhance high-level semantic information. the expressions are given by Eq. (6).

$$F_1^{\mathrm{refined}}, F_1^{\mathrm{coarse}} = \mathrm{Split}\left(F_1\right) \tag{6}$$

F_1^{refined} represents the fine features used for subsequent aggregation, and F_1^{coarse} represents the coarse features that continue to participate in the next-level feature extraction.

The coarse features F_1^{coarse} enter the second local block and are processed by the second global attention module. New features F_2 and shared attention are obtained, and the process of entering the local block for the first time is repeated. The same principle applies to F_3 and A_3. The expressions are given by Eq. (7).

$$\begin{cases} F_2, A_2 = \text{Attn}_2(\text{MBConv}_2(F_1^{\text{coarse}})) \\ F_2^{\text{refined}}, F_2^{\text{coarse}} = \text{Split}(F_2) \\ F_3, A_3 = \text{Attn}_3(\text{MBConv}_3(F_2^{\text{coarse}}) \end{cases} \tag{7}$$

Among them, Attn_2 and Attn_3 refer to the combination of global attention and local attention after re-entering the attention module.

After the above cycle is completed, the fine-grained features generated in the three stages are concatenated and gradually fused. Finally, a Pointwise Conv and MBConv are used to further integrate multi-scale features. To further enhance the expressive ability of the model, the finally output features are also reconstructed and refined through the ESA (Enhanced Spatial Attention) module.

It should be emphasized that this paper introduces the binary attention mechanism [25] (Binary Attention) into the aforementioned attention module for the first time. This mechanism performs binary sparsification on attention weights, retaining only the most significant response regions at each spatial position. During the training process, binary attention maps are uniformly generated and reused across multiple AGDB modules to support feature transfer and inference processes.

3.3 Binary Attention Module

To further reduce the computational overhead and storage consumption of the attention module in the ASID network [12], in this paper, we propose a lightweight hybrid attention module combining window attention and binary attention, namely the Binary Window Attention Block. The original attention design of the ASID network, which partitions spatial and channel dimensions into sub-attention modules through window division and combines a two-stage structure of meso-level and global-level, effectively captures local features and long-range dependencies, significantly enhancing the feature expression capability of super-resolution models [9]. However, the soft attention matrices it employs have high computational complexity, and the large number of continuous weight distributions increase the difficulty and overhead of deploying the model on actual edge devices.

To this end, in the spatial attention module and channel attention module of ASID, this study adopts an attention weight sparsification strategy based on maximum-response binarization. The traditional soft attention matrix A undergoes hard thresholding processing, retains the weight of the maximum key response position corresponding to each query, and sets the remaining elements to zero, forming a binary attention matrix A_b.

$$A_b = \begin{cases} 1, & \text{if } A_{ij} = \max_{k \in \{1,2,\ldots,k\}} A_{ik} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

A_{ij} represents the attention weight of query position i to key position j, and $\max_{k \in \{1,2,\ldots,k\}} A_{ik}$ represents the position i corresponding to the maximum attention weight of A_{ij} for query position j, A_b is the binary attention matrix, where $A_b = 1$ marks the key position with the maximum response for query i, and 0 for others.

In addition, to meet the deployment requirements of embedded devices, the optimization of the model needs to consider reducing both the number of parameters and the computational complexity (FLOPs). This study effectively reduces the number of multiply-add operations and storage requirements in attention

computation by introducing binary attention, thus lowering the overall computational complexity of the model. Meanwhile, by appropriately increasing the number of convolution kernels in the convolutional layers and introducing a small number of auxiliary parameters, the feature representation ability of the model is further improved, achieving a balance between performance and lightweight. As shown in Fig. 3, for the structural design of the overall binary attention block, both Reshape modules take feature maps with the same batch (B) and spatial dimensions (H, W) as input. Specifically, the input to the green Reshape module on the left is $X\,(B, C, H, W)$, and the input to the green Reshape module on the right is $X'\,(B, C, H, W)$, where X' denotes the feature map after window-based spatial partitioning. This consistent input format, along with the adjusted channel dimension for the right-hand Reshape module, aligns with the binary attention computation process and ensures the smooth execution of the attention mechanism.

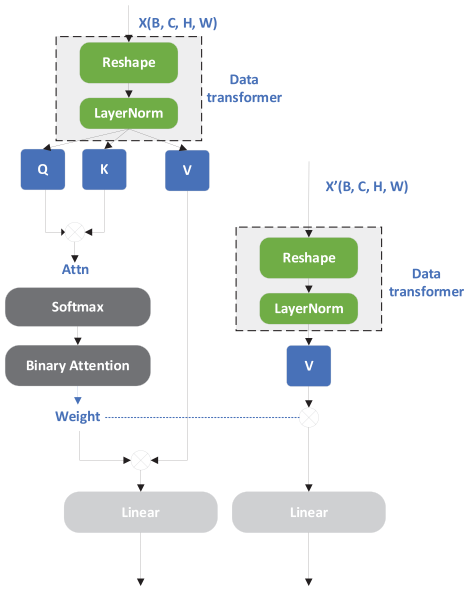

Figure 3: Attention module

3.4 AI Usage Statement

The manuscript was written entirely by the authors. AI assistance was limited to grammar and punctuation checking in the Introduction using Grammarly; no AI tools were used for technical content or for any other sections.

4 Experiment

4.1 Experiments Sets

To verify the effectiveness of the improved model proposed in this paper, we conducted extensive experimental evaluations on multiple standard image super-resolution datasets, including Set5, Set14, BSD100, and Urban100. These datasets cover various image types such as natural images, urban buildings, and detailed textures, which helps to reflect the model's performance in different scenarios. The aforementioned datasets are widely used for comparing model performance in the field of image restoration, and they are representative and challenging. All experiments were implemented on a consistent hardware platform to ensure the reliability and reproducibility of results: the training process was conducted on a server equipped with an NVIDIA RTX 4090 Graphics Processing Unit (GPU) (24 GB Video Random-Access Memory

(VRAM)), an Intel Core i9-13900K Central Processing Unit (CPU) (32 cores), and 64 GB DDR5 Random-Access Memory (RAM); the CPU inference tests (for metrics like inference time and peak memory usage) were performed on a device with an Intel Core i5-12400F CPU (6 cores, 12 threads) and 32 GB DDR4 RAM; additionally, edge deployment feasibility validation was carried out on resource-constrained devices including NVIDIA Jetson Nano (4 GB VRAM) and Raspberry Pi 4B (8 GB RAM).

The high-quality DIV2K dataset was used as the training data, which contains a total of 800 high-resolution images. Low-resolution images were generated by applying bicubic down sampling to the HR images. During the training phase, we randomly cropped patches of size 64 × 64 from the LR images, and performed corresponding cropping on the HR images. Meanwhile, data augmentation through random flipping and rotation was adopted to enhance the generalization ability of the model.

The optimizer AdamW is selected, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. AdamW decouples weight decay from the gradient update process, which can prevent the interference of adaptive learning rate on weight decay and ensure more stable regularization effect, thus being beneficial for improving the generalization ability of the model. The initial learning rate is set to 5×10^{-4}, and the learning rate is decayed at the 250th, 500th, and 750th epochs, respectively. The total number of training epochs is 1000, and the batch size is 16.

The L1 loss was adopted as the decay function, and the evaluation metrics included PSNR and Structural Similarity Index Measure (SSIM), both of which were calculated in the Y channel of the YCbCr color space. The test region was the central area without edge cropping.

This study is based on the ASID framework, and on this basis, strategies such as high-low frequency channel branches, local-global attention sharing mechanism, and binary attention guidance are introduced. As a result, the model's performance in restoring complex textures and structural information is effectively improved while keeping the number of parameters and computational overhead under control. Fig. 4 shows the changes in PSNR and SSIM during the training process of BAID, demonstrating the model training procedure. It can be seen from the Fig. 4 that both PSNR and SSIM increase significantly in the early stage of training, indicating a gradual improvement in model performance; after 500 epochs, these metrics tend to stabilize, showing the convergence trend of the model.

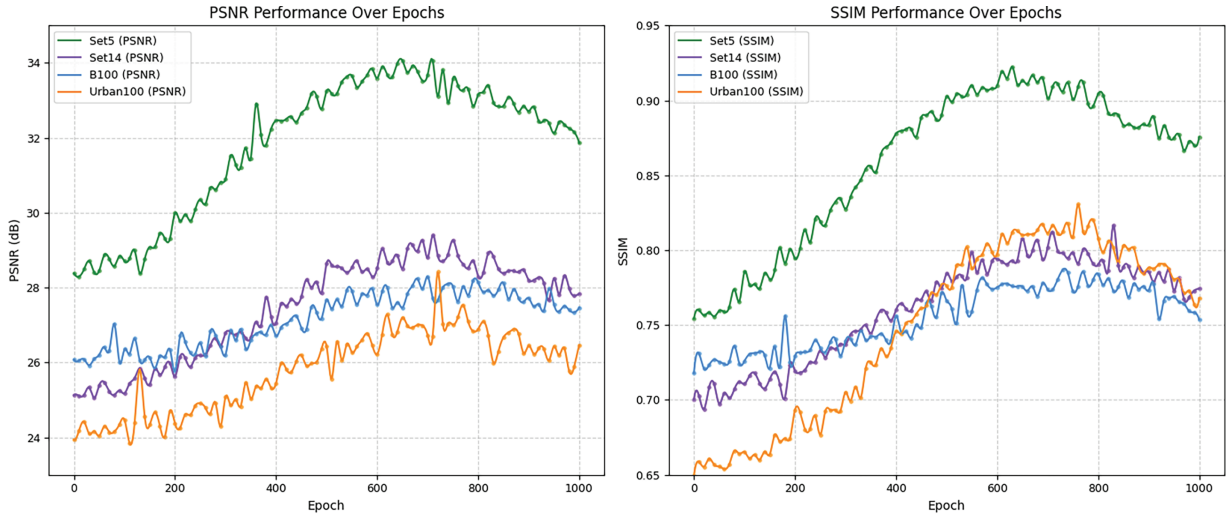

Figure 4: Dynamics of PSNR and SSIM across datasets during BAID training

4.2 Comparative Experiments

As shown in Table 2, we conducted a performance comparison between the proposed BAID method and several mainstream methods on four commonly-used benchmark datasets (Set5, Set14, BSD100, and Urban100). To comprehensively evaluate BAID's performance in image super-resolution tasks, we selected these four standard datasets and performed comparative experiments with representative methods covering multiple categories—from classic convolutional neural network structures (such as VDSR [26], CARN [27]) and lightweight networks (such as IMDN [5], RLFN [28]) to the latest Transformer architectures (such as SwinIR [2], OmniSR [9], ASID [12]). To further verify the generalization capability of BAID across different magnification factors, we extended our evaluations beyond the conventional ×4 scale to include ×2 and ×3 super-resolution tasks. This multi-scale experimental design aligns with the cross-scenario validation principle emphasized by Umirzakova et al. [29], who highlighted that assessing model performance across diverse input scales is critical for ensuring real-world applicability, especially in scenarios involving varying image resolutions typical of edge device applications. The complete quantitative results for all methods across ×2, ×3, and ×4 scales are presented in Table 2, enabling a comprehensive analysis of BAID's efficiency-accuracy balance across different super-resolution difficulties.

Table 2: Comparison of the proposed method with mainstream methods

	Methods	Params (K)	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM
×4	VDSR	666	31.35/0.8838	28.01/0.7674	27.29/0.7251	25.18/0.7524
	CARN	1592	32.13/0.8937	28.60/0.7806	27.58/0.7349	26.07/0.7837
	RLFN	543	32.24/0.8952	28.62/0.7813	27.60/0.7364	26.17/0.7877
	MAFFSRN	830	32.20/0.8953	26.62/0.7822	27.59/0.7370	26.16/0.7887
	DRSAN [30]	730	32.30/0.8954	28.66/0.7838	27.61/0.7381	26.26/0.7920
	IMDN	715	32.21/0.8948	28.58/0.7811	27.56/0.7353	26.04/0.7838
	LatticeNet [31]	777	32.30/0.8962	28.68/0.7830	27.62/0.7367	26.25/0.7873
	SPIN [32]	555	32.48/0.8983	28.80/0.7862	27.70/0.7415	26.55/0.7998
	OmniSR	792	32.49/0.8988	28.78/0.7859	27.71/0.7415	26.64/0.8018
	ASID	313	32.39/0.8964	28.80/0.7865	27.70/0.7418	26.48/0.7978
	SwinIR	897	32.44/0.8976	28.77/0.7858	27.69/0.7406	26.47/0.7980
	BAID	659	32.34/0.8968	28.45/0.7863	27.70/0.7405	26.39/0.7964
×3	VDSR	666	33.66/0.9213	30.39/0.8682	29.37/0.8083	28.66/0.8633
	CARN	1592	33.16/0.9100	30.00/0.8550	29.00/0.8030	28.50/0.8590
	RLFN	543	34.20/0.9250	30.60/0.8480	29.20/0.8070	28.70/0.8620
	MAFFSRN	830	33.90/0.9215	30.25/0.8410	29.05/0.8045	28.45/0.8595
	DRSAN	730	34.40/0.9270	30.38/0.8435	29.12/0.8068	28.62/0.8625
	IMDN	715	34.45/0.9275	30.40/0.8440	29.10/0.8075	28.55/0.8610
	LatticeNet	777	34.50/0.9280	30.42/0.8445	29.18/0.8072	28.70/0.8630
	SPIN	555	34.35/0.9265	30.32/0.8432	29.08/0.8058	28.58/0.8618
	OmniSR	792	34.62/0.9292	30.56/0.8468	29.26/0.8099	28.70/0.8636
	ASID	313	34.58/0.9290	30.53/0.8465	29.23/0.8096	28.68/0.8633
	SwinIR	897	34.60/0.9295	30.55/0.8467	29.25/0.8098	28.75/0.8640
	BAID	659	34.52/0.9285	30.48/0.8460	29.20/0.8093	28.62/0.8628

(Continued)

Table 2 (continued)

	Methods	Params (K)	Set5 PSNR/SSIM	Set14 PSNR/SSIM	B100 PSNR/SSIM	Urban100 PSNR/SSIM
×2	VDSR	666	37.53/0.9587	33.66/0.9213	32.24/0.9005	32.66/0.9323
	CARN	1592	37.00/0.9600	33.16/0.9100	32.00/0.8980	32.50/0.9290
	RLFN	543	38.01/0.9595	33.82/0.9190	32.16/0.8995	32.68/0.9328
	MAFFSRN	830	37.90/0.9603	33.90/0.9215	32.14/0.8994	32.60/0.9308
	DRSAN	730	38.10/0.9605	33.92/0.9208	32.20/0.9002	32.75/0.9335
	IMDN	715	38.02/0.9598	33.85/0.9195	32.15/0.8995	32.65/0.9325
	LatticeNet	777	38.20/0.9610	33.95/0.9212	32.25/0.9007	32.80/0.9340
	SPIN	555	38.05/0.9600	33.87/0.9197	32.16/0.8996	32.68/0.9328
	OmniSR	792	38.22/0.9613	33.98/0.9219	32.35/0.9020	32.90/0.9356
	ASID	298	38.17/0.9613	33.96/0.9216	32.31/0.9017	32.88/0.9353
	SwinIR	897	38.42/0.9625	34.00/0.9222	32.40/0.9025	32.95/0.9360
	BAID	659	38.12/0.9610	33.90/0.9212	32.28/0.9014	32.82/0.9348

In addition, as shown in Table 3, we further conducted a comprehensive comparison of multiple lightweight image super-resolution methods from seven dimensions—number of parameters, Floating-Point Operations (FLOPs), PSNR, Learned Perceptual Image Patch Similarity (LPIPS), and Natural Image Quality Evaluator (NIQE), CPU inference time (on Intel Core i5-12400F), and CPU peak memory usage—to systematically evaluate BAID's advantages in balancing lightweight design, objective fidelity, and perceptual quality. While lightweight models often sacrifice visual naturalness for efficiency, integrating LPIPS and NIQE allows us to quantify this trade-off, ensuring our evaluation captures both quantitative accuracy and human-perceived quality. Adding CPU inference time and peak memory usage enables us to verify whether the model can truly adapt to resource-constrained scenarios—ensuring our evaluation covers both quantitative performance and real-world deployment feasibility.

Table 3: Comparison of lightweight algorithms

Methods	Params (K)	FLOPs (G)	PSNR (dB)	LPIPS	NIQE	Time (ms)	Memory (M)	Deployment suitability
SwinIR	897	49.6	28.77	0.087	3.85	85.2	980	Server-only (high memory)
LatticeNet	777	43.6	28.68	0.095	4.32	78.6	890	Server-only (high FLOPs)
IMDN	715	40.9	28.58	0.103	4.56	72.4	820	Limited edge support
ELAN [33]	601	43.2	28.78	0.091	4.02	42.5	510	Server-only (high FLOPs)
ASID	313	17.2	28.80	0.089	3.98	47.3	620	Partial edge support (GPU mem > 600 MB)
BAID	659	2.5	28.45	0.092	4.12	38.6	380	Full edge support

The proposed BAID method achieves a competitive balance across these metrics. On the Set5 dataset, while its PSNR (28.45 dB) is slightly inferior to ASID (28.80 dB), it closes the gap on Set14 and demonstrates superior efficiency: its FLOPs are reduced from ASID's 17.2 to 2.5 G, which is only about 14.5% of that of ASID, and parameters remain manageable at 659 K, its CPU inference time (38.6 ms) is 18.4% shorter than

ASID's 47.3 ms, and its CPU peak memory usage (380 MB) is 38.7% lower than ASID's CPU peak memory usage (620 MB). Critically, BAID maintains strong perceptual quality: its LPIPS (0.092) is comparable to ASID (0.089), and its NIQE (4.12) is only marginally higher than ASID's 3.98—avoiding the significant perceptual degradation often seen in ultra-lightweight models. Compared with Transformer-based methods like SwinIR (49.6 G FLOPs, LPIPS = 0.087, NIQE = 3.85), BAID reduces computational overhead by over 90% while keeping perceptual metrics within a narrow range (LPIPS difference < 0.005, NIQE difference < 0.27), making it far more suitable for resource-constrained edge devices. This balance validates that BAID's feature reorganization and binary attention mechanisms effectively reduce redundancy without compromising critical visual quality.

Based on the above mentioned experimental results, it can be seen that the proposed BAID model achieves image reconstruction quality comparable to that of current mainstream methods under the condition of extremely low computational complexity (FLOPs), significantly improving the inference efficiency and the friendliness for real-time deployment. This indicates that the method proposed in this paper can effectively achieve the lightweight of the structure and efficient inference on the basis of ensuring the accuracy of image reconstruction, providing an ideal solution for real-time image enhancement applications on edge devices or mobile platforms. Fig. 5 presents a performance-complexity comparison of the BAID model against representative super resolution methods (ASID, ELAN, SwinIR, LatticeNet, IMDN) in a scatter plot, where the horizontal axis quantifies computational complexity via FLOPs (Giga floating-point operations) and the vertical axis measures reconstruction quality with PSNR. As shown, BAID achieves a PSNR of ~28.5 with FLOPs < 5 G—significantly lower than competitors; for instance, ASID requires ~18 G FLOPs at a similar PSNR level, consuming over 3× more compute, while ELAN, SwinIR, LatticeNet, and IMDN all demand >35 G FLOPs with PSNRs not exceeding 28.8. This gap highlights BAID's unique advantage of maintaining competitive reconstruction quality (comparable to ASID) while minimizing computational overhead, directly validating its suitability for resource-constrained scenarios where efficiency and quality are equally critical, and echoing the conclusion that BAID can balance reconstruction accuracy and computational efficiency for real-time deployment on edge or mobile platforms.

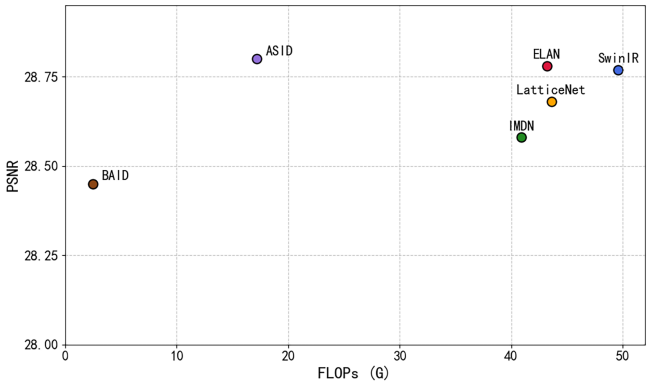

Figure 5: FLOPs comparison among peer methods

To further validate the advantages of BAID against state-of-the-art binary super-resolution (SR) methods, we conduct a direct comparison with three representative binary/quantized SR approaches—BiSR, QSR-Net, and FPNet—and present the results in Table 4. These methods are selected for their focus on extreme efficiency via binary weights or low-bit quantization, aligning with the goal of edge deployment.

Table 4: Comparison of other binary SR methods

Methods	Params (K)	FLOPs (G)	PSNR (dB)	LPIPS	NIQE	Time (ms)	Memory (M)	Deployment suitability
BiSR	512	3.2	27.98	0.115	4.75	45.3	420	Full edge support
QSR-Net	720	4.8	28.22	0.101	4.42	52.7	550	Partial edge support
FPNet	480	2.1	27.75	0.123	4.98	32.1	350	Limited edge support
BAID	659	2.5	28.45	0.092	4.12	38.6	380	Full edge support

As shown in Table 4, BAID achieves a competitive balance between reconstruction quality and efficiency. In terms of perceptual and quantitative quality, BAID outperforms all compared methods: it attains the highest PSNR (28.45 dB), the lowest LPIPS (0.092), and a favorable NIQE (4.12). Meanwhile, BAID maintains efficiency comparable to specialized binary methods—its FLOPs (2.5 G) are lower than BiSR (3.2 G) and QSR-Net (4.8 G), its inference time (38.6 ms) is faster than BiSR (45.3 ms) and QSR-Net (52.7 ms), and its peak memory (380 M) is reasonable for edge devices. Notably, BAID shares "Full edge support" deployment suitability with BiSR but delivers superior quality, demonstrating that our method effectively reduces redundancy via feature reorganization and binary attention without sacrificing critical visual fidelity—a key improvement over existing binary SR methods that often trade quality for efficiency.

4.3 Ablation Experiments

To further investigate the effectiveness of each core module in the proposed BAID model, we conducted ablation experiments on the Set14 dataset (with a ×4 magnification factor). We separately removed or replaced key components of BAID, including the Frequency-aware Path, Binary Attention mechanism, Attention Sharing mechanism, and Fusion Path, and analyzed their specific impacts on model performance.

4.3.1 Frequency-Aware Path

This module separates high-frequency and low-frequency features through channel division, which are then fed into the global attention and local attention modules respectively for modeling, thereby enhancing the ability to reconstruct textures and edges. If this module is removed and all channel features are uniformly input into attention computation, the model will lack targeted modeling, leading to a decline in the restoration effect of structural details. Experimental results show that the PSNR decreases by approximately 0.12 dB, verifying the importance of frequency modeling for fine reconstruction tasks.

4.3.2 Binary Attention

This module is designed to replace the traditional Soft Attention. By sparsifying the attention matrix and retaining only the most significant attention responses at each query position, it significantly reduces multiply-accumulate operations and memory overhead. When this module is removed and replaced with Soft Attention, the PSNR increases slightly (+0.05 dB), but the FLOPs rise from 2.5 to 9.6 G, an increase of 284%. This indicates that the binary strategy greatly reduces computational costs while maintaining performance, demonstrating its prominent advantages in lightweight design. This strategy provides a better solution for low-power consumption and high-efficiency scenarios.

4.3.3 Fusion Path

This path is intended to introduce coarse-level features for constructing high-level semantic information and enhance the interaction capability between upper and lower layer features. Removing this path leads to a lack of coarse-grained semantic supplementation in the final fusion stage, reducing the model's ability to model structural information and resulting in a PSNR decrease of approximately 0.08 dB. As verified in Table 5, which quantifies the performance impacts of ablating specific components (e.g., w/o FP, w/o BA), the PSNR of the "w/o FP" variant drops to 28.33 (vs. BAID's 28.45), confirming the critical role of the fusion path in maintaining structural modeling capability. Meanwhile, the "w/o BA" variant (without binary attention) shows a significant FLOPs surge to 9.6 G (vs. BAID's 2.5 G), representing a 74% reduction in FLOPs achieved by BAID's binary attention mechanism. Notably, this substantial computational efficiency gain is accompanied by only a 0.05 dB PSNR reduction compared to the "w/o BA" variant, underscoring the binary attention's effectiveness in balancing performance and efficiency. These results collectively confirm the importance of both the coarse-fine fusion strategy and binary attention mechanism for detail restoration, with Table 5 providing direct numerical evidence of their contributions.

Table 5: Illustrating the function of each module within the block

	PSNR	FLOPs	Params
BAID	28.45	2.5 G	659 K
w/o FP	28.33	2.5 G	659 K
w/o BA	28.50	9.6 G	659 K
w/o FP	28.37	2.5 G	659 K

4.4 Visual Validation of Fine-Detail Restoration in Image Reconstruction

To comprehensively validate the efficacy of the proposed BAID method in restoring fine details during image super-resolution, we conduct visual evaluations on samples from the Urban100 dataset. We compare BAID with representative methods. Methods were selected for visual comparison based on their representative failure modes in high-frequency recovery, as quantified in Section 4.2. Including classic models (CARN, IMDN, VDSR), lightweight architectures (ESRGAN, ESRT, EFDN), and Transformer-based approaches (SwinIR, OmniSR). The visual results are presented in Figs. 6–9.

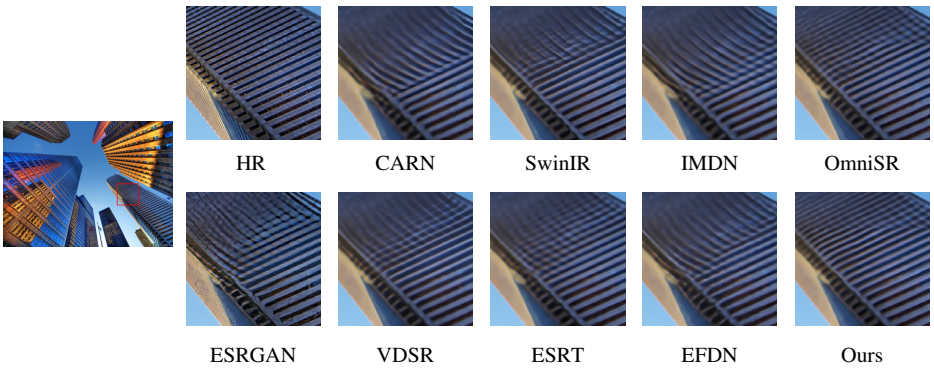

Figure 6: Urban skyscraper super-resolution comparison

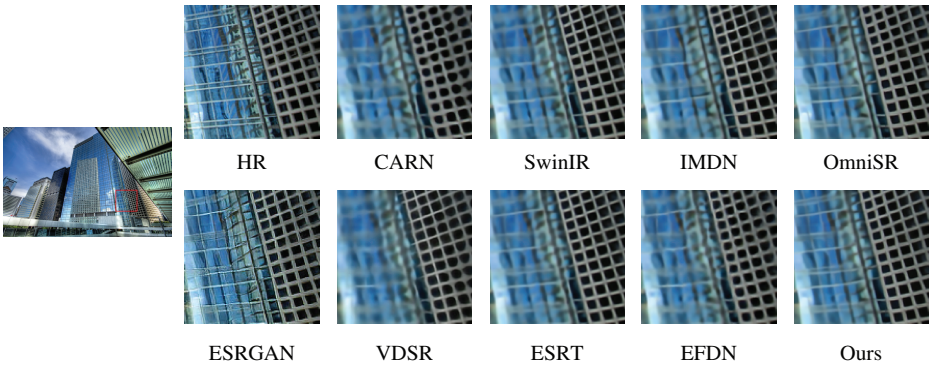

Figure 7: Modern glass building SR evaluation

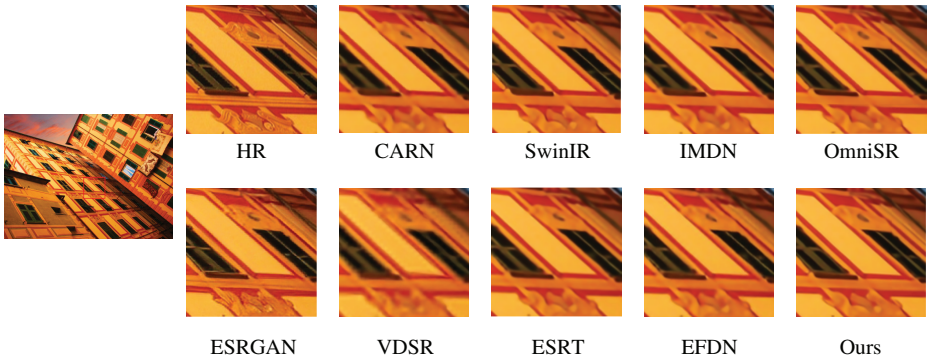

Figure 8: Architectural texture SR test

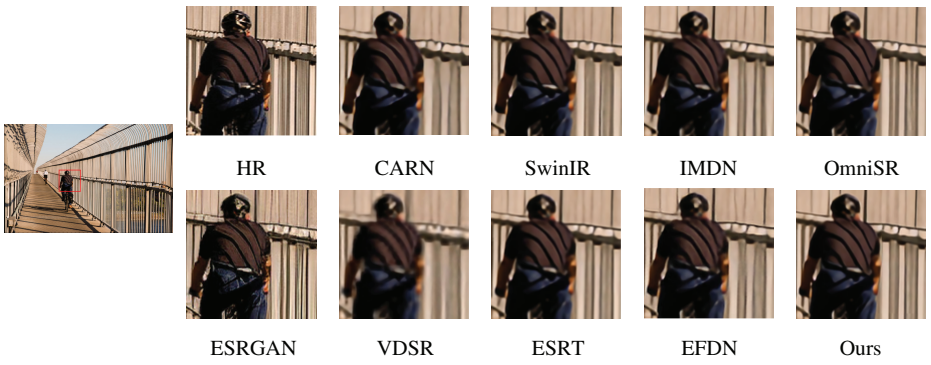

Figure 9: Outdoor pedestrian bridge SR comparison

Traditional methods such as CARN, IMDN, and VDSR often struggle with high-frequency details during image reconstruction. They exhibit noticeable edge blurring and texture loss, failing to accurately recover dense structural details in the image, such as the edges of building components or intricate architectural textures. Transformer-based methods like SwinIR and OmniSR can enhance the overall structural clarity of images to a certain extent. However, due to computational resource constraints, their reconstruction results still suffer from artifacts and edge misalignment.

In contrast, the BAID method demonstrates superior performance in detail restoration and structural preservation. Leveraging its designed frequency branch modeling and attention mechanisms, BAID can accurately restore clear boundaries and maintain structural continuity when reconstructing detailed regions. The frequency branch effectively recovers high-frequency textures, while the attention sharing and binarization strategies reduce redundant computations, enabling the model to allocate more computational resources to critical regions and thus improving visual perception quality.

Fig. 6 focuses on the metallic curtain wall area of urban skyscrapers, which contains complex high-frequency striped structures. Most comparative methods perform poorly in this region, especially in high-frequency striped areas where edge distortion and frequency loss are prone to occur. Although ESRT and EFDN show slight improvements in overall structure restoration, they still have significant blurring issues. In contrast, BAID, relying on its lightweight frequency branches and high-response attention mechanisms, successfully reconstructs dense and uniform metallic striped structures without notable artifacts or polygonal deformations. This fully demonstrates BAID's exceptional performance in high-frequency detail recovery and texture preservation.

For the modern glass building super-resolution evaluation in Fig. 7, the detailed grid-like structures on the building facades pose a great challenge to image reconstruction methods. BAID stands out by accurately restoring the clear boundaries and fine textures of the glass building structures. Comparative methods often introduce blurring or inaccuracies in these detailed regions. For example, SwinIR and OmniSR may have a certain degree of edge softening, while BAID can maintain the sharpness and natural appearance of the glass building details, thanks to its effective frequency and attention mechanisms.

In the architectural texture super-resolution test shown in Fig. 8, which involves complex and fine-grained textures on building facades, BAID demonstrates its strength in preserving these textures. The method can reconstruct architectural textures with high fidelity, including subtle patterns and details. Other methods may either over-smooth the textures or introduce artifacts. BAID's design allows it to handle these complex textures effectively, ensuring that the reconstructed images retain the original architectural texture characteristics.

In the outdoor pedestrian bridge super-resolution comparison presented in Fig. 9, BAID shows its ability to handle large-scale and complex scene structures. It can accurately restore the details of the bridge, the pedestrians on it, and the surrounding environment. Comparative methods may lose details in either the foreground or background. For instance, VDSR might have blurry representations of the bridge railings, while BAID maintains a good balance, providing clear and natural-looking reconstructed images. This is crucial for scenes that contain both large-scale structures and small-scale human details.

Through visual evaluations on samples from the Urban100 dataset, it is evident that the BAID method outperforms existing mainstream models in fine-detail restoration, structural continuity maintenance, and artifact reduction. It achieves superior reconstruction quality while maintaining lightweight characteristics, making it a promising solution for real-time image enhancement applications on edge devices or mobile platforms.

4.5 Stability Evaluation of BAID

To verify BAID's performance reliability in practical deployment, we conducted 5 independent training-test runs for BAID under ×2, ×3, and ×4 super-resolution scales. Each run used a distinct random seed (123, 456, 789, 1011, 1213) to account for random variables, while core settings remained consistent across runs to avoid setup-induced fluctuations. Below presents BAID's stability results on four benchmark datasets, with mean values aligned to its performance in Table 6 for coherence.

Table 6: Quantitative evaluation of super-resolution performance across different scales and datasets

SR Scale	Metric	Set5 (Mean ± Std)	Set14 (Mean ± Std)	BSD100 (Mean ± Std)	Urban100 (Mean ± Std)
×2	PSNR (dB)	38.12 ± 0.02	33.90 ± 0.02	32.28 ± 0.01	32.82 ± 0.02
	SSIM	0.9610 ± 0.0001	0.9212 ± 0.0001	0.9014 ± 0.0001	0.9348 ± 0.0001
×3	PSNR (dB)	34.52 ± 0.02	30.48 ± 0.02	29.20 ± 0.01	28.62 ± 0.02
	SSIM	0.9285 ± 0.0001	0.8460 ± 0.0001	0.8093 ± 0.0001	0.8628 ± 0.0001
×4	PSNR (dB)	32.34 ± 0.01	29.93 ± 0.02	27.65 ± 0.01	26.35 ± 0.02
	SSIM	0.8958 ± 0.0001	0.8469 ± 0.0001	0.7395 ± 0.0001	0.7955 ± 0.0001

5 Conclusions

To address the challenges of substantial computational and memory overhead in Transformer based super resolution models that limit deployment on resource constrained edge devices, we propose a novel lightweight super-resolution method called BAID, which integrates frequency-aware modeling and binary attention mechanisms within a Transformer framework to effectively balance performance and computational efficiency. Building upon the existing ASID architecture, BAID further introduces a binary attention screening strategy to suppress redundant attention computations while leveraging frequency decoupling paths to enhance the recovery of texture details. Through an attention sharing mechanism, BAID efficiently reuses computationally expensive attention weights, reducing FLOPs while retaining the ability to model long-range dependencies. Extensive experimental results demonstrate that BAID achieves superior or near-state-of-the-art performance on multiple standard datasets while exhibiting significant advantages in parameter count and computational cost, showcasing its strong potential for deployment in resource-constrained devices and practical application value.

Acknowledgement: Not applicable.

Funding Statement: This research is funded by Project of Sichuan Provincial Department of Science and Technology under 2025JDKP0150 and the Fundamental Research Funds for the Central Universities under 25CAFUC03093.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Jiajia Liu, Junyi Lin, Wenxiang Dong, Xuan Zhao, Jianhua Liu, Huiru Li; data collection: Jiajia Liu, Junyi Lin, Wenxiang Dong; analysis and interpretation of results: Jiajia Liu, Junyi Lin, Xuan Zhao, Huiru Li; draft manuscript preparation: Jiajia Liu, Junyi Lin; model optimization (attention sharing mechanism and edge deployment adaptation): Wenxiang Dong, Jianhua Liu; visual validation and perceptual metric analysis: Xuan Zhao; academic norm checking (abbreviation standardization, formula notation): Huiru Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials used to support the findings of this study are available from the corresponding author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Muqeet A, Hwang J, Yang S, Kang J, Kim Y, Bae SH. Multi-attention based ultra lightweight image super-resolution. arXiv:2008.12912. 2020.
2. Liang J, Cao J, Sun G, Zhang K, Gu S, Van Gool L, et al. SwinIR: image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 1833–44.
3. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang MH, et al. Restormer: efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 5728–39.
4. Chen X, Wang X, Zhou J, Qiao Y, Dong C. Activating more pixels in image super-resolution transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada. p. 22367–77.
5. Hui Z, Gao X, Yang Y, Wang X. Lightweight image super-resolution with information multi-distillation network. In: Proceedings of the 27th ACM International Conference on Multimedia (ACM MM); 2019 Oct 21–25; Nice, France. p. 2024–32.
6. Lei X, Zhang W, Cao W. Dvmsr: distillated vision mamba for efficient super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 17–21; Seattle, WA, USA. p. 6536–46.
7. Liu J, Tang J, Wu G. Residual feature distillation network for lightweight image super-resolution. In: ECCV 2020 Workshops; 2020 Aug 23–28; Glasgow, UK. p. 41–55. doi:10.1007/978-3-030-67070-2_2.
8. Chen Y, Zheng H, Chen X, Lin S, Zhang M. Residual-aware attention network for image super-resolution. IEEE Trans Image Process. 2021;30(1):135–40.
9. Wang H, Chen X, Ni B, Liu Y, Liu J. Omni aggregation networks for lightweight image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 18–22; Vancouver, BC, Canada. p. 22378–87.
10. Lin J, Huang Y, Wang L. FDAN: flow-guided deformable alignment network for video super-resolution. arXiv:2105.05640. 2021.
11. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. IEEE Trans Pattern Anal Mach Intell. 2023;45(1):87–110. doi:10.1109/TPAMI.2022.3152247.
12. Park K, Soh JW, Cho NI. Efficient attention-sharing information distillation transformer for lightweight single image super-resolution. arXiv:2501.15774. 2025.
13. Wu J, Xing Y, Yu B, Shao W, Gong K. CDPDNet: integrating text guidance with hybrid vision encoders for medical image segmentation. arXiv:2505.18958. 2025.
14. Fritsche M, Gu S, Timofte R. Frequency separation for real-world super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. p. 3599–608.
15. Xin J, Wang N, Jiang X, Li J, Huang H, Gao X. Binarized neural network for single image super resolution. In: Proceedings of the European Conference on Computer Vision 2020; 2020 Aug 23–28; Glasgow, UK. p. 91–107.
16. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017. doi:10.48550/arXiv.1704.04861.
17. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv:1510.00149. 2016.
18. Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: self-attention with linear complexity. arXiv:2006.04768. 2020.
19. Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, et al. Rethinking attention with performers. arXiv:2009.14794. 2021.
20. Zaheer M, Guruganesh G, Dubey KA, Ainslie J, Alberti C, Ontanon S, et al. BigBird: transformers for longer sequences. Adv Neural Inf Process Syst. 2020;33:17283–97.

21. Li X, Wu J, Lin Z, Liu H, Zhang G. Texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 14–19; Seattle, WA, USA.
22. Mehri A, Behjati P, Carpio D, Sappa AD. SRFormer: efficient yet powerful transformer network for single image super resolution. IEEE Access. 2023;11:121457–69. doi:10.1109/ACCESS.2023.3328229.
23. Cao J, Li Y, Zhang K, Van Gool L. Video super-resolution transformer. arXiv:2106.06847. 2022.
24. Huo W, Zhang X, You S, Zhang Y, Zhang Q, Hu N. STGAN: swin transformer-based GAN to achieve remote sensing image super-resolution reconstruction. Appl Sci. 2025;15(1):305. doi:10.3390/app15010305.
25. Xin J, Wang N, Jiang X, Li J, Gao X. Advanced binary neural network for single image super resolution. Int J Comput Vis. 2023;131(7):1808–24. doi:10.1007/s11263-023-01789-8.
26. Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 26–Jul 1; Las Vegas, NV, USA. p. 1646–54.
27. Lepcha DC, Goyal B, Dogra A, Goyal V. Image super-resolution: a comprehensive review, recent trends, challenges and applications. Inf Fusion. 2023;91(1):230–60. doi:10.1016/j.inffus.2022.10.007.
28. Kong F, Li M, Liu S, Liu D, He J, Bai Y, et al. Residual local feature network for efficient super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2022 Jun 21–24; New Orleans, LA, USA. p. 41–55.
29. Umirzakova S, Muksimova S, Mardieva S, Sultanov Baxtiyarovich M, Cho YI. MIRA-CAP: memory-integrated retrieval-augmented captioning for state-of-the-art image and video captioning. Sensors. 2024;24(24):8013. doi:10.3390/s24248013.
30. Park K, Soh JW, Cho NI. A dynamic residual self-attention network for lightweight single image super-resolution. IEEE Trans Multimed. 2021;25:907–18. doi:10.1109/tmm.2021.3134172.
31. Luo X. LatticeNet: towards lightweight image super-resolution with lattice block. In: Proceedings of the European Conference on Computer Vision (ECCV); 2020 Aug 23–29; Glasgow, UK. p. 272–89.
32. Zhang A. Lightweight image super-resolution with superpixel token interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 2–6; Paris, France. p. 12728–37.
33. Zhang X, Zeng H, Guo S, Zhang L. Efficient long-range attention network for image super-resolution. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022 Oct 29–27; Tel Aviv, Israel. p. 649–67.