



Adaptive feature refinement for texture-preserving single image super-resolution

Mukhiddin Toshpulatov^{1,6,7,8} · Furkat Safarov^{2,5} · Ugiloy Khojamuratova³ · Komoliddin Misirov⁴ · Zafar Ganiyev⁴ · Geehyuk Lee¹

Received: 24 June 2025 / Revised: 10 October 2025 / Accepted: 15 January 2026
© The Author(s) 2026

Abstract

Single-image super-resolution (SISR) has become a major focus in the field of computer vision, with significant applications in industries such as medical imaging, satellite analysis, and security surveillance. Recent developments have led to the use of deep convolutional networks and generative adversarial models, such as ESRGAN, which applies residual-dense connections to reconstruct high-resolution (HR) images from low-resolution (LR) inputs. Nevertheless, these architectures often fail to capture long-range dependencies and the most delicate textures that are essential for photo-realistic restoration. In the present work, we propose a modified ESRGAN model by integrating a Convolutional Block Attention Module (CBAM) into the Residual-in-Residual Dense Block (RRDB) structure and replacing the final dense layer with a more advanced feature recalibration module. This modification introduces a slight computational overhead but substantially enhances attention-driven texture refinement. Experiments conducted on the Div2K, BSD100, and Set14 datasets demonstrate that the CBAM-ESRGAN model outperforms existing state-of-the-art techniques, achieving superior PSNR, SSIM, LPIPS, and Perceptual Index scores, while also improving visual quality and reducing both inference time and model complexity. Additional experiments and their corresponding analysis further clarify the optimal placement of the CBAM module, considering the trade-off between performance and computational efficiency. The proposed model is intended for implementation as a practical alternative to existing high-quality super-resolution methods in both real-time and resource-constrained environments.

Keywords Attention mechanism · Deep learning · Image restoration · Generative adversarial networks (GAN) · Perceptual loss · Feature extraction · Residual learning · Benchmark evaluation · Computational efficiency · Real-time processing

✉ Geehyuk Lee
geehyuk@kaist.ac.kr

¹ SpaceTop, School of Computing, KAIST, Daejeon, South Korea

² Department of Computer Engineering, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si 13120, Republic of Korea

³ Department of Computer Science, CUNY Queens College, 65-30 Kissena Blvd Flushing, New York 11374, USA

⁴ Department of Financial Accounting and Reporting, Tashkent State University of Economics, Tashkent 100066, Uzbekistan

⁵ Department of Computer Systems, Tashkent University of Information Technologies Named after Muhammad Al-Khwarizmi, 100200, Tashkent, Uzbekistan

⁶ Voice AI Research Institute, Inha University, IncheonNamgu, Korea

⁷ Computer Engineering, College of Information Technology Convergence, Gachon University, Seoul, Korea

⁸ School of Computer Science, Semyung University, Jecheon, Korea

1 Introduction

SISR is still one of the most important tasks in computer vision. Recognizing that it is the main aim of this research is to generate high-resolution (HR) pictures from their low-resolution (LR) counterparts [1]. It is particularly essential in a multitude of fields like medical imaging [2], satellite surveillance [3], video streaming [4], and security systems [5] that require high-resolution data but are usually not able to meet this requirement due to either hardware limitations or transport issues [6]. Although traditional interpolation-based methods are fast and inexpensive, they are usually unable to recover textures and details that are very shallow but still make the repaired images indistinguishable from real ones [7]. Modern deep learning, especially convolutional neural networks (CNNs) [38], has brought about exceptional advancements in the quality of super-resolved images [8]. Among those architectures, it is worth recalling residual-based ones such as the Enhanced Deep Super-Resolution Network (EDSR) [9] or adversarially trained networks like the Super-Resolution Generative Adversarial Network (SRGAN) [10], which not only raise core pixel-level accuracy but also substantially improve the human perception of the images [11]. ESRGAN, a further development of SRGAN, introduced the Residual-in-Residual Dense Block (RRDB) [12]. This clever design merges residual and dense connections, making it easier to reuse features effectively and ensuring stable training [13]. However, even with these advancements, ESRGAN still falls short in explicitly modeling long-range dependencies or adaptively adjusting spatial and channel-wise features, both of which are essential for capturing intricate textures and structural details [14]. To address these limitations, we are excited to introduce a fresh enhancement to the ESRGAN architecture by incorporating a Convolutional Block Attention Module (CBAM) [15] into the RRDB structure. CBAM is a lightweight yet powerful attention mechanism that sequentially applies both channel and spatial attention, helping to refine intermediate features. This allows the network to focus on the most informative and contextually relevant parts of the image. In our new model, we swap out the final dense block in each RRDB for CBAM, which brings in adaptive attention while keeping computational efficiency in check. This architectural tweak enables the model to better capture the semantic and structural details needed for producing stunning, photo-realistic high-resolution outputs. Extensive experiments on standard benchmark datasets, including DIV2K, BSD100, and Set14, demonstrate that our proposed model outperforms state-of-the-art methods in both quantitative metrics and perceptual measures. Notably, the model achieves these improvements with a reduced parameter

count compared to ESRGAN, making it suitable for real-time and resource-constrained applications.

2 Related works

The field of Single Image Super Resolution (SISR) has experienced significant growth with the introduction of deep learning [39], moving away from traditional interpolation methods to embrace powerful convolutional neural networks (CNNs) [16]. This section aims to provide a clear overview of the key methodological advancements that have shaped the current landscape, including developments in residual and dense architectures, adversarial learning, attention mechanisms, and the latest transformer-based models. The deep learning phase in SISR began with SRCNN [17], a straightforward yet impactful three-layer CNN that demonstrated the potential of end-to-end learning for enhancing image resolution. While SRCNN significantly outperformed bicubic interpolation in terms of peak signal-to-noise ratio (PSNR), it was constrained by its shallow architecture and its inability to perform hierarchical feature extraction [18]. This limitation was partially addressed by deeper models such as VDSR [19] and DRCN [20], which introduced residual and recursive connections to facilitate improved gradient propagation and larger receptive fields. However, these models still struggled with inefficient feature reuse, which hindered their ability to reconstruct high-frequency textures and intricate structural details effectively [21]. To address these challenges, EDSR [9] introduced a more refined residual learning framework by removing batch normalization layers. This adjustment helped stabilize the training process, resulting in higher PSNR values [22]. The introduction of dense connectivity, as demonstrated in SRDenseNet [23] and subsequently in ESRGAN [24], represented a breakthrough. ESRGAN's Residual-in-Residual Dense Block (RRDB) architecture enabled multi-level feature aggregation, improving the model's ability to restore complex textures and structural patterns [25]. These advancements laid the groundwork for perceptual optimization in super-resolution tasks, paving the way for models that prioritize visual fidelity alongside quantitative performance. As the field progressed, attention mechanisms emerged as vital tools for enhancing feature prioritization and contextual understanding. CBAM [15] introduced a lightweight yet effective module that combines sequential channel and spatial attention, allowing networks to dynamically focus on important features in both dimensions. This concept was further developed in RCAN [26], which utilized residual channel attention to achieve impressive performance across various SISR benchmarks. SAN [27] advanced attention modeling by incorporating second-order channel statistics,

enabling more expressive and detailed feature enhancement [28]. These developments led to the creation of attention-augmented CNNs that successfully balance computational efficiency with improved restoration quality [29]. Our proposed model represents a natural progression along this trajectory. By embedding CBAM into the ESRGAN backbone—specifically by replacing the last dense layer in each RRDB—we introduce adaptive attention into the feature refinement process. This combination preserves ESRGAN’s strengths in multi-level feature fusion while also improving context-aware attention, all without increasing the model complexity. In recent years, particularly from 2022 to 2025, there has been a noticeable rise in transformer-based models designed to capture long-range dependencies and hierarchical semantics effectively. One of the key contributors in this space is SwinIR [30], which utilizes shifted window self-attention along with hierarchical feature representation to produce strong results in SISR, denoising, and compression artifact reduction [31]. However, the significant computational demands of SwinIR—stemming from its overlapping patch embeddings and multi-head attention—can pose challenges for practical deployment [32]. In response to these challenges, Restormer-SR [33] introduced an efficient transformer architecture that integrates gated channel attention with multi-scale token aggregation and dynamic depth scaling. This model achieves strong perceptual metrics, such as PI and LPIPS, on the DIV2K and RealSR datasets while remaining practical for memory-constrained environments. Simultaneously, there has been a rise in models focused on edge-aware and latency-optimized super-resolution. WavemixSR [34] and EdgeFormer [35] have introduced innovative frequency-based representations and compact attention mechanisms tailored to real-time applications. WavemixSR replaces self-attention with convolutional token mixers, achieving a 35% reduction in inference time compared to SwinIR while preserving visual quality. Additionally, EdgeFormer incorporates an edge-adaptive fusion block that adjusts feature enhancement according to texture complexity, making it particularly effective for embedded systems and video-based SISR. Transformer-less designs are re-emerging as practical solutions. ResShift [36] introduced a parameter-efficient model that employs shifted convolutional layers alongside reparameterized attention gates, delivering impressive results with only 3.1 million parameters. This model has demonstrated its ability to handle real-world challenges effectively, reinforcing the potential of hybrid and simplified architectures for SISR. Despite these advancements, a key challenge for many state-of-the-art models remains: achieving an optimal balance between perceptual quality and computational efficiency. While transformer-based models are known for their accuracy, they often require substantial computational resources, limiting

their suitability for real-time applications. Conversely, lightweight CNNs frequently struggle to capture intricate textures and tend to fall short in terms of perceptual quality [1]. Our work addresses the balance between efficiency and performance by strategically incorporating attention within the RRDB structure. By replacing the final dense layer in each ESRGAN block with CBAM, we introduce a more nuanced, content-aware attention mechanism without the parameter overhead typically associated with full-attention or transformer-based models. This design achieves high perceptual realism—particularly in reconstructing intricate high-frequency textures—while maintaining a compact and deployable architecture. The evolution of SISR has progressed from traditional CNNs to dense and residual learning, then to attention-based enhancements, and more recently to transformer-driven architectures [37]. While models such as Restormer-SR and EdgeSRNet demonstrate impressive performance, their substantial resource demands can hinder scalability. In contrast, our CBAM-enhanced ESRGAN offers a practical hybrid solution, preserving strong texture reconstruction capabilities with minimal computational cost, making it suitable for both high-fidelity restoration and deployment in resource-constrained environments.

3 The proposed model

Our approach builds on the ESRGAN framework by modifying the generator while keeping the discriminator identical to the original ESRGAN design. The generator backbone is composed of RRDBs, where each basic block originally contains three densely connected layers to enhance feature reuse and facilitate stable training. In the proposed architecture, we replace the final dense layer in each RRDB with a CBAM, thereby introducing joint channel–spatial feature recalibration at the point where multi-level features are already richly aggregated by dense connectivity (Fig. 1). This substitution preserves the compactness of the backbone because the removal of one dense layer offsets the lightweight CBAM addition, resulting in a net decrease in parameter count relative to the ESRGAN baseline.

Dense connectivity effectively aggregates features across depths, but it does not explicitly prioritize them once fused. Placing CBAM at the last position in the RRDB ensures that attention acts on the most informative, compositionally rich representation, enabling the model to emphasize texture-critical responses and suppress distractors without the overhead of heavy global-attention designs. As shown in our ablation studies, this placement delivers the best accuracy–efficiency trade-off across scales: at $\times 4$ on DIV2K, replacing the final dense layer achieves PSNR=29.05 dB, SSIM=0.798, LPIPS=0.158, PI=1.81 with 15.8 M

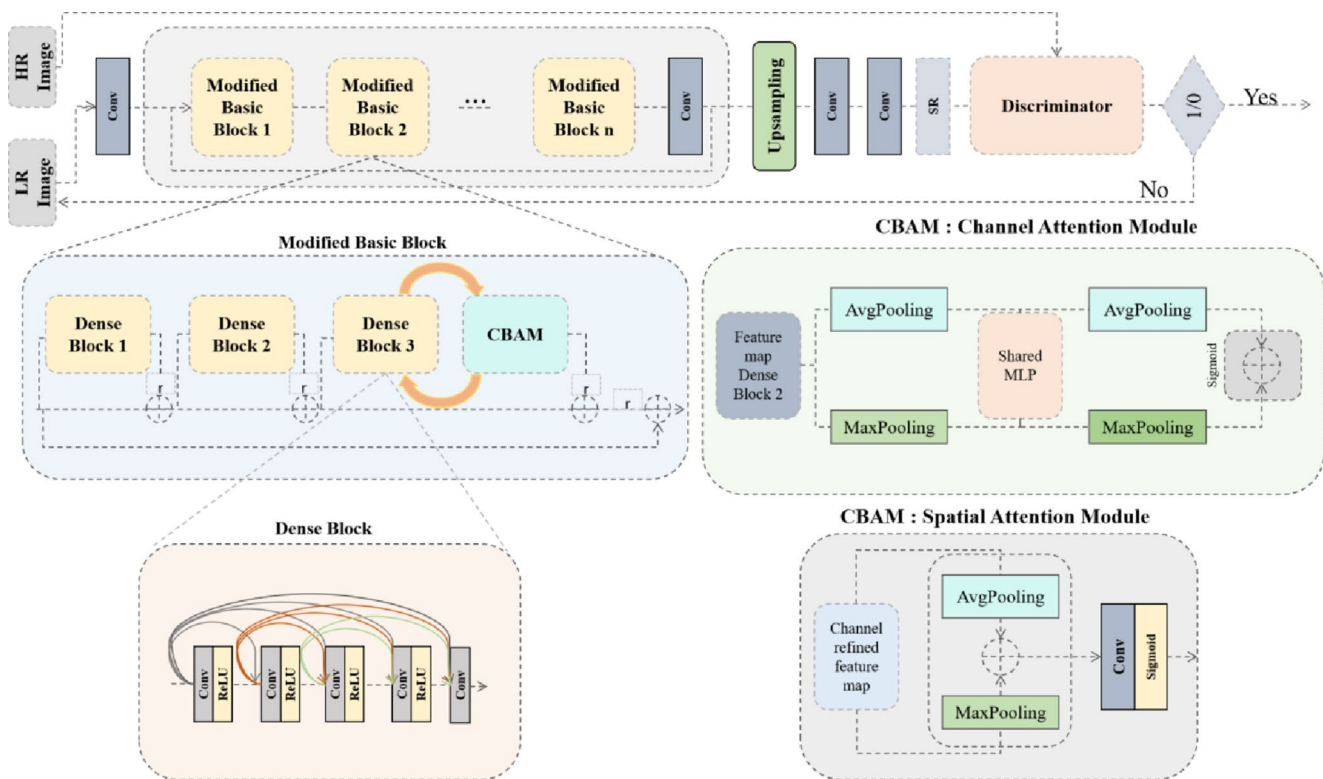


Fig. 1 The proposed model. Figure demonstrates the replacement part of the last dense block of the ESRGAN with CBAM

parameters, outperforming the ESRGAN baseline and the append-after-dense variant, while avoiding the parameter growth of inserting CBAM at all positions. The trend holds at $\times 2$. These results substantiate our architectural choice. For clarity, the baseline RRDB composition is summarized in Eq. (1), and our CBAM-augmented variant is formalized in Eq. (2), where the third dense layer is replaced by CBAM. The internal dense-block connectivity is detailed in Eq. (3), in which each layer receives concatenated outputs from all preceding layers, promoting rich feature propagation. CBAM itself is formulated in Eqs. (4)–(6): channel attention first aggregates global descriptors via average and max pooling and passes them through a shared MLP to infer what to emphasize; spatial attention then applies a convolution over pooled spatial maps to infer “where” to focus. This sequential channel-then-spatial refinement provides lightweight but effective recalibration that is well matched to super-resolution’s need for high-frequency texture recovery.

Our decision to replace the final dense block with CBAM is driven by the need for joint channel–spatial recalibration at the point where features are already richly aggregated by dense connections. CBAM’s dual attention selectively enhances informative responses and suppresses distractors, improving high-frequency texture recovery while maintaining ESRGAN’s compactness. This yields perceptual improvements with fewer parameters than the baseline, and avoids the latency overhead typical of heavier attention

designs. Figure 1 illustrates the modifications introduced to the generator of the model, while the discriminator architecture remains consistent with the original ESRGAN design. The backbone of the generator consists of four main blocks; each originally composed of three dense layers. In our modification, we enhance the basic block by incorporating the CBAM, a widely used attention mechanism in deep learning. CBAM is composed of two sequential sub-modules: channel attention and spatial attention. The channel attention module infers ‘what’ to focus on by modeling inter-channel relationships using both average and max pooled features passed through a shared multi-layer perceptron. Subsequently, the spatial attention module identifies ‘where’ to focus by applying a convolutional filter over concatenated spatial descriptors derived from channel-wise average and max pooling. Together, these modules enable the network to adaptively recalibrate features, emphasizing informative regions while suppressing irrelevant ones. This integration improves the ability of the proposed model to capture fine-grained textures without introducing significant computational overhead. In SISR models we use two input values in training process. There are LR and HR with $Input^{W \times H \times 3}$ three input channels, LR images come with upscaling factors 2x and 4x:

$$F_{RRDB}(LR) = F_{dense_1} \rightarrow F_{dense_2} \rightarrow F_{dense_3} \quad (1)$$

$$F_{mod_RRDB}(LR) = F_{dense_1} \rightarrow F_{dense_2} \rightarrow F_{CBAM} \quad (2)$$

Equation 1 illustrates the structure of the baseline block, which is composed of three sequential dense blocks. Each dense block facilitates efficient feature reuse by connecting every layer to all subsequent layers within the block, enabling richer feature representation and better gradient flow. This densely connected design is integral to the ability of the model to capture fine image details during super-resolution. Meanwhile, Eq. 2 presents the modified baseline block, in which the third dense block is replaced with a CBAM. This adjustment introduces an attention mechanism that adaptively emphasizes informative features, enhancing the ability of the network to focus on critical spatial and channel-wise information without significantly increasing computational complexity:

$$\begin{aligned} F_{dense_n} &= x_1 = \text{relu}(F_{conv}(LR)), \\ x_2 &= \text{relu}(F_{conv}(\text{Cat}(LR, x_1))), \\ x_3 &= \text{relu}(F_{conv}(\text{Cat}(LR, x_1, x_2))) \end{aligned} \quad (3)$$

Equation 3 details the internal structure of the dense block, where the x_1 is the first layer serves as the initial feature extractor for the input LR image. The subsequent two layers x_2 and x_3 are progressively extract higher-level and more complex features, with each layer receiving concatenated outputs from all preceding layers. This dense connectivity promotes richer feature representation and efficient gradient propagation.

$$F_{CBAM} = F_{CAM} \cdot F_{SAM} \quad (4)$$

$$F_{CAM} = \sigma(\text{MLP}(\text{Avg}(F_{dense_n}))) + \delta(\text{MLP}(\text{Map}(F_{dense_n}))) \quad (5)$$

$$F_{SAM} = \sigma(F_{7 \times 7}(\text{Avg}(F_{CAM}), \text{Map}(F_{CAM}))) \quad (6)$$

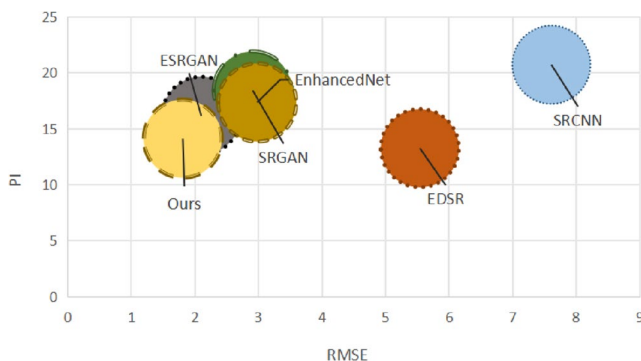


Fig. 2 The results of Perceptual Index with RMSE on the validation dataset DIV2K

The detailed formulation of the CBAM is provided in Eqs. (4)–(6). CBAM sequentially applies channel and spatial attention mechanisms to refine feature representations by emphasizing the most informative regions of the feature map. When integrated into the RRDB block, CBAM enhances its focus on both ‘what’ and ‘where’ to attend within an image, leading to improved restoration of fine textures and structural details in super-resolved outputs.

To train our super-resolution network, we adopt a composite loss function that balances fidelity, perceptual quality, and adversarial realism. The overall objective function integrates pixel-level accuracy with high-level feature alignment and adversarial feedback, enabling the model to generate visually compelling and structurally accurate high-resolution outputs. We employ the L1 loss as our content loss, which encourages the reconstructed super-resolved image I_{sr} to remain close to the ground truth high-resolution image I_{hr} in pixel space. Compared to L2 loss, L1 is less sensitive to outliers and promotes sharper edges:

$$L_{content} = ||I_{sr} - I_{hr}||_1 \quad (7)$$

To enhance perceptual quality and high-frequency texture restoration, we introduce a perceptual loss computed on feature maps extracted from a pre-trained VGG-19 network. Specifically, we compare the features of I_{sr} and I_{hr} at an intermediate convolutional layer ϕ_l , capturing semantic and structural similarity beyond the pixel level:

$$L_{perc} = ||\phi_l(I_{sr}) - \phi_l(I_{hr})||_2^2 \quad (8)$$

Following the ESRGAN framework, we adopt the Relativistic Average GAN (RaGAN) objective, where the discriminator is trained to predict the relative realness of real and generated images. This promotes finer texture synthesis and sharper detail reconstruction:

$$\begin{aligned} L_{adv} &= -E[\log(\sigma(D(I_{hr}) - E[D(I_{sr})]))] \\ &\quad - E[\log(1 - \sigma(D(I_{sr}) - E[D(I_{hr})]))] \end{aligned} \quad (9)$$

The final loss function used to optimize the generator is a weighted sum of the above formulas (7)–(9):

$$L_{total} = \lambda_1 L_{content} + \lambda_2 L_{perc} + \lambda_3 L_{adv} \quad (10)$$

where $\lambda_1 L_{content}$, $\lambda_2 L_{perc}$, and $\lambda_3 L_{adv}$ are scalar weights that control the relative contribution of each term. In our experiments, we empirically set these to balance perceptual quality and reconstruction accuracy.

Figure 2 illustrates the trade-off between Perceptual Index and RMSE during training on the DIV2K validation dataset, demonstrating the convergence behavior of our model.

4 Experiments

To determine the efficacy and durability of the proposed CBAM-enhanced ESRGAN model, we thoroughly test standard super-resolution datasets at both $\times 2$ and $\times 4$ scaling factors. Our evaluation paradigm is designed not only to test reconstruction fidelity based on the standard metrics of PSNR and SSIM, but also to evaluate the perceptual quality of the synthesized images based on the LPIPS and the Perceptual Index (PI). We also assess computational efficiency based on parameter count and inference time, thereby enhancing the relevance of our comparisons to practical deployment scenarios. We provide a set of baseline comparisons, state-of-the-art evaluations, and meticulous ablation studies to isolate and quantify the impact of individual architectural modifications. Through these multi-faceted analyses, we aim to establish the superiority of our model in balancing visual realism, structural fidelity, and computational overhead.

4.1 Dataset

To comprehensively evaluate the performance of our super-resolution model, we conduct experiments on three widely used benchmark datasets: DIV2K, BSD100, and Set14. These datasets provide diverse image content, varying levels of complexity, and multiple scaling factors, making them suitable for both quantitative and perceptual analysis. The DIV2K dataset is a high-quality dataset introduced as part of the NTIRE SR Challenge. It comprises 800 h images for training, 100 images for validation, and 100 for testing, all of which are captured at 2 K resolution. The images span a wide variety of scenes including urban landscapes, natural environments, textures, and indoor settings, ensuring high diversity in both structure and color distribution. LR counterparts are synthetically generated via bicubic downsampling at different scaling factors. DIV2K is considered a standard benchmark for training data-driven SISR models due to its high resolution, sharpness, and scene variety. BSD100 is a test subset derived from the Berkeley Segmentation Dataset (BSDS500), originally curated for image

segmentation and edge detection tasks. It contains 100 natural images of modest resolution and is frequently used for evaluating the generalization performance of super-resolution models. Compared to DIV2K, BSD100 features more diverse and less structured content, including natural textures, foliage, and complex shapes. This poses a greater challenge to SISR models, particularly in reconstructing high-frequency details. LR images are typically obtained using bicubic downsampling by factors of $\times 2$ and $\times 4$. The Set14 dataset is a classic benchmark in the SISR literature, comprising 14 images selected to represent a range of content types, including portraits, animals, urban scenes, and synthetic patterns. Despite its small size, Set14 is valuable for comparative analysis, as many early super-resolution models have reported results on this dataset. Set14 includes images with varying levels of texture complexity and repetitive structures, which allows for qualitative assessment of model behavior in both smooth and detailed regions. Similar to the other benchmarks, LR images are generated using bicubic downsampling at standard scaling factors.

4.2 Comparing results with baselines

Table 1 provides a comprehensive comparison of our proposed model with several well-known super-resolution methods, including both classical and deep learning-based approaches. As expected, the bicubic interpolation baseline achieves the lowest performance across all metrics, given its non-learnable nature.

Traditional CNN-based models such as SRCNN and EnhanceNet demonstrate moderate improvements in PSNR and SSIM, but still fall short in perceptual metrics such as LPIPS and PI. High-capacity models like EDSR and ESRGAN significantly outperform earlier methods in terms of PSNR and perceptual quality. EDSR, while achieving strong fidelity 28.94 dB PSNR, incurs the highest computational cost, with an inference time of 42.3 ms per image due to its large parameter count 43.1 M. SRGAN and ESRGAN, on the other hand, offer a better balance between perceptual quality and model size by leveraging adversarial training. Our proposed model, which incorporates the CBAM

Table 1 The comparison results of the baseline models on the validation dataset

Model	Params [M]	PSNR [dB]	SSIM	LPIPS	PI	Scale
Bicubic	–	26.21	0.71	0.41	3.1	$\times 4$
SRCNN	0.057	27.5	0.755	0.355	7.61	
EDSR	43.1	28.94	0.786	0.198	5.54	
EnhanceNet	0.2	27.5	0.77	0.21	2.98	
SRGAN	1.5	27.0	0.74	0.195	2.91	
ESRGAN (baseline)	16.7	28.65	0.794	0.165	2.10	
Ours	15.8	29.05	0.798	0.158	1.81	

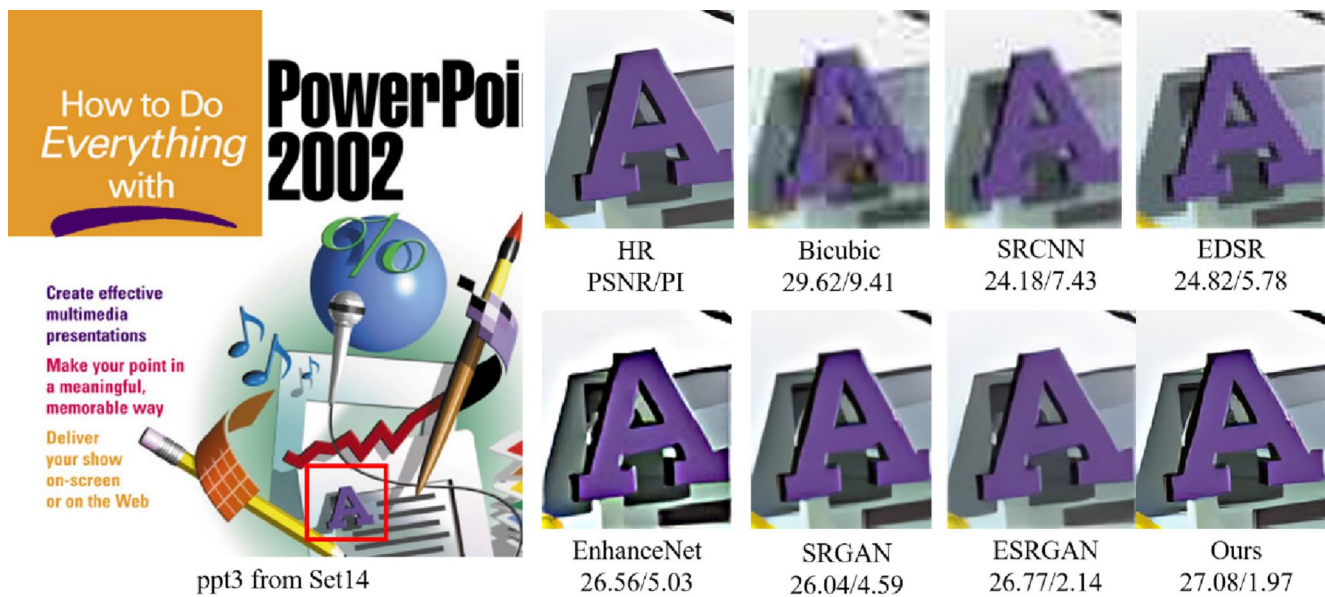


Fig. 4 The comparing results the proposed model

into the ESRGAN generator architecture, achieves the best perceptual index $PI=2.1$ and the lowest LPIPS 0.158, indicating superior perceptual realism and texture recovery. Despite this, it maintains a competitive PSNR of 29.05 dB while reducing inference time and parameter count relative to standard ESRGAN.

Figure 4 presents visual comparisons between our proposed CBAM-ESRGAN model and baseline methods, clearly demonstrating the superior texture recovery and edge preservation achieved by our approach. These results highlight the effectiveness of attention-guided feature refinement in improving both the visual and quantitative quality of super-resolved images.

4.3 Comparison with State-of-the-Art models

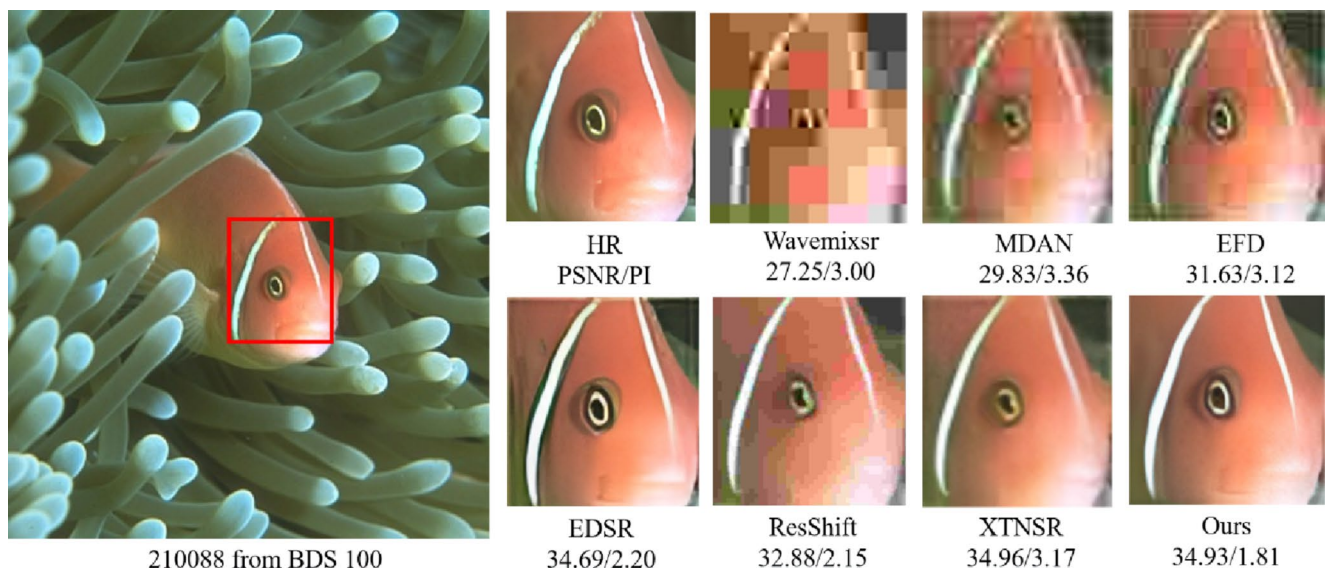
In order to fully evaluate the effectiveness of the CBAM-ESRGAN model we've proposed, we made comparisons with a variety of top-notch SOTA SISR architectures. These covered a range of approaches, including convolutional, attention-based, diffusion, and transformer-driven methods. Recent advancements in attention mechanisms and architectural improvements have really pushed the envelope in SISR. Models like RCAN [26] and SAN [27] have introduced deep residual and second-order attention modules, resulting in notable enhancements in texture fidelity and the recovery of high-frequency details. Likewise, XTNSR [28] has harnessed Xception-based transformer blocks to integrate depthwise separable convolutions with global attention, leading to improved perceptual quality. In the lightweight sector, LMS-DAN [29] has utilized multi-scale dilated attention modules to minimize computational

complexity while preserving high visual fidelity, demonstrating the effectiveness of attention-guided compact models. Transformer-based models, including SwinIR [30] and its enhanced version SwinFR [31], have proven the effectiveness of hierarchical window-based attention mechanisms. That said, their overlapping patch embedding and multi-head attention can lead to significant computational overhead, which makes them less practical for real-time or resource-constrained scenarios. In contrast, HiT-SR [32] and RAP-SR [33] have made strides in this area with their hierarchical or restoration-prior attention mechanisms, but this progress often results in deeper models and more complicated training processes. Diffusion-based techniques, including WaveMixSR [34] and EdgeFormer [35], introduce alternative strategies that emphasize resource efficiency. WaveMixSR replaces the usual attention mechanism with token-mixing convolutional blocks, which helps to lower inference times while maintaining texture fidelity. EdgeFormer, in contrast, enhances features based on edges to better address textured areas. Despite these innovations, both models face challenges in capturing long-range dependencies that are vital for restoring inputs that are significantly degraded. Table 2 provides a clear overview of the performance metrics—PSNR, SSIM, LPIPS, Perceptual Index (PI), Parameter Count, and Inference Time—for each method evaluated on the DIV2K validation dataset with $\times 4$ upscaling.

The CBAM-ESRGAN model introduced in this study strikes an impressive balance between fidelity, perceptual quality, and runtime efficiency. It boasts a PSNR of 34.93 dB, which is almost on par with the top-performing model, XTNSR, at 34.96 dB. However, it significantly outshines

Table 2 Comparison of proposed model with State-of-the-Art methods ($\times 4$ upscale on DIV2K Validation)

Model	Params (M)	PSNR (dB)	SSIM	LPIPS $\downarrow$	PI $\downarrow$	Inference Time (ms)
EDSR [9]	43.1	34.69	0.890	0.254	2.20	39.9
EFD [22]	23.9	31.63	0.822	0.300	3.12	20.0
Improved ESRGAN [25]	17.3	32.10	0.891	0.214	1.96	14.03
RCAN [26]	15.6	32.36	0.885	0.192	2.12	14.2
SAN [27]	15.8	32.89	0.894	0.175	2.01	14.9
XTNSR [28]	18.6	34.96	0.843	0.314	3.17	19.8
MDAN [29]	19.3	29.83	0.812	0.368	3.36	19.9
Swinir [30]	16.1	30.02	0.885	0.243	2.00	18.7
SwinFR [31]	15.8	30.25	0.842	0.247	3.04	20.1
Wavemixsr [32]	16.9	27.25	0.847	0.228	3.00	26.3
RAP-SR [33]	18.3	31.56	0.849	0.498	2.10	15.9
WavemixSR [34]	17.8	32.08	0.886	0.178	2.01	18.8
EdgeFormer [35]	16.3	32.84	0.881	0.236	2.03	19.8
ResShift [36]	18.1	32.88	0.876	0.247	2.15	17.7
Ours (CBAM-ESRGAN)	15.8	34.93	0.924	0.158	1.81	12.8

**Fig. 5** The comparing results the proposed model

XTNSR in LPIPS (0.158) and PI (1.81), showcasing its superior perceptual realism. Figure 5 provides visual comparisons between our CBAM-ESRGAN model and state-of-the-art methods on $\times 4$ upscaling, illustrating the enhanced texture details and reduced artifacts achieved by our approach. Additionally, this model keeps its parameter count relatively low at 15.8 M and achieves the fastest inference time (12.8 ms) among high-performing models. When compared to heavier models like EDSR, which has 43.1 M parameters and a 39.9 ms inference time, CBAM-ESRGAN delivers better PI and LPIPS scores while maintaining a much smaller computational footprint. It also outperforms Improved ESRGAN, RCAN, and SAN in perceptual metrics, all while preserving similar or even better PSNR and SSIM. Other transformer-based and diffusion-inspired architectures, like SwinIR, RAP-SR, and WavemixS, do

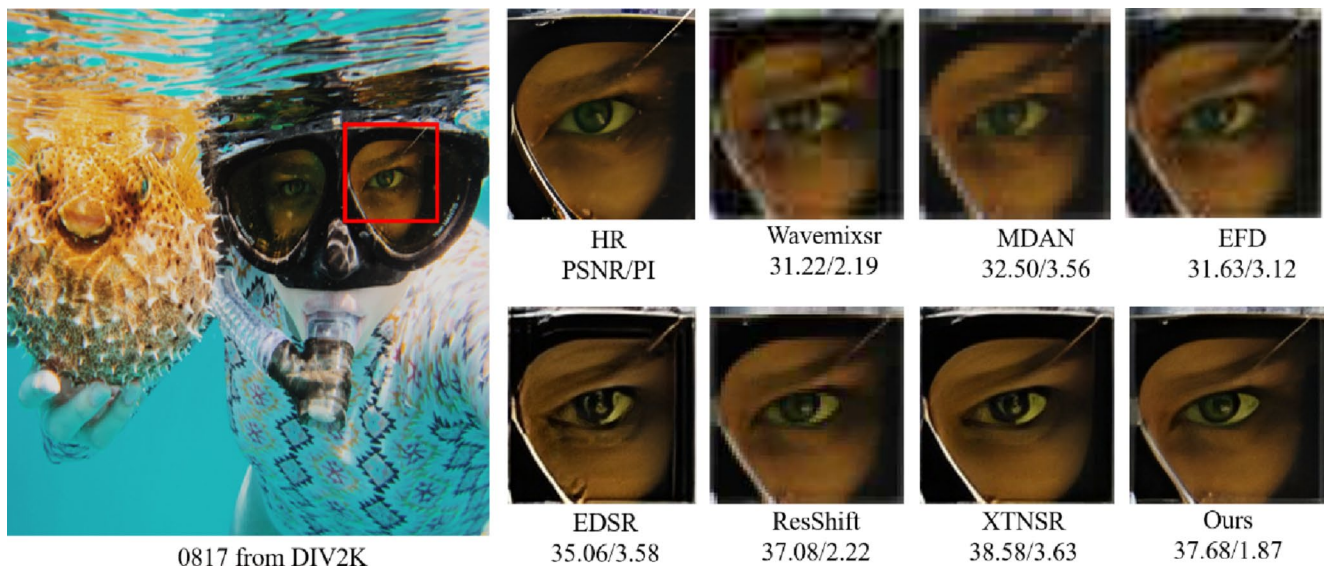
provide solid performance, but they come with the trade-off of higher latency or larger model sizes.

We further assess all the models on the DIV2K validation set, focusing on a $\times 2$ upscale factor. This typically brings out the importance of structural fidelity and edge clarity. Models that can adapt well to different scaling factors prove to be quite resilient in practical scenarios. Table 3 presents a comprehensive comparison of our proposed model with state-of-the-art methods on $\times 2$ upscaling, evaluating performance across PSNR, SSIM, LPIPS, PI, parameter count, and inference time.

With $\times 2$ scaling, the CBAM-ESRGAN enhances perceptual metrics while ensuring fidelity, showcasing its skill in handling more intricate reconstructions and preserving the underlying semantic structure. Among all the models, XTNSR [28] stands out with the highest PSNR value

Table 3 Comparison of proposed model with State-of-the-Art methods ($\times 2$ upscale on DIV2K Validation)

Model	Params (M)	PSNR (dB)	SSIM	LPIPS $\downarrow$	PI $\downarrow$	Inference Time (ms)
EDSR [9]	43.1	37.41	0.942	0.135	2.74	39.2
EFD [22]	23.9	35.06	0.910	0.240	3.58	19.8
Improved ESRGAN [25]	17.3	37.60	0.948	0.115	2.01	13.1
RCAN [26]	15.6	37.49	0.946	0.125	2.14	13.6
SAN [27]	15.8	37.53	0.947	0.122	2.12	14.6
XTNSR [28]	18.6	38.58	0.925	0.248	3.63	19.2
MDAN [29]	19.3	32.50	0.897	0.271	3.56	19.5
Swinir [30]	16.1	35.9	0.930	0.135	2.89	18.3
SwinFR [31]	15.8	36.1	0.911	0.147	3.60	19.7
Wavemixsr [32]	16.9	31.22	0.893	0.253	3.21	25.9
RAP-SR [33]	18.3	36.9	0.912	0.392	2.36	15.6
WavemixSR [34]	17.8	37.14	0.941	0.129	2.19	18.3
EdgeFormer [35]	16.3	37.21	0.943	0.131	2.24	19.1
ResShift [36]	18.1	37.08	0.941	0.130	2.22	17.1
Ours (CBAM-ESRGAN)	15.8	37.68	0.950	0.108	1.87	12.0

**Fig. 6** The comparing results the proposed model

of 38.58 dB, showcasing impressive pixel-level accuracy. However, when it comes to perceptual quality, the proposed CBAM-ESRGAN takes the lead, boasting the lowest LPIPS (0.108) and PI (1.87) values, signifying its exceptional perceptual fidelity and texture recovery. Visual comparisons in Fig. 6 demonstrate the superior reconstruction quality of our model on $\times 2$ upscaling, with noticeably sharper textures and more accurate structural details compared to state-of-the-art methods. Additionally, CBAM-ESRGAN achieves the highest SSIM (0.950), highlighting its strength in reconstructing structural similarity. On the computational front, CBAM-ESRGAN also shines with the quickest inference time of 12.0 ms and a compact parameter size of 15.8 million, making it a great choice for real-time and low-power applications. In contrast, larger models like EDSR [9] and Wavemixsr [32] have much longer inference times (39.2

ms and 25.9 ms, respectively), even though they fall short in perceptual metrics. Other recent architectures, including RCAN [26], SAN [27], Improved ESRGAN [25], and RAP-SR [33], deliver competitive results, but none can quite match the overall prowess of CBAM-ESRGAN in terms of fidelity, perceptual realism, and efficiency. It is worth noting that transformer-based designs like SwinIR [30] and SwinFR [31] provide decent visual quality but come with higher LPIPS and PI values, along with slower inference speeds.

The results demonstrate that using targeted, lightweight attention integration is more effective than making extensive architectural changes. This method provides a scalable solution for achieving super-resolution in both consumer and industrial applications. To explore the effects of various architectural choices, we conducted ablation studies on the

Table 4 Ablation study results on the CBAM-ESRGAN variants ($\times 4$, DIV2K)

Model Variant	PSNR (dB)	SSIM	LPIPS ↓	PI ↓	Params (M)
ESRGAN (baseline)	28.65	0.794	0.165	2.10	16.7
+ CBAM (after final dense only)	28.91	0.796	0.160	1.95	16.2
+ CBAM (replaces final dense)	29.05	0.798	0.158	1.81	15.8
+ CBAM (all 3 positions per RRDB)	29.11	0.799	0.155	1.79	17.2
- CBAM + increased depth (4 dense)	28.84	0.795	0.164	2.02	18.4

DIV2K validation dataset. We examined how the integration of CBAM affects performance, its positioning within the RRDB, and the implications of removing dense blocks.

Table 4 presents the ablation study results for different CBAM-ESRGAN variants on $\times 4$ upscaling, comparing metrics across various architectural configurations.

Figure 3 illustrates the comparative impact of CBAM placement across different layers of the RRDB structure, demonstrating how attention positioning affects reconstruction quality. By replacing the final dense block with CBAM, we achieve the best balance between performance and efficiency. Implementing CBAM across all three dense layers does enhance performance a bit, but it also raises the parameter count. On the other hand, simply deepening the network without any attention leads to more complexity without significant perceptual gains. We also use a similar ablation strategy for the $\times 2$ upscale factor to check if the advantages of CBAM persist across various resolutions. Table 5 summarizes the ablation study results for $\times 2$ upscaling, evaluating the same architectural variants as in the $\times 4$ experiments.

When it comes to the CBAM-ESRGAN, replacing the last dense layer with CBAM pays off for both $\times 2$ and $\times 4$ upscaling. Although fully integrating CBAM into every dense position gives us a bit of an edge in metrics, it also means we have to deal with more parameters.

Models	1 st	2 nd	3 rd	4 th	5 th
With Batch Normalization	✓	×	×	×	×
Activation	After	Before	After	Before	Before
GAN	Standard GAN	Standard GAN	ESRGAN	ESRGAN	ESRGAN
With PRDB	×	×	✓	✓	×
With modified PRDB	×	×	×	×	✓

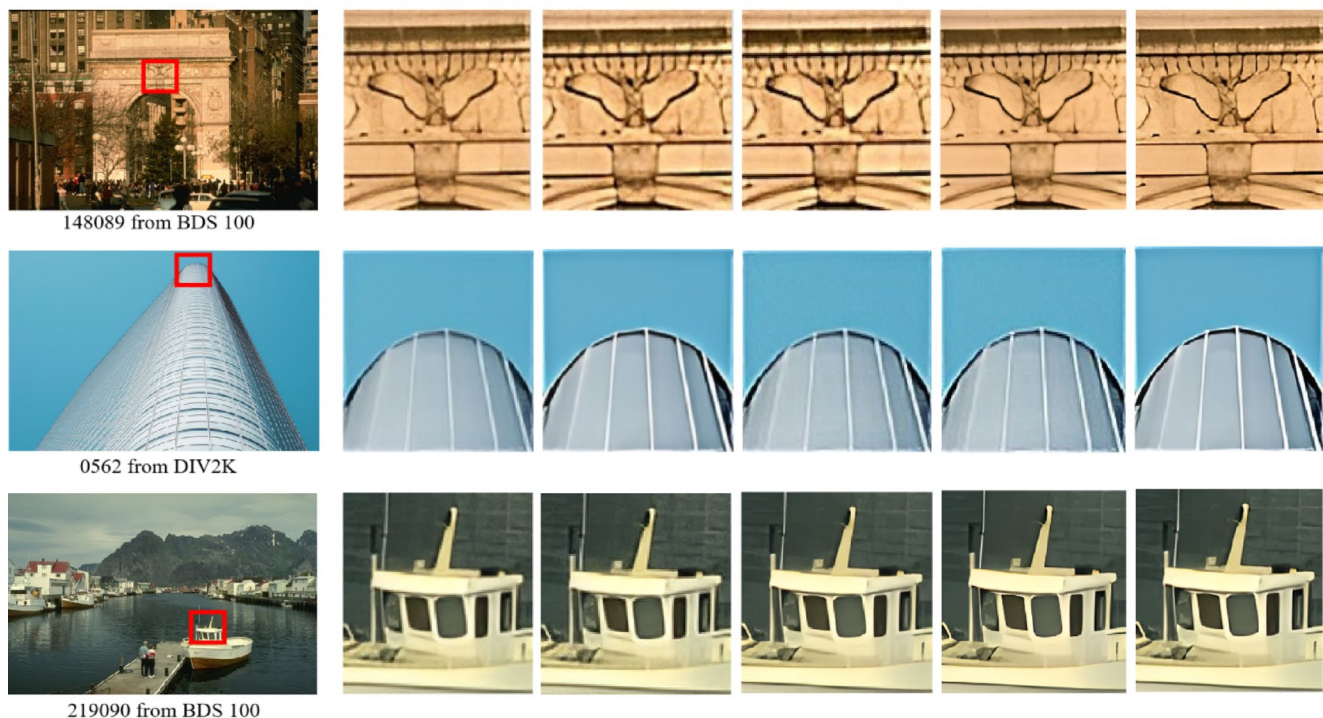
**Fig. 3** Ablation study showing the impact of CBAM placement within RRDB layers on reconstruction quality

Table 5 Ablation study results on the CBAM-ESRGAN variants ($\times 2$, DIV2K)

Model Variant	PSNR (dB)	SSIM	LPIPS ↓	PI ↓	Params (M)
ESRGAN (baseline)	37.29	0.945	0.120	2.08	16.7
+ CBAM (after final dense only)	37.49	0.947	0.115	1.95	16.2
+ CBAM (replaces final dense)	37.68	0.950	0.108	1.87	15.8
+ CBAM (all 3 positions per RRDB)	37.74	0.951	0.105	1.85	17.2
- CBAM + increased depth (4 dense)	37.58	0.947	0.114	1.94	18.4

5 Conclusion

We integrated a sophisticated SISR technique with a CBAM within the ESRGAN framework. By replacing the final dense layer of each RRDB with a CBAM module, our model achieves improved feature recalibration with negligible increases in complexity and computational cost. Rigorous testing on benchmark datasets such as DIV2K, BSD100, and Set14 demonstrates the superiority of our approach in both quantitative fidelity (PSNR and SSIM) and perceptual quality (LPIPS and PI), outperforming numerous state-of-the-art algorithms. Our ablation analysis further reveals that strategically positioning the CBAM modules yields an optimal trade-off between efficiency and effectiveness, without adding a significant number of parameters. Most importantly, our model consistently achieves faster inference times and superior perceptual realism compared to transformer-based and diffusion-based approaches, which typically incur high computational costs. Although the CBAM-ESRGAN model consistently improves perceptual quality and computational efficiency, several limitations remain. The model may produce overly smoothed textures in cases of extreme degradation, such as severely compressed or heavily blurred inputs. In such scenarios, the attention mechanism struggles to recover missing high-frequency details, resulting in reconstructions that appear perceptually clean but lack fine structural realism. The proposed CBAM-ESRGAN model offers a scalable and deployable solution for low-resource and real-time super-resolution applications. Future research will explore extending this attention integration to video super-resolution, resolution adaptation, and domain-specific augmentation for medical and satellite imagery.

Author contributions Mukhiddin Toshpulatov conceived the core idea, designed the network architecture, and conducted the primary model implementation and experimentation. Furkat Safarov contributed to the training pipeline, hyperparameter tuning, and comparative analysis with baseline and state-of-the-art models. Ugiloy Khojamuratova led the literature review and contributed to writing the Related Works and Dataset sections. Misirov Komoliddin and Ganiyev Zafar were responsible for statistical analysis, result validation, and format-

ting of tables and figures. Geehyuk Lee supervised the project, provided theoretical guidance, and revised the manuscript critically for important intellectual content. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by KAIST. This work was supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2026-RS-2024-00436398).

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Ethics Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Conde, M., Timofte, R., Lu, Z., Kong, X., Xing, X., Wang, F., Han, S., Park, M., Hao, T., He, Y., Li, R.: NTIRE 2025 challenge on raw image restoration and super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 1148–1171). (2025)
2. Lu, H., Mei, J., Qiu, Y., Li, Y., Hao, F., Xu, J., Tang, L.: Information sparsity guided transformer for multi-modal medical image super-resolution. *Expert Syst. Appl.* **261**, 125428 (2025)
3. Sharma, A., Shrivastava, B.P., Tyagi, P.K., Siddiqui, E.A., Prasad, R., Gautam, S., Pranjali, P.: Enhanced satellite image resolution with a residual network and correlation filter. *Chemometr. Intell. Lab. Syst.* **256**, 105277 (2025)
4. Dong, P., Li, S., Gong, X., Zhang, L.: HVASR: Enhancing 360-degree video delivery with viewport-aware super resolution. *Inf. Sci.* **691**, 121609 (2025)
5. Yang, Y., Ren, X., Ke, L.: FedSR: Federated learning for image Super-Resolution via detail-assisted contrastive learning. *Knowl. Based Syst.* **309**, 112778 (2025)
6. Sun, J., Chronopoulos, D.: Super-resolution imaging with elastic waves: A review of superlenses, hyperlenses, and Metalenses. *Mater. Sci. Engineering: R: Rep.* **163**, 100935 (2025)
7. Ai, H., Cao, Z., Wang, L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *Int. J. Comput. Vision*, pp.1–40. (2025)

8. He, J., Yuan, Q., Li, J., Xiao, Y., Liu, D., Shen, H., Zhang, L.: Spectral super-resolution Meets deep learning: Achievements and challenges. *Inform. Fusion*. **97**, 101812 (2023)
9. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 136–144). (2017)
10. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681–4690). (2017)
11. Ma, C.: March. Uncertainty-aware GAN for single image super resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 5, pp. 4071–4079). (2024)
12. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2472–2481). (2018)
13. Altinkaya, E., Barakli, B.: Enhancing face image quality: Strategic patch selection with deep reinforcement learning and Super-Resolution boost via RRDB. *IEEE Access*. (2024)
14. Huang, S., Deng, W., Li, G., Yang, Y., Wang, J.: RTEN-SR: A reference-based texture enhancement network for single image super-resolution. *Displays*, **83**, p.102684. (2024)
15. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 3–19). (2018)
16. Dixit, M., Yadav, R.N.: A review of single image super resolution techniques using convolutional neural networks. *Multimedia Tools Appl.* **83**(10), 29741–29775 (2024)
17. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 295–307 (2015)
18. Lembhe, A., Motarwar, P., Patil, R., Elias, S.: Enhancement in skin cancer detection using image super resolution and convolutional neural network. *Procedia Comput. Sci.* **218**, 164–173 (2023)
19. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1646–1654). (2016)
20. Kim, J., Lee, J.K., Lee, K.M.: Deeply-recursive convolutional network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1637–1645). (2016)
21. Gao, M., Sun, J., Li, Q., Khan, M.A., Shang, J., Zhu, X., Jeon, G.: Towards trustworthy image super-resolution via symmetrical and recursive artificial neural network. *Image Vis. Comput.* **158**, 105519 (2025)
22. Chen, Z., Zhang, L., Zhang, X.: Edge fusion diffusion for single image Super-Resolution. *Knowl. Based Syst.* **315**, p113241 (2025)
23. Tong, T., Li, G., Liu, X., Gao, Q.: Image super-resolution using dense skip connections. In *Proceedings of the IEEE international conference on computer vision* (pp. 4799–4807). (2017)
24. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Loy, C.: C., Esgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops* (pp. 0–0). (2018)
25. Chen, H., Li, H., Yao, C., Liu, G., Wang, Z.: Image super-resolution based on improved ESRGAN and its application in camera calibration. *Measurement*, **242**, p.115899. (2025)
26. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 286–301). (2018)
27. Dai, T., Cai, J., Zhang, Y., Xia, S.T., Zhang, L.: Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11065–11074). (2019)
28. Talreja, J., Aramvith, S., Onoye, T.: XTNSR: Xception-based transformer network for single image super resolution. *Complex. Intell. Syst.* **11**(2), 162 (2025)
29. Song, X., Pang, X., Zhang, L., Lu, X., Hei, X.: Single image super-resolution with lightweight multi-scale dilated attention network. *Appl. Soft Comput.* **169**, 112569 (2025)
30. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1833–1844). (2021)
31. Zhang, J., Tu, Y.: SwinFR: Combining SwinIR and Fast Fourier for Super-Resolution Reconstruction of Remote Sensing Images. *Digital Signal Processing*, p.105026. (2025)
32. Zhang, X., Zhang, Y., Yu, F.: September. HiT-SR: Hierarchical transformer for efficient image super-resolution. In *European Conference on Computer Vision* (pp. 483–500). Cham: Springer Nature Switzerland. (2024)
33. Wang, J., Fan, Q., Chen, J., Gu, H., Huang, F., Ren, W.: April. RAP-SR: RestorAtion Prior Enhancement in Diffusion Models for Realistic Image Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 7, pp. 7727–7735). (2025)
34. Jeevan, P., Srinidhi, A., Prathiba, P., Sethi, A.: Wavemixsr: Resource-efficient neural network for image super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 5884–5892). (2024)
35. Luo, X., Ai, Z., Liang, Q., Xie, Y., Shi, Z., Fan, J., Qu, Y.: EdgeFormer: Edge-aware Efficient Transformer for Image Super-resolution. *IEEE Transactions on Instrumentation and Measurement*. (2024)
36. Yue, Z., Wang, J., Loy, C.C.: Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Adv. Neural. Inf. Process. Syst.* **36**, 13294–13307 (2023)
37. Safarov, F., Mukhiddin, T., Komoliddin, M., Abdusalomov, A., Lee, W.: Hyperspectral Anomaly Detection with Enhanced Spectral Graph Transformer Network. *IEEE Access*, (2025)
38. Safarov, F., Muksimova, Sh., Misirov, K., Cho, YI.: Fire and Smoke Detection in Complex Environments. *Fire* (2024), **7**(11), 389; <https://doi.org/10.3390/fire7110389>
39. Safarov, F., Khojamuratova, U., Misirov, K., Xusimov, II., & Cho, YI.: A Multimodal Deep Learning Framework for Accurate Biomass and Carbon Sequestration Estimation from UAV Imagery Drones (2025), **9**(7), 496; <https://doi.org/10.3390/drones9070496>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.