KMANPUB | AI and Tech in Behavioral and Social Sciences Vol. 4 No. 2 (2026): 1-15

Article history:
Received 26 August 2025
Revised 19 October 2025
Accepted 25 November 2025
Published online 01 April 2025

Attention-Guided Dynamic Model Selection for Single Image Super-Resolution Using Deep Ensemble Learning

Faraz Mohammadian Jadval Ghadam[1], Sattar Hashemi[2*], Karamollah Bagherifard[3], Samad Nejatian[4]

[1] Ph.D. Candidate, Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran
[2] Professor of Artificial Intelligence, Department of Artificial Intelligence, Shiraz University, Shiraz, Iran
[3] Associate Professor, Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran
[4] Assistant Professor, Department of Computer Engineering, Yas.C., Islamic Azad University, Yasuj, Iran

*** Corresponding author email address**: ka.bagherifard@iau.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Mohammadian Jadval Ghadam, F., Hashemi, S., Bagherifard, K., & Nejatian, S. (2026). Attention-Guided Dynamic Model Selection for Single Image Super-Resolution Using Deep Ensemble Learning. *AI and Tech in Behavioral and Social Sciences, 4*(2), 1-15.
https://doi.org/10.61838/kman.aitech.5375

A B S T R A C T

The rapid growth of digital imaging technologies has made high-quality visual data increasingly accessible; however, the storage, transmission, and restoration of high-resolution images remain challenging in bandwidth-limited and resource-constrained environments. Although compression methods reduce file size, they may remove critical details required for scientific, medical, remote-sensing, and security applications. To address this limitation, this study proposes an attention-guided dynamic ensemble framework for Single Image Super-Resolution (SISR). The proposed method integrates several representative super-resolution models, including LapSRN, SRResNet, ResNeXt-based SR, SRCNN/FSRCNN, and ESPCN, and uses an attention-guided selection module to assign the most suitable model to different image regions based on local characteristics such as edges, textures, and smooth areas. The selected outputs are then fused by a convolutional integration network to generate the final high-resolution image. Experiments on DIV2K and BSDS300 show that the proposed method improves reconstruction quality, particularly in terms of structural similarity and texture preservation. On DIV2K, the proposed method achieved 33.40 dB PSNR and 0.9172 SSIM; on BSDS300, it achieved 28.13 dB PSNR and 0.8497 SSIM. These findings indicate that dynamic model selection can reduce the limitations of individual super-resolution models and improve detail recovery in feature-diverse images.

Keywords: *Super-resolution; Single Image Super-Resolution; Attention Mechanism; Ensemble Deep learning; Model Selection; Image Reconstruction*

1. Introduction

There is a growing demand for high-quality images in commercial, scientific, medical, and security-related applications. Capturing or transmitting such images can be costly or technically difficult. For example, mobile devices cannot always accommodate large optical sensors, satellite and remote-sensing systems may face bandwidth limitations, and medical imaging systems may produce images whose spatial resolution is insufficient for accurate interpretation. These constraints have encouraged researchers in computer vision and artificial intelligence to develop software-based methods for enhancing low-resolution (LR) images. Super-resolution (SR) aims to reconstruct a high-resolution (HR) image from one or more LR observations, and Single Image Super-Resolution (SISR) is particularly important when only one LR input is available (Dong, Loy, He, et al., 2016; Wang et al., 2021). Conventional SR methods, including interpolation, edge-based reconstruction, patch-based learning, sparse representation, and statistical estimation, have contributed to the field but often fail to recover realistic high-frequency

details (Freeman et al., 2002; Glasner et al., 2009; Keys, 1981; Yang et al., 2010; Yue et al., 2016). Deep learning has therefore become the dominant paradigm for SISR because it can learn nonlinear mappings between LR and HR image spaces. In this study, the original SISR framework is strengthened by using attention-guided dynamic model selection within a deep ensemble architecture. Rather than applying one fixed model to every region, the proposed framework adaptively selects the most suitable model for each region according to local image content.

In recent years, deep learning has substantially improved SISR performance. Convolutional models such as SRCNN, FSRCNN, VDSR, LapSRN, EDSR, and transformer-based approaches have demonstrated strong reconstruction ability (Dong, Loy, He, et al., 2016; Dong, Loy, & Tang, 2016; Kim et al., 2016a, 2016b; Lai et al., 2017; Liang et al., 2021). Ensemble learning is also relevant because it combines complementary models to improve robustness and reduce the bias of any single architecture. In image restoration, ensembles can exploit the strengths of different networks: some architectures preserve edges more effectively, whereas others reconstruct textures or smooth regions more reliably. The attention-guided mechanism proposed in this article extends this idea by selecting and fusing model outputs in a region-adaptive manner.

In general, ensemble learning combines multiple predictors and aggregates their outputs to produce a more reliable final result. Unlike classification-only ensembles, super-resolution ensembles operate on continuous image outputs. Bagging, boosting, stacking, and dynamic model selection are common ensemble strategies; however, SISR requires spatially aware fusion because different regions of the same image may require different reconstruction priors. The present framework therefore uses attention maps to guide model selection at the regional level and then applies a CNN-based integration stage to combine the selected outputs.

Our proposed approach seeks to enhance image resolution by mitigating the limitations of individual models through ensemble learning. By integrating SISR with deep ensemble learning, we aim to combine the strengths of deep learning and ensemble techniques, resulting in a model with enhanced generalization capabilities. This paper demonstrates how this synergy between SISR and ensemble learning techniques leads to the production of high-quality images.

The structure of the paper is as follows: Section 2 reviews related work in the literature. Section 3 details the proposed method, explaining our novel approaches. Section 4 presents the experimental setup and results. Section 5 discusses the findings, comparing them with state-of-the-art methods. Finally, Section 6 concludes the paper and explores directions for future research.

2. Related Work

The development of SISR has moved from interpolation and sparse representation toward deep, attention-based, and ensemble architectures. Early CNN-based approaches such as SRCNN established the feasibility of end-to-end LR-to-HR mapping (Dong, Loy, He, et al., 2016). FSRCNN improved computational efficiency by operating directly on LR images and using a deconvolution layer for upsampling (Dong, Loy, & Tang, 2016). VDSR and DRCN showed that deeper architectures and residual learning could improve reconstruction quality (Kim et al., 2016a, 2016b). Later, LapSRN used a progressive Laplacian pyramid structure to reconstruct high-frequency residuals at multiple scales (Lai et al., 2017). GAN-based models, especially SRGAN, shifted attention from pixel-level fidelity alone to perceptual realism (Ledig et al., 2017). More recent attention-based and transformer-based models, such as non-local sparse attention and SwinIR, further improved long-range dependency modeling and texture reconstruction (Ledig et al., 2017; Liang et al., 2021; Mei et al., 2021). This literature supports the need for architectures that are not only deep but also adaptive to image content.

Generating HR images from LR inputs is an ill-posed inverse problem because multiple plausible HR images can correspond to the same LR observation. Traditional SISR methods usually apply a fixed reconstruction rule across the entire image. However, feature-diverse images contain smooth regions, sharp edges, repeated textures, and complex patterns, each of which may require a different reconstruction strategy. The attention-guided dynamic selection mechanism used in this study was designed to address this limitation by assigning specialized models to the regions where they are expected to perform best.

Among existing methods, interpolation techniques are widely used due to their simplicity and ease of implementation (Wang et al., 2004). However, these linear methods often lack the representational power to capture complex details, resulting in blurry HR outputs. To address these limitations, sparsity-based methods [5, 15] have been

AITBSS
E-ISSN: 3041-9433

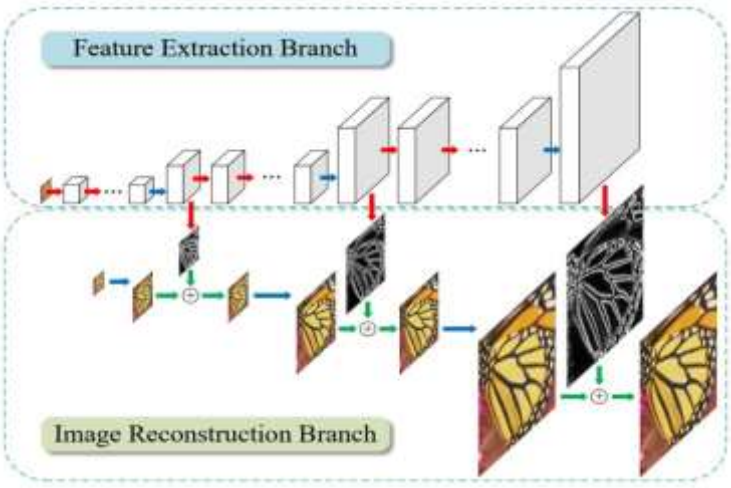
developed to enhance linear models by incorporating rich image priors, allowing for more accurate reconstructions. For instance, the Bicubic Super-Resolution (SR) method (Freeman et al., 2002) is often used as a benchmark for evaluating model performance. Bicubic SR estimates pixel values by fitting a surface among neighboring pixels, considering vertical, horizontal, and diagonal gradients, as well as intensity values. This method is popular due to its relatively low computational complexity and satisfactory results, making it a useful baseline for analyzing the gap between unsupervised and supervised SR frameworks.

Since the introduction of GANs, adversarial learning has been widely used to improve the perceptual quality of super-resolution outputs (Ledig et al., 2017). In SISR, SRGAN demonstrated that perceptual and adversarial losses can generate visually realistic textures, although such models may introduce artifacts or reduce pixel-level fidelity in some cases. Therefore, GAN-based components are best used with complementary CNN-based models and carefully designed fusion mechanisms.

LapSRN is not a GAN-based method; rather, it is a progressive CNN architecture based on a Laplacian pyramid that reconstructs high-frequency residuals at multiple scales (Lai et al., 2017). Because of its coarse-to-fine design, LapSRN is useful for recovering structural details and reducing computational cost compared with methods that require pre-upsampling by bicubic interpolation. In the present framework, LapSRN is treated as one of the candidate reconstruction models within the ensemble rather than as an adversarial generator.

Figure 1

LapSRN (Denton et al., 2015)



High-resolution (HR) images, particularly in fields such as aerial and medical imaging, are often expensive to produce and suffer from lower signal-to-noise ratios. To address these challenges, a variety of machine learning models have been proposed, achieving promising results. Among these, deep learning networks have demonstrated significant success in processing both two-dimensional and three-dimensional data. However, achieving optimal performance with deep networks requires careful alignment of the model architecture and objective functions with the inherent characteristics of the data. Regularization techniques are often employed to ensure this alignment.

Advanced deep network architectures, such as Inception, ResNet (Residual Neural Network) (He et al., 2016), and DCCN (Dense Connected Cascade Network) (Mei et al., 2021), have shown potential for improving image quality when combined with perceptually driven target functions. These architectures, along with their ability to capture intricate patterns and structures in data, highlight the need for continued research and development to fully realize their capabilities in high-quality image reconstruction (Haris et al., 2018; Tong et al., 2017).

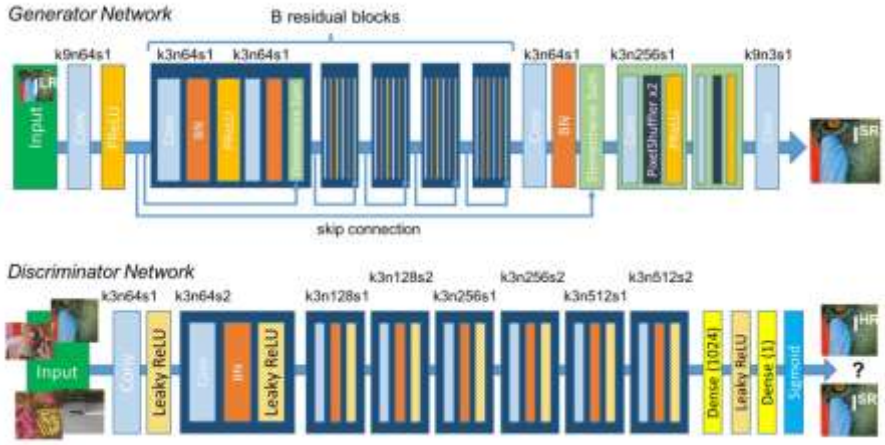
Recent architectures such as SRResNet, ESPCN, and related CNN-based models use specialized network designs to improve reconstruction quality (Ledig et al., 2017; Mei et al., 2021; Shi et al., 2016). However, applying these models uniformly across all image regions can be suboptimal, particularly when an image contains both smooth and highly textured areas. To address this limitation, the present study introduces an attention-guided

AITBSS
E-ISSN: 3041-9433

dynamic model selection mechanism that adaptively selects the most suitable model for each region. Figure 2 illustrates the SRResNet structure used as one of the candidate reconstruction models.

Figure 2

SRResNet (Denton et al., 2015)



In SRGAN and SRResNet-style models, the generator-discriminator framework is used to distinguish HR ground-truth images from super-resolved outputs, while the generator reconstructs HR images from LR inputs (Ledig et al., 2017). SRResNet itself is the residual generator architecture without the adversarial discriminator. This distinction is important because the proposed ensemble includes both adversarial and non-adversarial SR components.

Each residual block in SRResNet typically consists of convolutional layers, batch normalization, and rectified linear unit (ReLU) activations, with skip connections facilitating gradient flow and feature preservation. The original SRResNet architecture uses residual learning to preserve image structure and stabilize training (Ledig et al., 2017). The content loss used in SRResNet-style reconstruction is shown in Equation (1).

$$l^{SR}_{\frac{VGG}{i}.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\emptyset_{i,j}(I^{HR})_{x,y} - \emptyset_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2 \quad (1)$$

Another network incorporated into this model is ResNeXt, which has been optimized to address challenges in super-resolution tasks. Figure 3 illustrates the architecture of this network. ResNeXt is specifically designed to overcome several limitations of generative adversarial networks (GANs) used in super-resolution (SRGAN), such as high computational complexity, network instability, and slow learning rates.

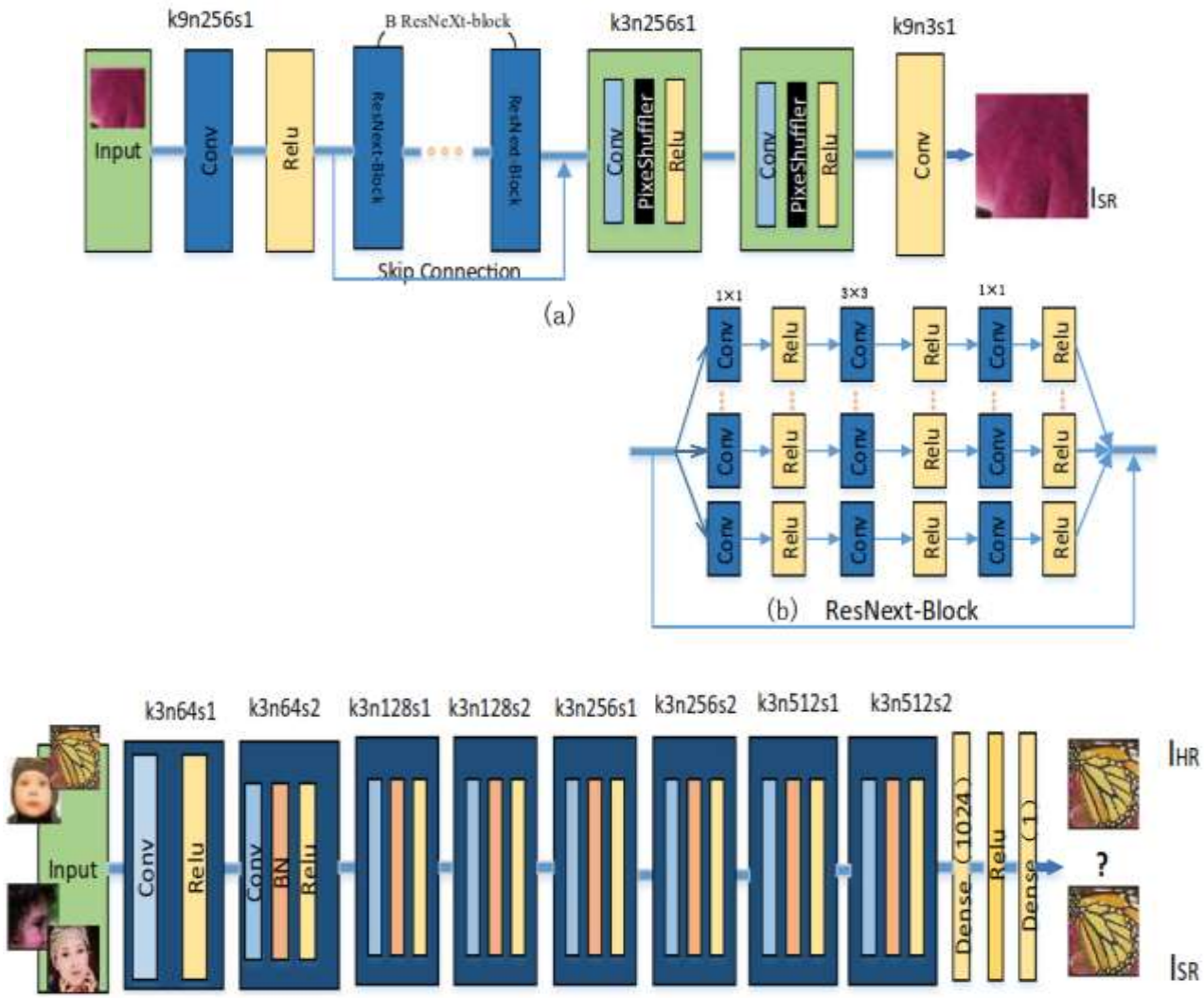
By employing the ResNeXt architecture, the computational complexity of the generator is reduced significantly to nearly one-eighth of the original SRGAN. Additionally, to address instability issues inherent in SRGAN, researchers implemented a discriminator based on the Wasserstein GAN (WGAN). This adjustment enhances stability and improves training efficiency. Further, the learning rate was accelerated by omitting the normalization operation in the residual layers of the network.

Experimental evaluations of Res_WGAN demonstrated both subjective and objective improvements over existing models. These assessments, conducted using five benchmark datasets, highlighted the superiority of the proposed approach compared to state-of-the-art super-resolution methods (Zhang, Li, et al., 2018).

AITBSS
E-ISSN: 3041-9433

Figure 3

RESNeXt (Timofte et al., 2014)



The loss function in the ResNeXt-based SR component is computed using Equation (2).

$$L_{Gen}^{SR} = \sum_{n=1}^{N} = D_{\theta_D}\left(G_{\theta_G}(I^{LR})\right) \quad (2)$$

Another important model for super-resolution is the Super-Resolution Convolutional Neural Network (SRCNN), together with its faster variant, FSRCNN. Figure 4 illustrates the architecture of SRCNN/FSRCNN-style reconstruction. SRCNN learns an end-to-end mapping from interpolated LR inputs to HR outputs, whereas FSRCNN improves efficiency by operating directly on LR images and upsampling only at the final stage (Dong, Loy, He, et al., 2016; Dong, Loy, & Tang, 2016).

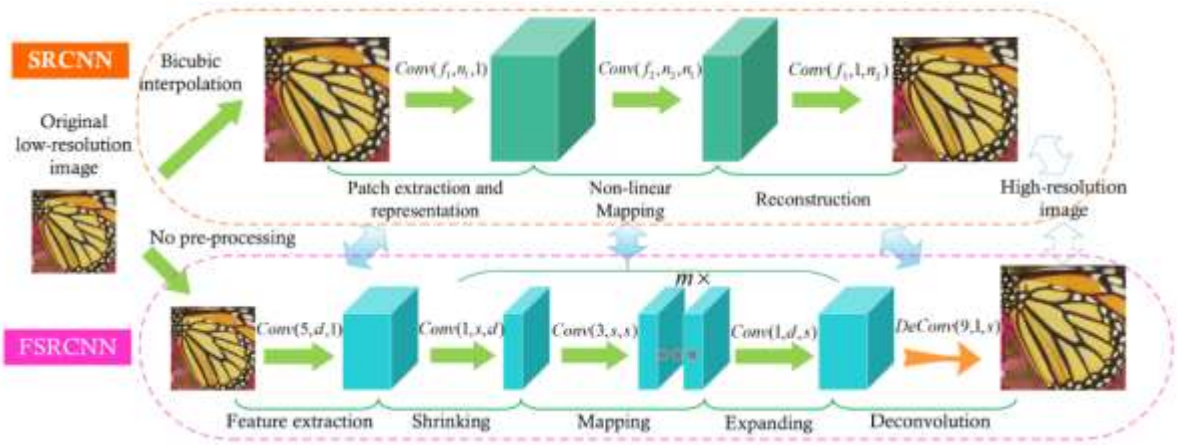
Firstly, FSRCNN directly uses the original low-resolution image as input, bypassing the need for bicubic interpolation. Secondly, it incorporates a deconvolution layer at the end of the network for upsampling, allowing for more precise resolution enhancement. Thirdly, FSRCNN replaces SRCNN's non-linear mapping steps with a three-step process comprising shrinking, mapping, and expanding layers. Additionally, FSRCNN utilizes smaller-sized filters and a deeper network architecture, which significantly boosts performance while reducing computational costs.

These improvements make FSRCNN more efficient for real-world applications than the original SRCNN, particularly when computational cost is a practical concern (Dong, Loy, & Tang, 2016).

AITBSS
E-ISSN: 3041-9433

Figure 4

SRCNN (Huang et al., 2017)



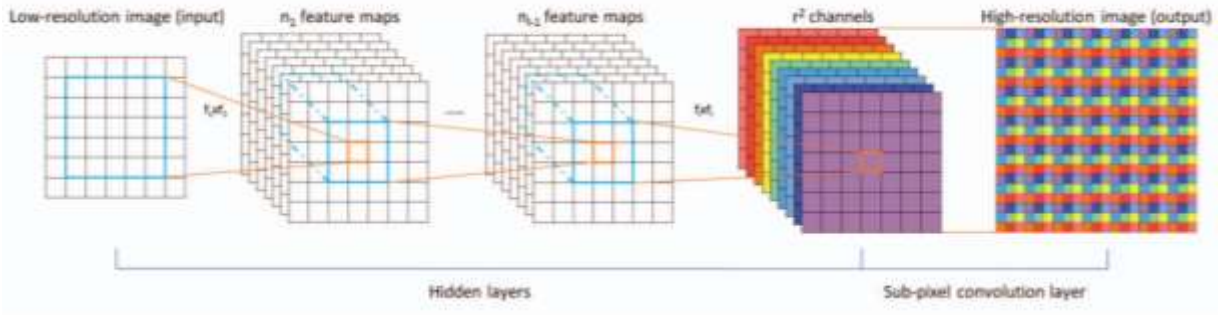
The loss function in SRCNN is computed using this formula:

$$loss = \sum_{i=1}^{N} \quad ay_i \log \log (p_i) + (1-\alpha)(1-y_i) \log \log (1-p) \qquad (3)$$

The final component of the proposed model is the Efficient Sub-Pixel Convolutional Neural Network (ESPCN), which performs efficient upscaling by learning feature maps in LR space and rearranging them into HR outputs through a sub-pixel convolution layer (Shi et al., 2016). In the SISR task, an HR image is reconstructed from its LR counterpart, which is typically generated by downsampling the HR image using a known degradation operator. ESPCN avoids the need for early interpolation and can therefore reduce computational complexity while preserving useful image features. The operational flow of this network is illustrated in Figure 5.

Figure 5

ESCPN



In the conventional process, the LR image is first upscaled to match the dimensions of the HR image using interpolation techniques. This results in the LR image data being mapped onto a surface, or a manifold, within the image space. However, this procedure differs from that of the HR image, as the LR image lies at a certain distance from the HR image after interpolation. This distinction is crucial, as without it, the value of machine learning approaches would be undermined. The dimensionality of the LR image is clearly different from that of the HR image, with the HR image containing more detail. If manifold learning is applied to the LR image, the coordinates of the LR image within this manifold are obtained. These coordinates must then be mapped to the corresponding coordinates of the HR image in the HR manifold, which can be achieved through feature learning techniques. Finally, the HR image can be reconstructed

AITBSS
E-ISSN: 3041-9433

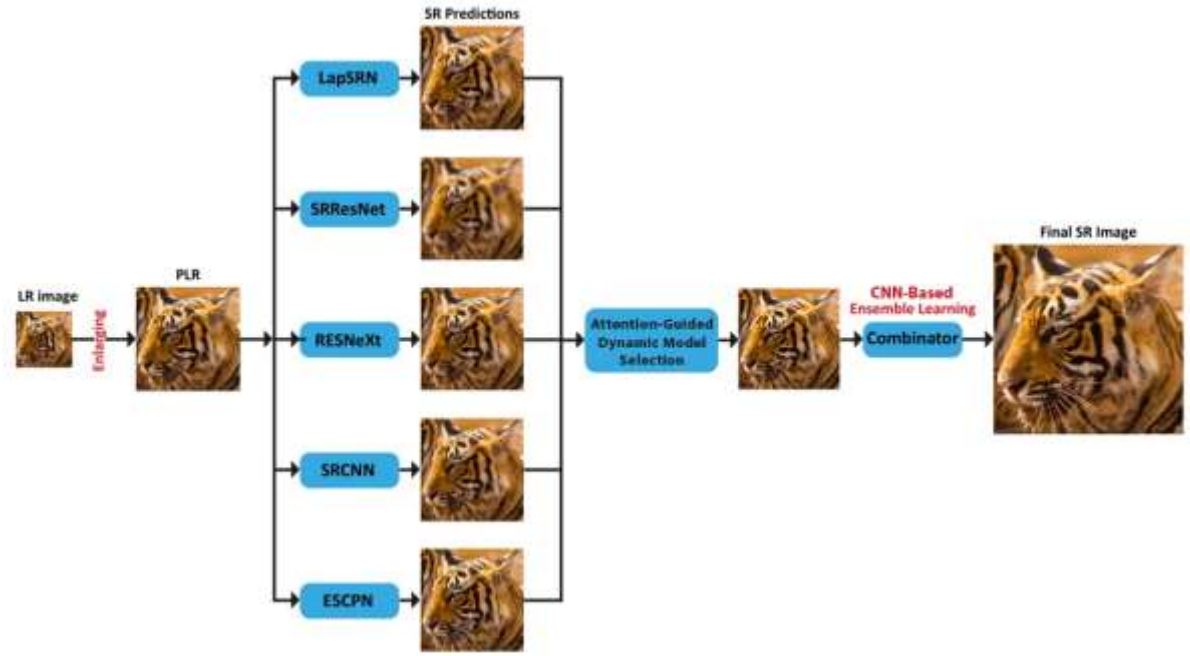
from these learned HR manifold coordinates (Wang et al., 2018; Zhang, Tian, et al., 2018).

The classification-oriented explanation in the original version has been revised because SISR is a regression and image reconstruction problem rather than a classification task. In the proposed framework, the learned function maps LR image features to HR image estimates. The dynamic selection module can be interpreted as a model-selection function that assigns the most appropriate SR predictor to each image region.

High-resolution (HR) images play a crucial role in various domains, such as satellite and aerial imagery, ultrasound imaging, medical imaging, traffic monitoring, security surveillance, ground-based remote sensing, astronomical observations, and biometric detection. However, due to hardware and physical constraints, producing high-quality images is often expensive and can result in issues such as low signal-to-noise ratio (SNR) and extended image production times (Ding et al., 2021).

Figure 6

Workflow of the SR method used in this study



$g \approx f^{-1}$To reconstruct I_HR, the model approximates the inverse transformation from the LR observation to the HR image estimate, as shown in Equation (5).

$$I_{HR} \approx g(I_{LR}) \quad (5)$$

3. Proposed Method

In this study, an ensemble deep learning approach with dynamic model selection and attention-based guidance is used to improve image resolution. The framework combines multiple pre-trained SR models to optimize SISR quality by exploiting the distinct characteristics of different image regions. The intended reconstruction task is to upscale LR inputs to HR outputs while recovering fine details in edges, textures, and structural patterns. Figure 6 demonstrates the workflow of the proposed method.

$p \times qm \times n$Let I_LR ∈ R^{p×q} denote a low-resolution image and I_HR ∈ R^{m×n} denote its corresponding high-resolution image. The degradation process from I_HR to I_LR can be modeled as a forward operator that includes blurring, downsampling, and possible noise. This relationship is represented in Equation (4).

$$I_{LR} = (I_{HR}) \quad (4)$$

f^{-1}Because the exact inverse of the degradation function is not available, different image priors and learned mappings are used to approximate the HR reconstruction. The proposed method consists of three main stages: attention-map generation, dynamic selection among

AITBSS
E-ISSN: 3041-9433

candidate SR models, and CNN-based ensemble integration.

3.1. *Pre-processing and Attention Map Generation:*

$p = \frac{m}{2}$ $q = \frac{n}{2}$In the first step, HR images are degraded into LR images for training and evaluation. The experiments use standard benchmark degradation settings associated with DIV2K and BSDS300 comparisons, while the Fourier-domain description is retained as an auxiliary degradation simulation. An attention network is then applied to generate an attention map for each LR image. This map highlights feature-rich regions, such as edges and textures, that require specialized reconstruction.

1. Candidate SR Models with Dynamic Selection Mechanism:

Five distinct super-resolution models - LapSRN, SRResNet, a ResNeXt-based SR model, SRCNN/FSRCNN, and ESPCN - are pre-trained or fine-tuned for reconstruction. These models are not all GANs; rather, they include both CNN-based and adversarially inspired architectures. Using the attention map, the dynamic selection mechanism assigns the most suitable model to each image region. For example, texture-rich areas can be processed by detail-preserving models, while edge-dominant regions can be assigned to residual or pyramid-based networks.

3.2. *Ensemble Learning through CNN Integration:*

The dynamically selected GAN outputs are integrated using a Convolutional Neural Network (CNN) to generate the final high-resolution image. In this ensemble learning process, the outputs from each model are concatenated and input into the CNN for training, validation, and testing. The CNN then combines these diverse inputs into a unified image, effectively leveraging the strengths of each model to maximize reconstruction quality. The process is guided by minimizing the mean-absolute-error (MAE) as the loss function, defined as:

$$I_{SR} = (\sum I_{SR,i}\ , \Theta) \qquad (6)$$

where $I_{SR} \in \mathbb{R}^{m \times n}$ indicates ensemble learning outputs (our SR outcomes), ϕ: $I_{SR},i \in \mathbb{R}^{m \times n} \rightarrow I_{SR} \in \mathbb{R}^{m \times n}$ suggests the process of ensemble learning, Θ represents the parameters of the CNN model.

4. Findings and Results

In this section, we present the datasets used and the evaluation metrics employed to assess the proposed method.

4.1. *Datasets*

To evaluate the effectiveness of the proposed approach and compare it with state-of-the-art methods, the DIV2K and BSDS300 datasets were used. Table 1 summarizes the dataset specifications used in this paper.

Table 1

Datasets used in this study

Dataset	Training data	Validation data	Experimental data
DIV2K (RGB images)	800 images (with three low-quality images equivalent)	100 data (with equivalent low-quality images.)	100 images (with equivalent low-quality)
BSDS300	200 images	----------	100 images

1. **DIV2K**: This dataset consists of RGB images divided into three subsets:
   - **Training data**: 800 images, each paired with three low-quality versions corresponding to different quality reduction coefficients.
   - **Validation data**: 100 images, each with equivalent low-quality versions.
   - **Test data**: 100 images, each with equivalent low-quality versions. Low-quality images are generated using the bicubic downsampling method as indicated in the file names.
2. BSDS300: Originally developed for segmentation and boundary detection research, this dataset was repurposed here for super-resolution evaluation by generating LR-HR image pairs. It contains 200 training images and 100 test images.

AITBSS E-ISSN: 3041-9433

4.2. *Evaluation Metrics*

$A_{i,j}B_{i,j}$Mean Squared Error (MSE), a widely used full-reference image quality metric, is calculated by averaging the squared differences between corresponding pixels of the reference image and the reconstructed image. In Equation (7), A represents the original HR image and B represents the reconstructed SR image.

$$MSE = \frac{\sum_{j=1}^{N} \left(\sum_{i=1}^{M} (A_{i,j} - B_{i,j})^2 \right)}{MN} \quad (7)$$

Here, M represents the number of rows in image A, and N denotes the number of columns in image B. Additionally, the Peak Signal-to-Noise Ratio (PSNR), which is derived as a function of the Mean Squared Error (MSE), is employed as an evaluation criterion in this research:

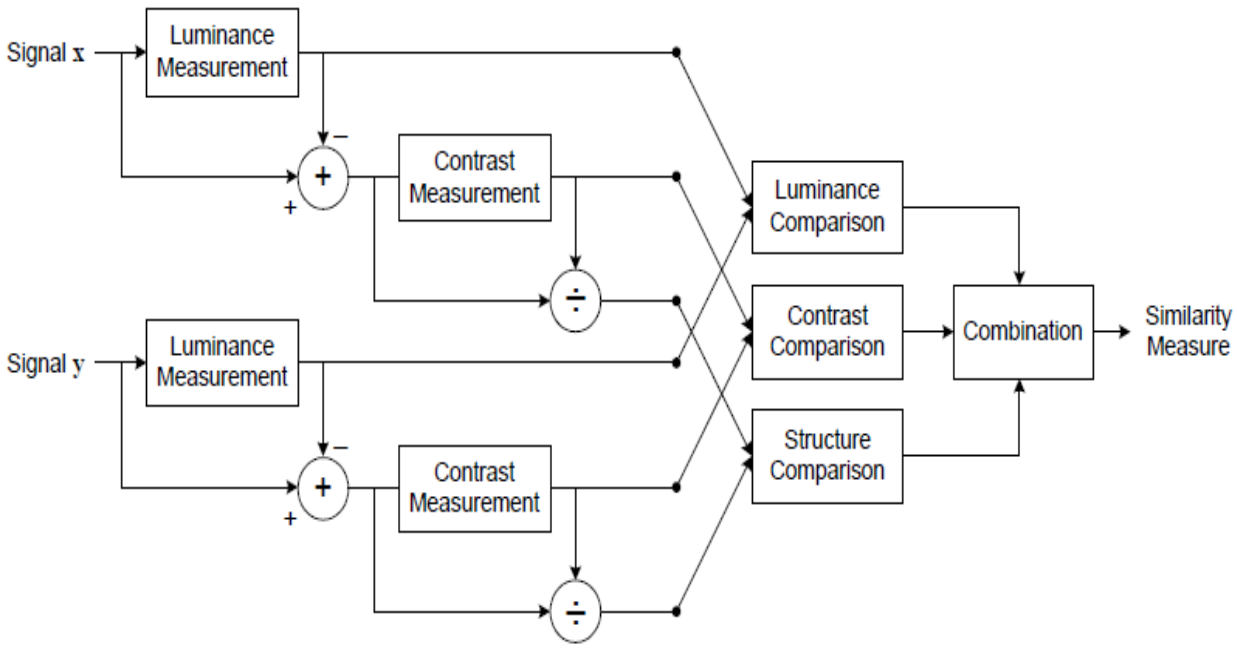
$$PSNR = 10log_{10} \frac{(maximum\ pixel\ value)^2}{MSE} \quad (8)$$

PSNR and SSIM are widely used in SR evaluation, but they capture different aspects of image quality. PSNR is derived from pixel-wise error and is useful for measuring fidelity, whereas SSIM estimates structural similarity and better reflects perceptual consistency (Wang et al., 2004). Because MSE and PSNR do not always align with human visual perception, SSIM is also used as a complementary metric.

Most of the work has focused on the first approach, with significant efforts directed at enhancing MSE by incorporating penalties to constrain its free parameters (Zhang, Tian, et al., 2018). Within this framework, the Structural Similarity Index Measure (SSIM) was introduced in (Wang et al., 2018), designed based on the human visual system's functionality. The performance of SSIM is demonstrated in Figure 7. While SSIM is rooted in the HVS-based approach, it also aligns with the second approach as it aims to capture the structural similarity within the data.

Figure 7

SSIM performance diagram (Martin et al., 2001)



The structure of an object in a scene is independent of its brightness. Therefore, to analyze the structural information in an image, the influence of brightness must be isolated. Since brightness and contrast vary across scenes, local brightness and contrast are used to define the similarity criterion. As shown in Figure 7, the similarity criterion is divided into three components: brightness, contrast, and structure. First, brightness is compared. Let xxx and yyy represent two sub-images with non-negative values from the same image. The average values of these sub-images are then used to compare their brightness:

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \quad (9)$$

AITBSS
E-ISSN: 3041-9433

The brightness comparison function $l(x,y)$ depends on μ_x and μ_y, the mean values of sub-images x and y, respectively. By subtracting the mean from each sub-image, the resulting sub-image is mapped onto the following scatter plot:

$$\sum_{i=1}^{N} \quad x_i = 0 \quad (10)$$

The standard deviation is then used to estimate the contrast of the image:

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N} \quad (x_i - \mu_x)^2\right)^{\frac{1}{2}}$$

(11)

The contrast comparison function $c(x,y)$ is defined as a function of σ_x and σ_y, the standard deviations of sub-images x and y, respectively.

The signal is normalized by dividing it by the standard deviation, allowing the structure comparison $s(x,y)$ to be performed on this normalized signal. These three components brightness $l(x,y)$, contrast $c(x,y)$, and structure $s(x,y)$—are then combined to form the overall similarity criterion $S(x,y)$. The definition of these functions and their combination, as detailed in (Wang et al., 2004), is based on the following three conditions:

1. Symmetry: S (x, y) = S (y, x)
2. Boundary: S (x, y) ≤1
3. Having a maximum value: S (x, y) = 1 if and only if x = y

The function l (x, y) is defined as follows:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (12)$$

The constant C_1 is defined to prevent values close to zero of the denominator. This constant is considered as follows:

$$C_1 = (K_1 L)^2 \quad (13)$$

Where L is the image signal range (for example, for a grayscale image with 8 bits per pixel, this value is 255) and K_1 is a small value less than 1: K≪1. The function l precisely satisfies the three required conditions. In addition to these three conditions, this criterion is also compatible with the Weber rule or a light mask used to model light in the human visual system. The contrast comparison function also has a similar structure:

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (14)$$

Where C_2 is defined as $C_2 = (K_2 L)^2$. This function is also compatible with the contrast mask used in HVS.

After the image of each of the sub-images below the space equivalent to the equation $\sum_{i=1}^{N} \quad x_i = 0$ and normalizing it with standard deviation, to measure the similarity of the structure, we can find the correlation between the two Used images. Since the correlation between $\frac{x-\mu_x}{\sigma_x}$ and $\frac{y-\mu_y}{\sigma_y}$ is equal to the Pearson correlation coefficient between x and y, the function can be defined as follows:

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (15)$$

Where σ_xy is defined as follows:

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N} \quad (x_i - \mu_x)(y_i - \mu_y) \quad (16)$$

The general criterion defined in (Agustsson & Timofte, 2017) combines the above three parts as follows and then specifies the default value for the specified parameters α, β, and γ, and the criterion used in the article is achieved:

$$SSIM(x,y) = l(x,y)^{\alpha}.c(x,y)^{\beta}.s(x,y)^{\gamma} \quad (17)$$

These three parameters in the article are all considered equal to 1, and in addition, the value $C_3 = \frac{C_2}{2}$ is set:

$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$In general, PSNR is used to evaluate pixel-level fidelity, whereas SSIM is used to evaluate structural similarity. These two criteria provide complementary evidence for comparing different SR methods. Figures 8 and 9 illustrate examples of reconstruction behavior under noisy and noise-free conditions.

In general, PSNR is used to evaluate the similarity of the gray value and SSIM is used to show the structural similarity. Using these two criteria, different methods are compared together. Other algorithms can be used to calculate MSE. When images are corrupted by noise, l_1 software can not eliminate dots, but preserves the texture of the image. In contrast, l_1 software has many problems in dealing with Gaussian noise distribution, and l_p software achieves better results in terms of visual impact and quantitative indicators. For example, in Figure 8, the soft l_p with $p = 1.3$, and in Figure 9, the soft $l_p$with $p = 1.5$, are provided. Also, it can be clearly seen in Figure 8 that l_1 and l_p are stronger than l_2 (Agustsson & Timofte, 2017; Wang et al., 2018; Zhang, Tian, et al., 2018).

AITBSS
E-ISSN: 3041-9433

Figure 8

Results of reconstruction of transferability of Lena image using (a) Bilinear interpolation and (b) soft l2, (c) soft l1, and (d) soft lp with p = 1.3. As shown in Figure 8, image (a) has noise, and its edges are not obvious. The input parameters in (b), (c), and (d) determine the clarity value of the edges of these images; however, these parameters increase the complexity and operation time of the algorithm. In fact, there is a trade-off between time, complexity, and quality.

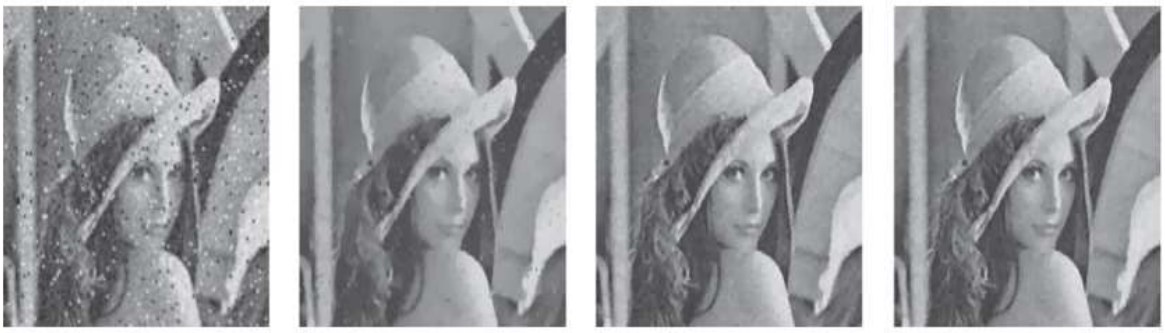
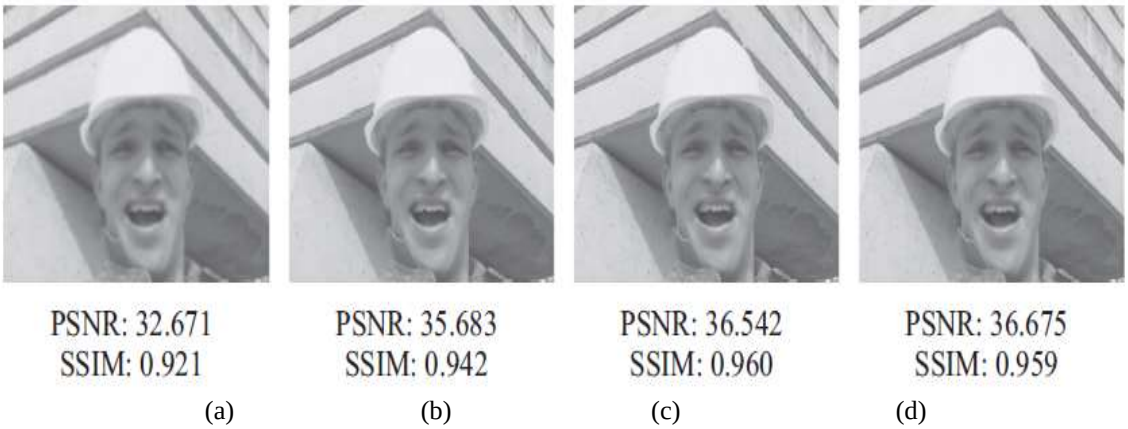


Figure 9

Results of reconstruction of the resolution of the image without noise using (a) Bilinear interpolation and (b) soft l2, (c) soft l1, and (d) soft lp with p = 1.5



4.3. Results

In the experiments, the integration of attention-guided dynamic model selection improved overall reconstruction quality. The improvement was most evident in texture preservation, edge recovery, and SSIM performance. On DIV2K, the proposed method achieved a PSNR of 33.40 dB and SSIM of 0.9172, outperforming the compared methods in both metrics. On BSDS300, the proposed method achieved the best SSIM and MSE values and a PSNR value that was competitive with the best-performing baseline. Therefore, the results support the effectiveness of the attention-guided ensemble strategy without changing the original numerical findings.

The proposed algorithm was applied to several test images, and representative visual results are shown in Figures 10 and 11.

AITBSS
E-ISSN: 3041-9433

Figure 10

Results of our proposed method against other methods

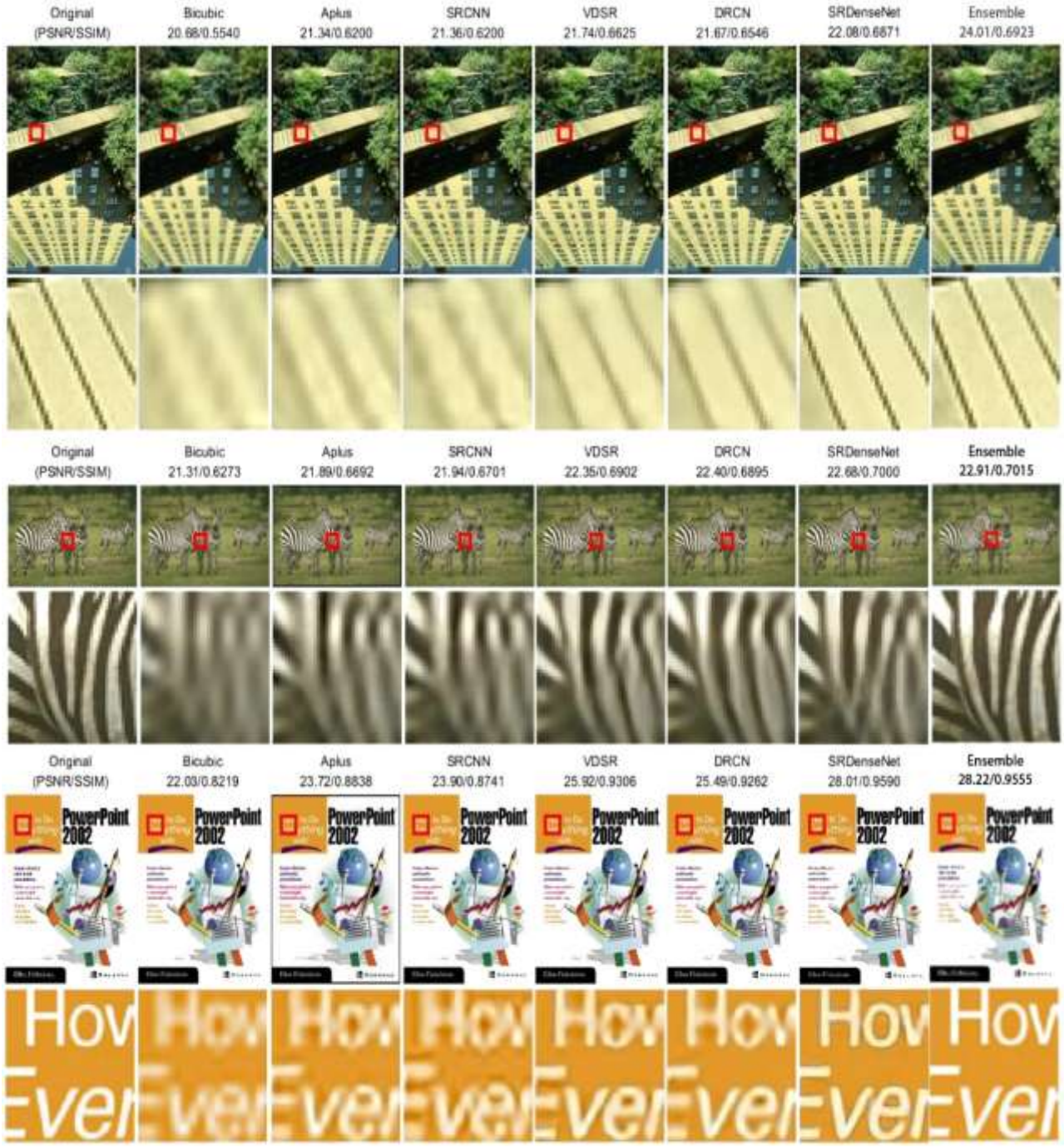
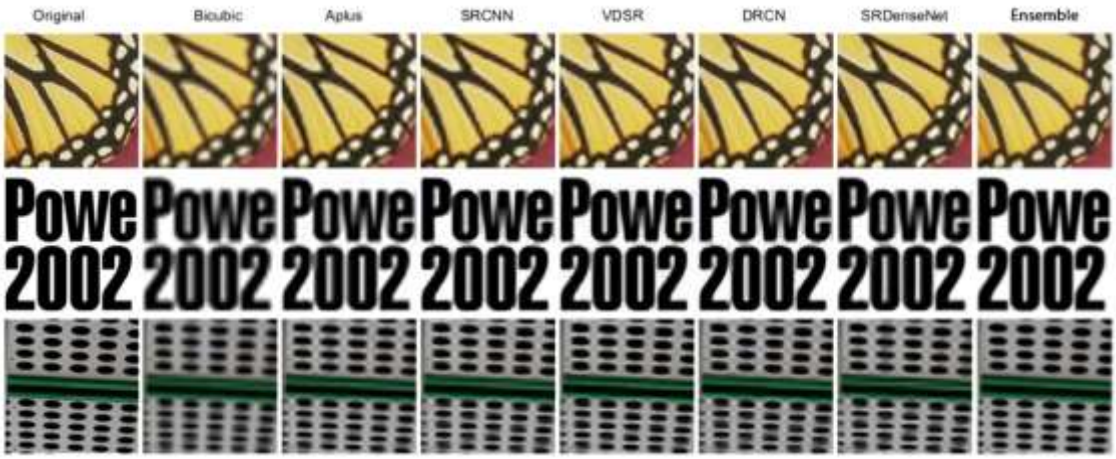


Figure 10 compares reconstructed images obtained from different methods. The results from previous methods show visible distortions and weaker texture recovery, whereas the proposed method reconstructs texture patterns more effectively and minimizes distortions. For example, in the last image of Figure 10, the PSNR is 28.22 and the SSIM is 0.9555, indicating high reconstruction quality for that sample.

AITBSS
E-ISSN: 3041-9433

Figure 11

Results of our proposed method



The results indicate that the blurring effect was reduced, and the color quality remained close to the ground truth, with minimal differences. This slight variation could be attributed to the fact that the color means obtained from the networks were not directly used in the ensemble method, but instead, the optimal results from the networks were combined.

As demonstrated, the model leveraged the extracted knowledge and shared it across different networks, leading to significantly better results compared to previous methods. While individual network weaknesses could negatively impact performance, using a combination of networks helps mitigate these weaknesses, ultimately producing high-quality results.

Table 2

The comparison of the results of our proposed method with those of others using DIV2K and BSDS300 datasets

	nearest	bicubic	Aplus	SRCNN	VDSR	DRCN	SRDenseNet	Proposed method
DIV2K								
PSNR	26.36	28.33	30.17	30.33	31.52	30.76	32.05	33.40
SSIM	0.7542	0.8221	0.8617	0.872	0.8938	0.8784	0.9019	0.9172
MSE	22.21	23.11	22.70	21.594	22.55	21.25	20.47	21.7
BSDS300								
PSNR	24.46	26.29	26.48	26.85	27.92	28.16	27.19	28.13
SSIM	0.7100	0.7586	0.7861	0.7962	0.8174	0.8004	0.8084	0.8497
MSE	23.43	25.57	24.75	23.96	25.41	24.34	23.59	22.05

The use of attention mechanisms in the proposed method significantly enhanced the resolution of fine details in the super-resolution images. The attention layers dynamically prioritized the most important features across different GAN outputs, which facilitated better texture preservation and reduced artifacts in the final image. This improvement was particularly evident in regions with high-frequency details, where conventional methods often produce blurred results.

In addition to pixel-based comparisons, the study used perceptual and structural evaluation criteria to assess reconstruction quality. As shown in Table 2, the proposed method was superior in four of the six reported quantitative criteria: PSNR and SSIM on DIV2K, and SSIM and MSE on BSDS300. Its PSNR on BSDS300 was also competitive, although DRCN produced a slightly higher value. This interpretation resolves the inconsistency in the original wording while preserving the original numerical results.

Overall, the attention-guided dynamic model selection, combined with the ensemble learning of GAN networks, provides substantial improvements in super-resolution accuracy. The dynamic selection process ensures that only

AITBSS
E-ISSN: 3041-9433

the most relevant model outputs are used, while the attention mechanism refines the output, leading to higher fidelity in texture recovery and a reduction in common artifacts.

5. Conclusion and Future Work

This study investigated the integration of ensemble learning and super-resolution for generating HR images from LR inputs. The proposed framework used attention-guided dynamic model selection and CNN-based fusion to combine complementary reconstruction models. The model progressively exploited feature-specific outputs to recover fine details and reduce artifacts. Five candidate SR models were fine-tuned or integrated using the DIV2K and BSDS300 datasets.

The results showed that the proposed method achieved an average PSNR of 33.40 dB and SSIM of 0.9172 on DIV2K, and a PSNR of 28.13 dB and SSIM of 0.8497 on BSDS300. Compared with the baselines reported in Table 2, the method improved structural similarity and produced visually sharper outputs. The earlier claim of a generic 0.1 dB improvement was therefore replaced with dataset-specific results directly supported by the reported table.

Future work may investigate replacing the CNN fusion module with a trainable adversarial or transformer-based fusion network. Additional research may also evaluate computational cost, inference time, and user-controlled output scales, which were not fully analyzed in the present study.

Authors' Contributions

Faraz Mohammadian: Conceptualization, Methodology, Software, Validation, Writing the original draft

Sattar Hashemi and Karamollah Bagherifard: Supervision, conceptualization, validation, investigation, writing, review, and editing.

Samad Nejatian: Consultation, validation, writing, review, and editing.

Declaration

Artificial intelligence tools were used only for language polishing, consistency checking, and formatting support. The study design, equations, figures, tables, numerical results, and scientific interpretation remained based on the original manuscript and author-provided results.

Transparency Statement

Data are available for research purposes upon reasonable request to the corresponding author.

Acknowledgments

The authors thank the providers of the DIV2K and BSDS300 benchmark datasets and the researchers whose open scientific contributions supported this study.

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

This study used publicly available image datasets and did not involve human participants or animal experiments. Therefore, formal ethics approval and informed consent were not required.

References

Agustsson, E., & Timofte, R. (2017). NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and study. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,

Denton, E. L., Chintala, S., Szlam, A., & Fergus, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. Advances in Neural Information Processing Systems,

Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2021). Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, *129*, 1258-1281. https://doi.org/10.1007/s11263-020-01419-7

Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 295-307. https://doi.org/10.1109/TPAMI.2015.2439281

Dong, C., Loy, C. C., & Tang, X. (2016). Accelerating the super-resolution convolutional neural network. European Conference on Computer Vision,

Freeman, W. T., Jones, T. R., & Pasztor, E. C. (2002). Example-based super-resolution. *IEEE Computer Graphics and Applications*, *22*(2), 56-65. https://doi.org/10.1109/38.988747

Glasner, D., Bagon, S., & Irani, M. (2009). Super-resolution from a single image. Proceedings of the IEEE International Conference on Computer Vision,

Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Deep back-projection networks for super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

AITBSS
E-ISSN: 3041-9433

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(6), 1153-1160. https://doi.org/10.1109/TASSP.1981.1163711

Kim, J., Lee, J. K., & Lee, K. M. (2016a). Accurate image super-resolution using very deep convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Kim, J., Lee, J. K., & Lee, K. M. (2016b). Deeply-recursive convolutional network for image super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Lai, W. S., Huang, J. B., Ahuja, N., & Yang, M. H. (2017). Deep Laplacian pyramid networks for fast and accurate super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using Swin Transformer. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops,

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. Proceedings of the IEEE International Conference on Computer Vision,

Mei, Y., Fan, Y., & Zhou, Y. (2021). Image super-resolution with non-local sparse attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

Timofte, R., De Smet, V., & Van Gool, L. (2014). A+: Adjusted anchored neighborhood regression for fast super-resolution. Asian Conference on Computer Vision,

Tong, T., Li, G., Liu, X., & Gao, Q. (2017). Image super-resolution using dense skip connections. Proceedings of the IEEE International Conference on Computer Vision,

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. European Conference on Computer Vision Workshops,

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600-612.

Wang, Z., Chen, J., & Hoi, S. C. H. (2021). Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3365-3387. https://doi.org/10.1109/TPAMI.2020.2982166

Yang, J., Wright, J., Huang, T. S., & Ma, Y. (2010). Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, *19*(11), 2861-2873. https://doi.org/10.1109/TIP.2010.2050625

Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., & Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. *Signal Processing*, *128*, 389-408. https://doi.org/10.1016/j.sigpro.2016.05.002

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. European Conference on Computer Vision,

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,

AITBSS
E-ISSN: 3041-9433