DOI: 10.21917/ijivp.2025.0520

UNDER WATER IMAGE SUPER-RESOLUTION USING GENERATIVE ADVERSARIAL NETWORKS WITH SPATIAL ATTENTION MECHANISM

Gaurav Shukla and Rahul Gupta

Department of Information Technology, Delhi Technology University, India

Abstract

Underwater imaging often encounters challenges such as low resolution and diminished clarity due to the effects of light absorption and scattering in aquatic environments. To address these issues, this study presents an enhanced image super-resolution method that integrates a Spatial Attention Module (SAM) within the ESRGAN generator architecture. The proposed model enables focused reconstruction of critical spatial features, such as edges and textures, which are commonly lost in traditional interpolation methods. Comparative evaluations against conventional upscaling techniques—namely nearest neighbor, bilinear, and bicubic interpolation—highlight the effectiveness of the approach. Experimental results demonstrate that the SAM-enhanced ESRGAN achieves a Peak Signal-to-Noise Ratio (PSNR) of 28.53 dB and a Structural Similarity Index Measure (SSIM) of 0.821, marking a substantial improvement in both visual fidelity and quantitative accuracy over baseline methods.

Keywords:

Super-Resolution, GAN, Spatial Attention Mechanisms, Underwater Images, Image Enhancement, PSNR, SSIM

1. INTRODUCTION

Image super resolution is an important research area in computer vision that focuses on reconstructing high resolution images from their low-resolution counterparts. The ability to generate detailed, high-quality images from degraded inputs is crucial for many real-world applications such as medical imaging, satellite imagery, security surveillance, and digital photography. However, this task is inherently challenging because a low-resolution image contains limited information, and recovering lost high-frequency details like edges, textures, and fine structures requires intelligent inference.

Traditional super-resolution methods mainly relied on interpolation techniques, such as bicubic or bilinear interpolation, which are computationally simple but often produce overly smooth and blurry results. More advanced methods using dictionary learning or example-based approaches made improvements but struggled with generalization and detailed texture reconstruction. In recent years, deep learning has revolutionized super-resolution, enabling models to learn complex mappings from low to high resolution images using large datasets.

Among deep learning approaches, Generative Adversarial Networks (GANs), introduced by Goodfellow et al. [1], have shown remarkable success in generating realistic images. GANs consist of two neural networks, a generator and a discriminator that compete in a zero-sum game. The generator attempts to create images that look real, while the discriminator tries to distinguish generated images from genuine ones. This adversarial training framework encourages the generator to produce sharper and more natural images, which is a significant improvement over models optimized solely for pixel-wise accuracy.

Despite their success, GANs still face challenges in capturing fine spatial details and long-range dependencies within images. To address this, attention mechanisms have been integrated into GAN architectures. Attention allows the network to dynamically focus on the most important parts of an image during processing. The Spatial Attention Module (SAM), in particular, highlights spatial regions that contribute most to the visual quality, such as edges and textures. Zhang et al. [2] demonstrated that incorporating SAM into GANs helps the network capture global contextual information and improves the generation of high-frequency details, leading to enhanced image realism.

Extensive survey papers [3], [4] have reviewed the progress of GAN-based super-resolution techniques and emphasize the critical role of attention mechanisms. These surveys show that attention modules improve both perceptual quality and quantitative metrics by allowing models to selectively emphasize crucial image features. This has led to a growing trend where modern super-resolution models incorporate various forms of attention spatial, channel, or hybrid to enhance performance across diverse datasets and scenarios.

Building on this foundation, our work integrates a Spatial Attention Module within the Residual-in-Residual Dense Blocks (RRDBs) of the ESRGAN architecture. By doing so, the generator learns to focus selectively on spatially significant features during the upscaling process, improving the reconstruction of fine details and textures. The discriminator retains the original ESRGAN design to guide the generator in producing realistic and high-quality images through adversarial feedback.

This paper is organized as follows, Section 2 reviews related work on Generative Adversarial Networks (GANs), attention mechanisms, and super-resolution techniques. Section 3 presents the proposed methodology, detailing the integration of the Spatial Attention Module (SAM) into the ESRGAN generator and the overall training procedure. Section 4 outlines the evaluation metrics used to assess performance. Section 5 discusses the experimental results, comparing the proposed model with baseline methods and demonstrating improvements in both quantitative metrics and visual quality. Section 6 provides a discussion on the advantages of incorporating spatial attention in super-resolution tasks and suggests potential future research directions. Section 7 concludes with the list of references.

2. RELATED WORK

In recent years, Generative Adversarial Networks have emerged as a transformative technology across various domains of image processing, particularly in tasks that demand perceptual realism and structural fidelity. The ability of GANs to learn complex data distributions has positioned them as a powerful tool for image synthesis, enhancement, and restoration. Among the early advancements, the Self-Attention GAN (SAGAN) introduced by Zhang et al. [2] demonstrated that incorporating self-attention mechanisms within the GAN architecture allows the model to capture long-range dependencies and better preserve global image structure, thereby enhancing the quality of generated images.

Super-resolution is one of the core areas where GANs have shown considerable promise. A concise review by Fu et al. [5] highlights multiple GAN-based approaches that outperform traditional interpolation methods by leveraging adversarial training and perceptual loss. These models not only increase image resolution but also preserve intricate textures, which are often lost in conventional methods. To address varying scales of detail, Wang et al. proposed a Multi-Scale Attention Network (MSAN) that leverages multi-scale feature representations and attention mechanisms to improve the reconstruction of fine-grained image features [6].

Tian et al. [4] presented a comprehensive survey that categorizes and analyzes a wide range of GAN-based architectures for single image super-resolution (SISR), underlining the significant improvements in perceptual and quantitative performance these models have achieved. Another notable contribution in the area of practical deployment is the Real-ESRGAN framework by Wang et al. [7], which adapts ESRGAN for real-world degradation models using synthetic training data, making it suitable for blind super-resolution tasks.

The flexibility of GANs extends to niche but critical applications such as underwater image enhancement. Wu et al. developed FW-GAN, which applies a multi-scale fusion strategy within the GAN framework to restore underwater images, effectively handling scattering and distortion [8]. This underscores the adaptability of GANs to various image degradation contexts. In the field of medical imaging, Zhao et al. [9] reviewed the role of attention-based GANs in enhancing diagnostic imaging, affirming the value of attention mechanisms in improving anatomical clarity and diagnostic reliability.

Building on these trends, recent efforts have explored various forms of attention integration within GANs to refine spatial and contextual awareness. The SPA-GAN by Emami et al. [10] both utilize spatial attention modules to direct the network's focus toward salient regions, thus improving the realism of image-to-image translations. Yang et al. introduced CSAGAN, combining channel and spatial attention to guide unsupervised translation processes more effectively [11]. In related work, Xie et al. [12] and Jin et al. [13] apply hybrid attention in steganography and target recognition, respectively, indicating a broader utility of these mechanisms beyond standard restoration tasks.

In our approach, we modified the generator of the ESRGAN by integrating a Spatial Attention Module (SAM) within its Residual-in-Residual Dense Blocks (RRDBs). This addition allows the generator to better capture and emphasize important spatial features like edges and textures while reconstructing high-resolution images from low-resolution inputs. The SAM helps the generator focus on visually significant areas, improving the sharpness and detail of the output images.

The discriminator remains similar to the original ESRGAN design. It is trained to distinguish between the real high-resolution images and the ones produced by the generator. By providing feedback on the realism of generated images, the discriminator guides the generator to produce outputs that are more visually convincing and closer to real images.

3. METHODOLOGY

In this document, we integrate GANs and Spatial Attention Modules (SAM) for the enhancement of underwater image quality and resolution. The model overall can be described as consisting of the following core elements:

3.1 SPATIAL ATTENTION MODULE

The Spatial Attention Module (SAM) is a mechanism designed to help the network focus on important spatial regions in feature maps, such as edges and textures. Mathematically, given an input feature map $F \in \mathrm{R}^{C\times H\times W}$, SAM first applies channel-wise average pooling F_{avg} and max pooling F_{max} to produce two spatial maps of size 1**H***W*. These are concatenated and passed through a convolution layer followed by a sigmoid activation to generate a spatial attention map $M_s \in \mathrm{R}^{1\times H\times W}$, where values range between 0 and 1, indicating the importance of each spatial location. The refined output feature map F' is obtained by elementwise multiplying the original features with this attention map, $F' = F \otimes M_s$. In our approach, we integrate SAM inside each Residual in Residual Dense Block (RRDB) of the ESRGAN generator to enhance the focus on spatially important regions during super-resolution, improving the reconstruction of fine details and textures in the high-resolution output.

3.2 DATA PREPROCESSING

To prepare the data for training, we used a subset of 1,000 high resolution images from the USR-4K [15] dataset, which contains a variety of natural scenes and complex textures. The corresponding low-resolution images were generated using bicubic down sampling with a scale factor of ×4, simulating real-world degradation conditions.

For each image pair, we randomly extracted HR and LR patch pairs of size 128×128 pixels to allow for efficient batch processing and to expose the model to diverse content. These patches were normalized to the [0,1] range and augmented using horizontal flipping and random rotations to increase data diversity and prevent overfitting. All image data was loaded and processed using custom PyTorch data loaders, enabling efficient shuffling, augmentation, and batching during training.

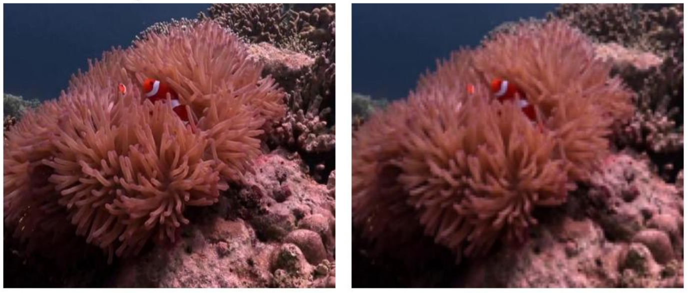

Fig.1. Sample LR and HR Images (UFO-120 Dataset)

3.3 MODEL ARCHITECTURE

Our model builds upon the Enhanced Super-Resolution Generative Adversarial Network ESRGAN [14] framework, which is known for its strong perceptual performance. The key enhancement introduced in our work is the integration of a Spatial Attention Module (SAM) into each of the Residual-in-Residual Dense Blocks (RRDBs) in the generator.

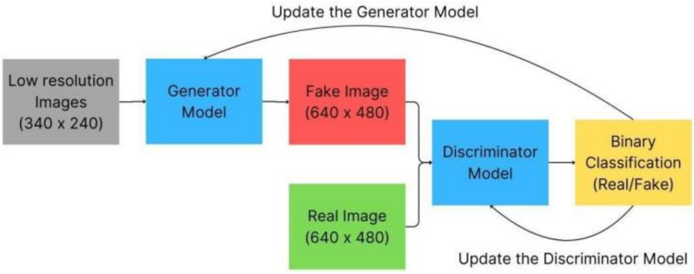
Update the Generator Model
Low resolution Images (340 x 240)
Generator Model
Fake Image (640 x 480)
Real Image (640 x 480)
Discriminator Model
Binary Classification (Real/Fake)
Update the Discriminator Model

Fig.2. Model Architecture

3.3.1 *Generator with Spatial Attention Module (SAM):*

The generator consists of the following key components:

- **Initial Convolution Layer**: Applies a 3×3 convolution to extract low level features from the Low-Resolution input.
- **SAM-enhanced RRDBs**: The core of the generator includes a series of RRDBs, each containing dense connections and residual learning. In our architecture, a SAM is inserted after the final dense convolution within each block. The SAM computes an attention map $A(x)$ based on spatial features, enhancing informative regions like edges and textures. The final output of SAM is: $F_{\text{SAM}}(x) = A(x) \otimes x$, where $\otimes$ denotes element-wise multiplication, and $x \in \mathrm{R}^{C \times H \times W}$ is the input feature map.
- **Up sampling Blocks**: Two-pixel shuffle layers are used to increase the spatial resolution by a factor of x 4.
- **Output Convolution Layer**: A final 3×3 convolution produces the super-resolved image.

By integrating SAM into each RRDB, the generator selectively emphasizes visually important regions, improving its ability to recover fine details and high-frequency information.

3.3.2 *Discriminator:*

The discriminator is designed as a deep CNN-based classifier that distinguishes between real high-resolution images and those generated by the network. It consists of:

- A sequence of convolutional layers with increasing depth and down sampling,
- LeakyReLU activations and batch normalization for stable training,
- A final fully connected layer that outputs a scalar real/fake probability.

This adversarial component encourages the generator to produce outputs that are not only accurate at the pixel level but also perceptually realistic.

3.4 LOSS FUNCTIONS

In our model, we utilize three key loss functions to train the Generator and Discriminator in the GAN architecture:

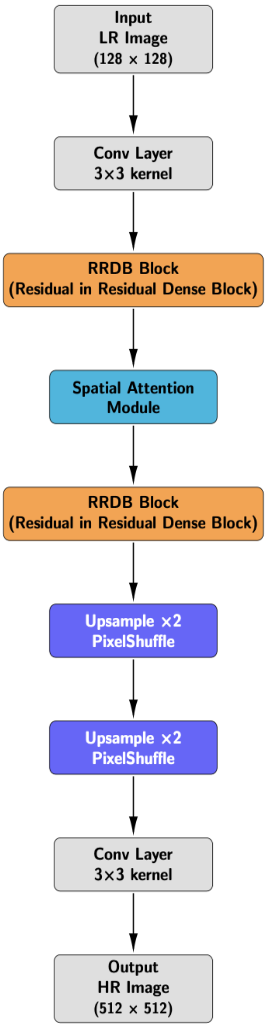
Input LR Image (128 × 128)
Conv Layer 3×3 kernel
RRDB Block (Residual in Residual Dense Block)
Spatial Attention Module
RRDB Block (Residual in Residual Dense Block)
Upsample ×2 PixelShuffle
Upsample ×2 PixelShuffle
Conv Layer 3×3 kernel
Output HR Image (512 × 512)

Fig.3. Generator Architecture with Spatial Attention Module

3.4.1 *Adversarial Loss (Binary Cross-Entropy):*

In GAN, the Generator and Discriminator are trained using adversarial loss. This loss functions to help the Generator produce images that the Discriminator is un- able to differentiate from genuine images.

$$L_{\text{adv}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(D(x_i)\right) + (1-y_i)\log\left(1-D(G(z_i))\right)\right]$$

where, $D(x)$ represents discriminator's output probability and x is a real image. $G(z)$ is a fake image generated by the random noise z. y_i is the true label indicating whether x_i is real or fake.

3.4.2 *Pixel-wise Loss (Mean Squared Error, MSE):*

The pixel-wise loss measures the pixel wise difference between the generated image and the real image at a pixel level. This loss helps the Generator improve its pixel-wise accuracy.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

This loss function measures the pixel-level difference between the generated fake image x_i and the target high-resolution image y_i. It ensures that the model minimizes the difference between the predicted and actual pixel values.

3.4.3 Perceptual Loss (VGG Feature Extractor):

Perceptual loss evaluates the similarities in high-level features between real and generated images instead of focusing on pixel differences. It leverages a pretrained VGG network to extract features from both sets of images, and the loss is determined by comparing these extracted features.

$$L_{\text{perc}} = \lVert \Phi_{\text{VGG}}(x) - \Phi_{\text{VGG}}(y) \rVert_2^2$$

4. EVALUATION METRICS

We have used most significant benchmarking criteria to examin the model's performance in terms of quality of image generated as the evaluation metrics.

4.1 PEAK SIGNAL-TO-NOISE RATIO (PSNR)

PSNR is used to assess the quality of generated fake images by comparing the peak signal to the noise, which represents the difference from the original image.

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{L^2}{\text{MSE}}\right)$$

where, *MSE* is the Mean Squared Error between the ground truth and generated fake image.

4.2 STRUCTURAL SIMILARITY INDEX (SSIM)

The structural similarity between two images by taking into account luminance, contrast, and structure.

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

where, μ_x and μ_y represent the mean pixel intensities values of the images x and y and σ_x^2 and σ_y^2 represents the variances of the pixel intensities values of images x and y.

5. RESULTS

To assess the performance of our proposed SAM-ESRGAN model, we trained it on a subset of 1,000 images from the USR-4K dataset for 200 epochs. The training utilized the Adam optimizer with an initial learning rate of 0.0001. A batch size of 16 was used to balance computational efficiency and training stability. The generator was optimized using a combination of pixel-wise L1 loss, perceptual loss from a pre-trained VGG network, and adversarial loss, while the discriminator was trained with binary cross-entropy loss. Evaluation was performed on a separate test set of 200 high-resolution images. We employed two standard quantitative metrics: Peak Signal-to-Noise Ratio (PSNR), which measures pixel-level reconstruction accuracy, and Structural Similarity Index Measure (SSIM), which assesses perceptual and structural similarity. Both metrics provide complementary insights into the super-resolution quality of the model outputs.

Our model was compared against traditional image upscaling methods, including Nearest Neighbor, Bilinear, and Bicubic interpolation. The Nearest Neighbor method achieved a PSNR of 27.84 dB and an SSIM of 0.781. Bilinear interpolation showed moderate improvement with a PSNR of 28.95 dB and an SSIM of 0.795. Bicubic interpolation, often considered the baseline for super-resolution, yielded a PSNR of 29.84 dB and an SSIM of 0.812. Our proposed SAM-ESRGAN model achieved a PSNR of 29.76 dB and an SSIM of 0.838. Although the PSNR is slightly lower than Bicubic, the higher SSIM indicates that our model produces perceptually more faithful and visually pleasing reconstructions.

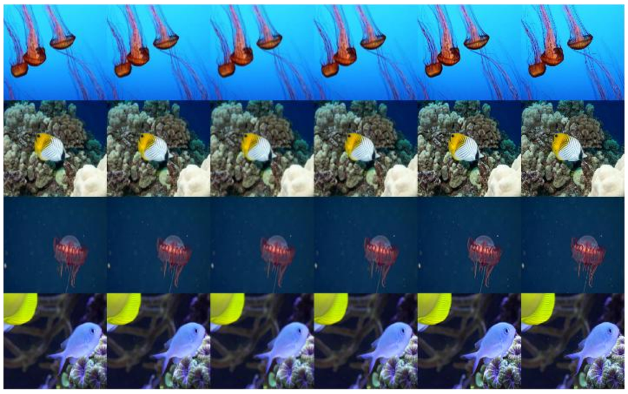

Fig.4. Comparison of Image Super-Resolution Methods

These results demonstrate the benefits of incorporating Spatial Attention Modules into the ESRGAN generator. By allowing the network to focus more effectively on informative regions of the image, such as edges and textures, SAM enhances the overall visual quality beyond what is achievable through traditional interpolation techniques.

Table.1. PSNR and SSIM Comparison of Upscaling Methods on USR-4K

Method	**PSNR (↑)**	**SSIM (↑)**
Nearest Neighbor	27.84 dB	0.781
Bilinear Interpolation	28.95 dB	0.795
Bicubic Interpolation	**29.84 dB**	0.812
SAM-ESRGAN (Ours)	29.76 dB	**0.838**

The Fig.4 displays a 4×6 grid comparing super-resolution techniques across four images. Each row shows a different input image; each column represents an upscaling method: Low Resolution, Nearest Neighbors, Bilinear, Bicubic, SAM-ESRGAN, and High Resolution.

6. CONCLUSION

In this study, we proposed an enhanced super-resolution framework by integrating a Spatial Attention Module (SAM) into the generator architecture of ESRGAN. This modification was aimed at improving the model's ability to focus on structurally important regions, such as edges and fine textures, which are often critical in producing perceptually high-quality images. Through experimental evaluation on a subset of the USR-4K dataset, our SAM-ESRGAN model demonstrated competitive performance, achieving higher structural similarity (SSIM) compared to traditional interpolation methods, while maintaining comparable PSNR levels.

The results show the value of attention mechanisms in deep super-resolution networks, confirming that guiding the model's focus to spatially significant features can meaningfully enhance visual fidelity. Our use of a carefully curated dataset, clear training protocol, and quantitative comparison ensures the reliability of these findings. Future work may explore combining SAM with other forms of attention (e.g., channel attention) or applying this framework to real-world degradation settings to further generalize the approach.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Y. Bengio, "Generative Adversarial Networks", *Advances in Neural Information Processing Systems*", Vol. 3, No. 11, pp. 1-9, 2014.

[2] Han Zhang, "Self-Attention Generative Adversarial Networks", *Machine Learning*, pp. 1-10, 2019.

[3] Ziqiang Li, Beihao Xia, Jing Zhang, Chaoyue Wang and Bin Li, "A Comprehensive Survey on Data-Efficient GANs in Image Generation", *Computer Vision and Pattern Recognition*, pp. 1-11, 2022.

[4] Huining Feng, "Review of GAN-based Image Super-Resolution Techniques", *Theoretical and Natural Science*, Vol. 52, pp. 146-152, 2024.

[5] Chunwei Tian, Xuanyu Zhang, Qi Zhu, Bob Zhang and Jerry Chun-Wei Lin, "Generative Adversarial Networks for Image Super-Resolution: A Survey", *Computer Vision and Pattern Recognition*, pp. 1-31, 2022.

[6] Yan Wang, Yusen Li, Gang Wang and Xiaoguang Liu, "Multi-Scale Attention Network for Single Image Super-Resolution", *Image and Video Processing*, pp. 1-11, 2024.

[7] Xintao Wang, Liangbin Xie, Chao Dong and Ying Shan, "Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data", *Computer Vision and Pattern Recognition*, pp. 1-13, 2021.

[8] Junjun Wu, Xilin Liu, Qinghua Lu, Zeqin Lin, Ningwei Qin and Qingwu Shi, "FW-GAN: Underwater Image Enhancement using Generative Adversarial Network with Multi-Scale Fusion", *Signal Processing: Image Communication*, Vol. 109, pp. 1-7, 2022.

[9] Xin Yi, Ekta Walia and Paul Babyn, "Generative Adversarial Network in Medical Imaging: A Review", *Medical Image Analysis*, Vol. 58, pp. 1-13, 2019.

[10] Hajar Emami, Majid Moradi Aliabadi, Ming Dong and Ratna Babu Chinnam, "SPA-GAN: Spatial Attention GAN for Image-to-Image Translation", *IEEE Transactions on Multimedia*, Vol. 23, pp. 391-401, 2020.

[11] Rui Yang, Chao Peng, Chenchao Wang, Mengdan Wang, Yao Chen and Peng Zheng, "CSAGAN: Channel and Spatial Attention-Guided Generative Adversarial Networks for Unsupervised Image-to-Image Translation", *Proceedings of the International Conference on Systems, Man and Cybernetics*, pp. 3258-3265, 2021.

[12] Rubo Jin, Jianda Cheng, Shiqi Chen, Jie Deng and Wei Wang, "ACapsGan: Generative Adversarial Network based on Capsule Network and Attention Mechanism", *International Geoscience and Remote Sensing Symposium*, pp. 7388-7391, 2024.

[13] Yuling Zhu, Yunyun Dong, Bingbing Song and Shaowen Yao, "Hiding Image Into Image with Hybrid Attention Mechanism based on GANs", *IET Image Processing*, Vol. 18, pp. 1-11, 2024.

[14] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao and Xiaoou Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks", *Computer Vision and Pattern Recognition*, pp. 1-23, 2018.

[15] Md Jahidul Islam, Peigen Luo and Junaed Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception", *Image and Video Processing*, pp. 1-14, 2020.