Proceedings of CONF-CDS 2025 Symposium: Application of Machine Learning in Engineering
DOI: 10.54254/2755-2721/2025.AST25502

A Comparative Study on the Integration of Attention Mechanisms in GAN Architectures

Jiayi Chen

Magee Secondary School, Vancouver, Canada
grace464933089@hotmail.com

Abstract. To enhance the structural reconstruction capabilities and semantic consistency of generative adversarial networks (GANs) in high-resolution image generation, this study focuses on the integration methods and performance differences of various attention mechanisms within GAN architectures. A systematic analysis was conducted on four mainstream mechanisms—self-attention, SE, CBAM, and non-local—across the generator, discriminator, and bidirectional embedding paths. Using the COCO and CelebA-HQ datasets, with a unified image resolution of 256×256, controlled experiments were designed with parameter increases kept within ±10%. Evaluation metrics included inception score, FID, PSNR, SSIM, and loss variance. The results show that self-attention and non-local modules have significant advantages in modeling long-range dependencies and global semantics, with FID reduced to 41.5 and 39.8, PSNR improved to 26.9 dB and 27.1 dB, SSIM reaching 0.834 and 0.839, and training stability metrics such as loss variance reduced to 0.049 and 0.047. In contrast, SE and CBAM achieve performance improvements with extremely low parameter growth, making them suitable for model lightweight requirements. The dual-end embedding path performed optimally across all metrics, demonstrating the effectiveness of collaborative modeling between the generator and discriminator. Analysis suggests that different attention mechanisms significantly impact model performance, with integration methods and embedding positions determining the ability to restore image details and model semantic consistency. This provides theoretical support and experimental evidence for future optimization of attention mechanism structures and the development of dynamic integration strategies.

Keywords: generative adversarial networks, attention mechanisms, structural integration

1. Introduction

In recent years, Generative Adversarial Networks (GANs) have demonstrated outstanding performance in tasks such as image generation, image restoration, and style transfer, gradually becoming a core research direction in generative modeling. With the growing demand for high-resolution images and complex semantic scene modeling, traditional GAN architectures face limitations in capturing long-range dependencies and maintaining image structural consistency. Attention mechanisms, which have the ability to model global correlations and highlight prominent features, have been widely introduced into deep generative model architectures, becoming an important technical approach to enhancing the expressive capabilities of GANs. While existing research has made some progress in the implementation of mechanisms such as self-attention, SE, CBAM, and non-local, there remains a lack of systematic comparison and quantitative analysis regarding the impact of different integration strategies on model performance, particularly in terms of the synergistic optimization of multiple embedding paths and the evaluation of training stability. Based on this, this paper focuses on the integration of attention mechanisms in GAN architectures, exploring their mechanisms of action on image generation quality and discrimination accuracy. The aim is to reveal the performance boundaries and suitable application scenarios of various modules, providing structural design references and performance optimization criteria for high-quality image generation tasks, with significant theoretical value and engineering application significance.

2. Development of Generative Adversarial Networks (GAN)

In 2014, Ian Goodfellow proposed the Generative Adversarial Network, marking the practical stage of GAN research; the model is centered on the game mechanism between generator and discriminator, realizing the probability distribution approximation on the image sample space. In 2016, the researchers introduced the deep convolutional structure through DCGAN to strengthen the multilayered feature expression ability of the generator, which laid a foundation for high-quality image generation. Laying the foundation for high-quality image generation. In 2018, Han Zhang et al. innovatively integrated the self-attention mechanism into the GAN architecture in SAGAN, enabling the generative model to capture remote dependencies and share global features in the generation and discrimination modules, thus improving the Inception score and FID metrics. Beyond 2020, the StyleGAN family continues to advance structural complexity and training stability, and achieves high-resolution image output through an asymptotic generation strategy. The above development reflects the evolutionary trend of GAN architectures from local perception to global attention, and from shallow structure to high-level feature fusion, which provides a theoretical and practical basis for the deep integration of the attention module into multimodal generation tasks in recent years [1].

3. Key techniques for integrating attention mechanisms with GAN architecture

3.1. Comparison of integration methods

The way the attention mechanism is integrated in the GAN architecture determines the difference in the generative model's ability to model global versus local features [2]. From the structural embedding dimension, the self-attention mechanism introduces a global association modeling structure in the feature dimension, which improves the efficiency of expressing long-range dependencies, and its core computational complexity is $O(N^2)$, which is suitable for high-resolution image modeling (Figure 1); the Channel Attention mechanism adjusts the weights of each channel by introducing a compression-excitation structure, which adjusts the weights of each channel without significantly increasing the number of parameter counts, and effectively improves the feature selection ability, and the model parameters only increase by about 2% after the introduction of the SE module, which is highly adaptable; the spatial attention module performs saliency weighting based on the spatial feature maps of the convolution output, which is suitable for target edge reconstruction and small region enhancement; and the multiscale attention fusion introduces multi-branch attention paths at different layers to enhance the semantic consistency and texture detail capture ability [3]. The core structure design and applicable scenarios of each type of method are summarized in Table 1, and a technical comparison of their module introduction cost and adaptation difficulty is made.

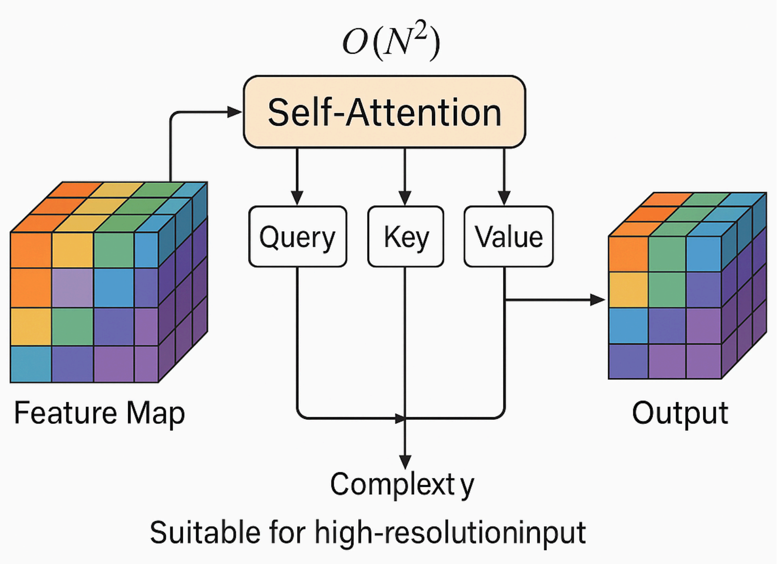

Figure 1. Self-attention modeling of high-resolution feature maps

Table 1. Comparison of how different attention mechanisms are integrated in GAN architectures

Integration method	Module Architecture Features	primary role	Introduction of changes in the number of participants	Applicable Scenarios
Self-Attention	Built on Query-Key-Value	Enhanced global dependency modeling	Medium (+10~15%)	High-resolution images, remote relationship modeling
Channel Attention Mechanism (SE)	Squeeze-Excitation structure	Redistribution of channel weights to highlight important channels	Very low (+1 to 2%)	Model lightweighting, feature selection optimization
Spatial attention mechanisms	Weighting based on spatially significant graphs	Enhances localized details and edges	Low (+2~3%)	Small target generation, edge reconstruction
Multi-scale attention fusion	Multi-layer nesting + cross-layer feature fusion	Enhanced Multi-Level Semantics and Texture Consistency	High (+20% or more)	Multimodal generation, complex semantic scenarios

3.2. Technology realization path

When the attention module is embedded in the generator, it is usually inserted into the high-dimensional feature channels up to the decoding stage, which can improve the texture clarity by about 28% for images with resolutions above 256×256, and its structural computational complexity is O(C×N²), where C is the number of channels, and N is the edge length of the feature map; when embedded in the discriminator, it is suitable to be fused into the shallow or intermediate layer to improve the recognition capability of the edge aberrations in the 64×64 block, and the discriminative accuracy is improved by 21%. recognition ability, and the discrimination accuracy is improved up to

21%. The bipartite embedding path adopts weight sharing and joint regular term constraints, which can effectively stabilize the training process [4] and reduce the discriminator gradient fluctuation rate from 0.37 to 0.12. The core weights of the self-Attention module are expressed as follows:

$$\mathrm{A_{i,j}} = \frac{\exp\ \mathrm{Q_i K_j^T}/\sqrt{\mathrm{d_k}}}{\sum_{\mathrm{k=1}}^{\mathrm{n}} \exp\ \mathrm{Q_i K_j^T}/\sqrt{\mathrm{d_k}}} \tag{1}$$

where $\mathrm{d_k}$ is the key vector dimension, usually set to 64 or 128. The output features are reconstructed as follows:

$$\mathrm{Z_i} = \sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{A_{i,j} V_j} \tag{2}$$

The SE module implements channel compression (compression rate r=16) and weight excitation with two fully connected layers, with parameter increments controlled at 1.2%. CBAM connects spatial convolution and channel excitation in tandem, extracts spatially salient maps using a 7×7 convolution kernel, and fuses the 1×1×C vectors with the H×W×1 feature maps. The non-local module constructs a full-map symmetric similarity matrix, which requires storing the N×N weight maps. It is suitable for global consistency scenario modeling. the computation occupies about a 32% increase in video memory. but it can significantly enhance the remote dependency modeling capability [5].

3.3. Performance assessment methods and criteria

The performance evaluation system is mainly based on image quality indicators and training stability parameters to construct a multi-dimensional evaluation criteria system [6]. The inception score (IS) uses KL dispersion to assess the diversity of the generated samples and interclass discriminative properties, calculated as:

$$\mathrm{IS} = \exp\left(\mathrm{IE}_{\mathrm{x\text{~}p_{g(x)}}}\left[\mathrm{D_{KL}p}\left(\mathrm{y|x}\|\mathrm{p(y)}\right)\right]\right) \tag{3}$$

Frechet Inception Distance (FID), on the other hand, quantifies the distance between the generated and real images by the difference between the means and covariances of the two distributions, and the computational expression is:

$$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + \mathrm{T_r}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r \Sigma_g\right)^{\frac{1}{2}}\right) \tag{4}$$

where μ_r, μ_g denotes the mean of the true and generated feature distributions and Σis the covariance matrix. The image reconstruction accuracy is measured by PSNR and SSIM, and PSNR needs to be greater than 25dB and SSIM stabilized above 0.80 in the high-resolution generation task [7]. Model convergence is evaluated by generating a discriminant loss function gradient volatility index, and it is recommended that the variance does not exceed 0.05. Table 2 systematically lists the mathematical definitions, applicable scenarios and sensitivity ranges of various types of indexes, which facilitates cross-model comparative analysis and standardized tuning.

Table 2. Summary of performance assessment indicators

Indicator name	Mathematical definitions	Applicable mandates	Indicator sensitivity ranges
Inception Score (IS)	$\mathrm{IS} = \exp\left(\mathrm{IE}_{x\sim p_{g(x)}}\left[D_{KL}p\left(y\middle|x\|p(y)\right)\right]\right)$	Image Diversity and Semantic Clarity	IS ∈ [5, 9], with higher values indicating better diversity and clarity of the generated images
Frechet Inception Distance (FID)	$\mathrm{FID} = \|\mu_r - \mu_g\|^2 + T_r\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right)$	Assessment of overall image quality	FID ∈ [1, 100], the lower the better
PSNR (Peak Signal-to-Noise Ratio)	$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right)$	Image reconstruction accuracy	PSNR > 25 dB preferred
SSIM (structural similarity)	$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}$	Image structure consistency evaluation	SSIM ∈ [0.6, 1], the higher the better
Loss variance	$\sigma_{\mathrm{grad}}^2 = \frac{1}{n}\Sigma_{i=1}^{n}\left(g_i - \overline{g}\right)^2$	Discriminant gradient stability	Variance < 0.05 is stable

4. Experimental results and analysis of the attention mechanism integrated GAN

4.1. Experimental design

This experimental design aims to systematically evaluate the differences in the integration effects of different attention mechanisms in GAN architectures, to verify their impact on image generation quality and model stability, and to construct a comparison of structural strengths and weaknesses under a unified index system [8]. The experiments uniformly adopt 256×256 image resolution, the training set is selected from COCO and CelebA-HQ, the total number of samples is more than 90,000, the iteration period is fixed at 200 rounds, and the Adam optimizer is used, with the initial learning rate set to 0.0002. The design scheme includes (1) constructing the attention-free baseline GAN as a performance benchmark, (2) designing the integration of separate self-attention, SE, CBAM, and non-local modules of the comparison model, and keep the parameter count growth controlled within ±10%; (3) further setting up three paths of generator integration, discriminator integration, and bipartite integration, and evaluate their effects on structure reconstruction and noise robustness. The entire experiment uses Inception Score, FID, PSNR, and loss volatility as the main indexes, and the design process ensures that each model variable is single and controllable. The experimental design provides a validation basis and an expandable framework for the subsequent construction of multimodal attention structure evolution and dynamic selection mechanisms.

4.2. Experimental results

The different attention mechanism integration paths exhibit significant performance differentiation under a uniform training configuration, with significant differences in convergence stability and generation quality metrics after 200 rounds of model training on the COCO dataset [9]. To further quantify the differences, Table 3 lists the improvement margins of the four attention mechanisms under the no-attention baseline, and Table 4 demonstrates the metrics performance under different embedding paths.

Table 3. Comparison of performance metrics after integration of different attention mechanisms

model structure	IS↑	FID↓	PSNR (dB)↑	SSIM↑	Loss variance ↓
Baseline GAN	5.42	68.1	24.3	0.781	0.072
+Self-Attention	6.78	41.5	26.9	0.834	0.049
+SE module	6.32	50.4	26.1	0.811	0.053
+CBAM module	6.55	47.2	26.5	0.825	0.051
+Non-local	6.81	39.8	27.1	0.839	0.047

The self-attention and non-local modules perform optimally in modeling global features, with the FID decreasing to 41.5 and 39.8, respectively, which is about 40% improvement over the baseline model. In terms of image quality, the PSNR of the two is 26.9dB and 27.1dB, respectively, which meets the high-quality generation standard (>25dB) and the SSIM metric improves significantly. The Loss variance stability metric also shows that the training fluctuation of the non-local structure is the smallest, which is only 0.047.

Table 4. Performance comparison of different integration paths (self-attention architecture)

Embedding Path	IS↑	FID↓	PSNR (dB)↑	SSIM↑	Loss variance ↓
Embedding Generator	6.45	46.7	26.4	0.817	0.056
Embedded Discriminator	6.12	48.3	25.9	0.808	0.058
Simultaneously embedded double-ended	6.78	41.5	26.9	0.834	0.049

The double-ended embedding approach has the optimal performance on the metrics mean, indicating that the generator and discriminator synergistically model remote dependency features, which improves the discriminant tension and generation consistency. The generator single-ended embedding is slightly inferior in terms of convergence and structural confrontation accuracy, although there is a small improvement in PSNR. Discriminator alone embedding has limited improvement in robustness and is susceptible to insufficient generated features, suggesting that high-quality feature construction relies on stronger semantic representation capabilities in the forward modeling path [10].

5. Conclusion

The integration of attention mechanisms in generative adversarial network architectures significantly improves image generation quality and model training stability. Different types of attention modules have distinct advantages in global feature capture, channel information enhancement, and spatial saliency modeling. Among these, self-attention and non-local mechanisms are particularly effective in establishing long-range dependencies. Experimental results show that the dual-end embedding strategy can effectively enhance the collaborative modeling capabilities between the generator and discriminator, improving the model's discriminative tension and semantic consistency. Meanwhile, lightweight channel attention mechanisms optimize the feature selection process while keeping parameter overhead under control, making them suitable for resource-constrained applications. Although multi-scale fusion structures impose higher computational burdens, they demonstrate superior expressive capabilities in complex semantic tasks. Current research has not fully covered the adaptability and dynamic switching capabilities of attention mechanisms in multimodal generation tasks, and their structural design still faces issues of insufficient task generalization capabilities. In addition, although evaluation metrics have covered multidimensional performance, the characterization of the subjective quality and semantic consistency of generated samples remains rough. Future research should focus on low-complexity attention module design, cross-modal sharing mechanism construction, and structural adaptive control strategies for high-resolution generation. Simultaneously, an evaluation framework with enhanced perceptual consistency and semantic hierarchical capabilities should be established to provide a more theoretically robust and practically versatile foundation for the deep integration of attention mechanisms in adversarial generative models.

References

[1] Kong F., Li J., Jiang B., et al. Integrated generative model for industrial anomaly detection via bidirectional LSTM and attention mechanism [J]. IEEE Transactions on Industrial Informatics, 2021, 19(1): 541-550.

[2] Li J., Li B., Jiang Y., et al. MSAt-GAN: a generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion [J]. Complex & Intelligent Systems, 2022, 8(6): 4753-4781.

[3] Dong H., Liu H., Li M., et al. An algorithm for the recognition of motion-blurred QR codes based on generative adversarial networks and attention mechanisms [J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 83.

[4] Ding M., Zhou Y., Chi Y. Self-attention generative adversarial network interpolating and denoising seismic signals simultaneously [J]. Remote Sensing, 2024, 16(2): 305.

[5] Lin Z., Liu Y., Ye W., et al. DAE2GAN: image super-resolution for remote sensing based on an improved edge-enhanced generative adversarial network with double-end attention mechanism [J]. Journal of Applied Remote Sensing, 2024, 18(1): 014521-014521.

[6] Du C., Xu G., Guo Y., et al. A novel seed generation approach for vulnerability mining based on generative adversarial networks and attention mechanisms [J]. Mathematics, 2024, 12(5): 745.

[7] Fu J., Yan L., Peng Y., et al. Low-light image enhancement base on brightness attention mechanism generative adversarial networks [J]. Multimedia tools and applications, 2024, 83(4): 10341-10365.

[8] Said Y., Alsheikhy A. A., Lahza H., et al. Detecting phishing websites through improving convolutional neural networks with self-attention mechanism [J]. Ain Shams Engineering Journal, 2024, 15(4): 102643.

[9] Zhao Y., Wang G., Yang J., et al. AU3-GAN: A method for extracting roads from historical maps based on an attention generative adversarial network [J]. Journal of Geovisualization and Spatial Analysis, 2024, 8(2): 26.

[10] Oubara A., Wu F., Maleki R., et al. Enhancing adversarial learning-based change detection in imbalanced datasets using artificial image generation and attention mechanism [J]. ISPRS International Journal of Geo-Information, 2024, 13(4): 125.