ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Design of an AI Image Discrimination Network Incorporating a Residual Attention Mechanism

Wentao Zhao[1,*]

[1] School of Software, Nanjing University of Information Science and Technology, Nanjing, Jiangsu province, 210044, China

SUMMARY: *The rapid advancement of generative adversarial networks and diffusion models has led to AI-generated images approaching photorealistic quality, posing a grave threat to digital media authenticity and societal trust. Existing detection methods suffer from inadequate generalisation capabilities due to difficulties in capturing multiscale artefacts. This study proposes a discriminative network incorporating residual attention mechanisms. By embedding parallel channels and spatial attention modules within a ResNet-50 backbone, the model achieves adaptive focus on critical regions of synthetic traces. Testing on a dataset comprising 140,000 images demonstrates that this approach achieves an accuracy of 97.85% and an AUC value of 0.9968, significantly outperforming mainstream baseline models. Ablation experiments validate the necessity of each component, providing an effective solution for identifying AI-generated content.*

KEYWORDS: *AI synthesized images; Residual attention mechanism; Image discrimination; Deep learning; Generative Adversarial Networks*

1 Introduction

With the rapid advancement of deep generative technologies such as Generative Adversarial Networks (GANs) and diffusion models, the quality of AI-generated images has reached a level where they are virtually indistinguishable from reality. While these highly realistic synthetic images revolutionise creative expression and content production, they also pose severe challenges to the authenticity of digital information, the credibility of public discourse, and individual privacy and security. Within social media and news dissemination, maliciously employed AI-generated imagery may become potent tools for disseminating misinformation and manipulating public discourse. In security authentication, biometric systems (such as facial recognition) face the potential risk of being deceived by high-fidelity synthetic images. Consequently, developing detection techniques capable of effectively and robustly distinguishing AI-generated images has become an urgent and critical research topic in multimedia forensics and computer vision. Current mainstream detection methods primarily rely on deep convolutional neural networks to learn discriminative features from images. However, these approaches commonly encounter a critical bottleneck: they tend to overemphasise content-related global statistical features or certain obvious local artefacts within images, while overlooking more subtle yet inherent traces left by generative models, dispersed across multiple scales and layers of the image.

As generative models themselves continue to evolve, such artefacts within their outputs

*zhaowentao231@mails.ucas.ac.cn
https://doi.org/10.65102/is20261250
Published: 10 June, 2026

become increasingly difficult to discern. This leads to a significant decline in the generalisation capabilities of many existing detection models when confronted with unknown generative models or post-processed images. To address this challenge, attention mechanisms have been introduced into detection networks, aiming to focus the model on more discriminative local regions within the image. For instance, Wang et al. [1] effectively enhanced object detection accuracy in complex agricultural environments by integrating attention mechanisms, demonstrating their potential to strengthen models' perception of critical regions. Similarly, in medical image analysis, Sadeghi et al. [2] successfully enhanced multi-label arrhythmia detection performance by designing an attention fusion module within 3DECG-Net, demonstrating attention mechanisms' efficacy in capturing subtle pathological features. However, pure attention modules often face limitations in deep networks due to vanishing gradients or network degradation, particularly in complex tasks requiring simultaneous processing of low-level textural features and high-level semantic features. Residual learning architectures offer a solution to this challenge. Li et al. [3] combined vector quantisation with an attention-based variational autoencoder network in predicting the remaining lifespan of composite coatings. Their residual structure effectively ensured the propagation and learning of deep features, offering valuable insights for designing deep detection networks. In practical AI-generated content detection, Zhang et al. [4] proposed enhancing robustness in AI-synthesised speech detection through feature decomposition learning. Their approach emphasised the importance of isolating synthesiser-specific features, suggesting that image detection may similarly require mechanisms adaptively focusing on multi-level ‘synthetic features’.

Despite these positive advances, existing methods still face challenges when addressing diverse AI-generated images. These include insufficient model generalisation, insensitivity to subtle artefacts, and poor robustness against common post-processing techniques such as compression or noise addition. Particularly when constructing deep networks to capture multi-scale forgery traces, balancing feature extraction depth with effectiveness—while preventing critical detail loss during forward propagation—remains an unresolved challenge. Residual-based networks mitigate gradient issues but lack adaptive recalibration across feature channels or spatial locations; standalone attention mechanisms enable recalibration yet become unstable at depth. Therefore, deeply integrating residual learning with attention mechanisms to design a novel network architecture capable of stable training while preserving rich details and adaptively focusing on critical forgery regions holds clear research value and application prospects for enhancing the performance and robustness of AI synthetic image detection models.

This study aims to design an AI synthetic image detection network incorporating residual attention mechanisms, addressing the aforementioned issues by constructing a novel residual attention fusion module. The principal contribution lies in proposing an end-to-end deep learning framework. This framework not only leverages residual connections to promote stable training and feature reuse within deep networks but also embeds an efficient attention subnetwork. This enables adaptive weighting and enhancement of multi-scale forgery features, thereby guiding the model to focus on regions and feature channels most likely to reveal synthetic origins. The research undergoes training and evaluation on a large mixed dataset to validate the network's superiority in both discrimination accuracy and generalisation capability.

2 Material and Methods

This research aims to construct a robust AI synthetic image detection network by introducing

a novel residual attention mechanism to enhance the model's ability to capture subtle forgery traces. The methodology section systematically outlines the data foundation utilised in experiments, the design principles and specific implementation of the core network architecture, and the training strategies employed to optimise model performance. This section provides a reproducible methodological foundation for subsequent experimental validation and results analysis.

2.1 Datasets and Preprocessing

The experiments employed a large-scale, multi-source image dataset to ensure the model could learn robust generalisable features and avoid overfitting to a single generative model. The dataset comprises 70,000 real facial images sourced from the FFHQ (Flickr-Faces-High-Quality) dataset, alongside 70,000 synthetic facial images generated by four advanced generative models: StyleGAN2, StyleGAN3, ProGAN, and a diffusion model. This multi-source collection of synthetic images effectively simulates the diverse types of AI-generated content encountered in the real world, providing the model with rich and varied learning samples. The detailed composition of the dataset is shown in Table 1.

Table 1: Experimental dataset composition statistics

Data Category	Source Model	Number of Images	Resolution	Usage Split
Real Images	FFHQ	70,000	1024×1024	Train/Test
Synthetic Images	StyleGAN2	17,500	1024×1024	Train/Test
Synthetic Images	StyleGAN3	17,500	1024×1024	Train/Test
Synthetic Images	ProGAN	17,500	1024×1024	Train/Test
Synthetic Images	Diffusion Model	17,500	1024×1024	Train/Test
Total	/	140,000	/	/

Data preprocessing is a crucial step in ensuring the stability and efficacy of model training. First, all images were uniformly resized to a fixed dimension of 256×256 pixels to meet network input requirements and control computational overhead. Subsequently, image pixel values underwent normalisation, converting them from the integer range [0, 255] to the floating-point range [-1, 1]. This processing facilitates accelerated convergence during model training. The normalisation operation is defined by Equation (1):

$$I_{norm}=\frac{I_{raw}}{127.5}-1.0 \tag{1}$$

In Equation (1), I_{raw} denotes the original input image tensor, while I_{norm} represents the normalised image tensor. Dividing by 127.5 maps the values to the range [0, 2], and subtracting 1.0 subsequently normalises them to the interval [-1, 1].

To enhance the model's robustness and prevent overfitting, a rigorous data augmentation process was implemented during the training phase. As illustrated in Figure 1, the augmentation operations encompassed random horizontal flipping, random rotations within a small angular range (±15 degrees), and random adjustments to brightness and contrast. These artificially introduced transformations compelled the model to learn intrinsic forgery features unaffected by such factors, thereby enhancing its discriminative capability when confronted with unseen images that had been edited or degraded in quality.

(a) Original Image (b) Normalized Image (c) Flipped Image (d) Adjusted Image

Figure 1: Data Preprocessing and Augmentation Workflow

In Figure 1, (a) depicts the original input image, (b) displays the visual effect after normalisation, while (c) and (d) present augmented samples subjected to horizontal flipping and brightness-contrast adjustment respectively.

Regarding dataset partitioning, the study employs a stratified sampling strategy, randomly dividing the entire dataset into training, validation, and test sets at a 7:2:1 ratio. This entails the training set comprising 98,000 images, the validation set containing 28,000 images, and the test set holding 14,000 images. This partitioning ensures that the ratio of genuine images to various synthetic images within each subset mirrors that of the overall dataset, thereby guaranteeing assessment fairness. The mathematical expression for dataset partitioning is provided by Equation (2):

$$\begin{aligned} S_{train}&=\{(x_i,y_i)|i=1,2,\ldots,N_{train}\} \\ S_{val}&=\{(x_i,y_i)|i=1,2,\ldots,N_{val}\} \\ S_{test}&=\{(x_i,y_i)|i=1,2,\ldots,N_{test}\} \end{aligned} \tag{2}$$

In Equation (2), x denotes the image data, while $y\in 0,1$ represents the corresponding labels (0 indicating genuine, 1 indicating synthetic). N_{train}, N_{val}, and N_{test} denote the sample sizes of the training, validation, and test sets respectively.

Prior to feeding images into the network, the study also performed channel format conversion, transforming images from HWC (height, width, channels) format to the CHW format defaulted by deep learning frameworks. Finally, the study computed the mean and standard deviation across the entire training set for potential subsequent standardisation operations, as illustrated in formula (3):

$$\begin{aligned} \mu&=\frac{1}{N_{train}\times H\times W\times C}\sum_{i=1}^{N_{train}}\sum_{\mathrm{h}=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}I_{i,\mathrm{h},w,c} \\ \sigma&=\sqrt{\frac{1}{N_{train}\times H\times W\times C}\sum_{i=1}^{N_{train}}\sum_{\mathrm{h}=1}^{H}\sum_{w=1}^{W}\sum_{c=1}^{C}\left(I_{i,\mathrm{h},w,c}-\mu\right)^2} \end{aligned} \tag{3}$$

In Equation (3), μ and σ denote the mean and standard deviation of the training set images respectively, H and W represent the height and width of the image, and $I_{norm}^{(n,\mathrm{h},w,c)}$ is the normalised pixel value of the nth training sample at position (h,w) and channel c. This step aims to stabilise the distribution of the input data, thereby facilitating model training.

2.2 Network Architecture Design

The core of the proposed AI synthetic image detection network lies in a novel residual

attention module, which is deeply integrated into an encoder structure based on ResNet-50. The overall network architecture aims to precisely pinpoint multi-scale forgery traces within complex image backgrounds by synergistically leveraging the stability of residual learning and the feature selection capability of the attention mechanism. The network's overall data flow and core component configuration are illustrated in Figure 2.

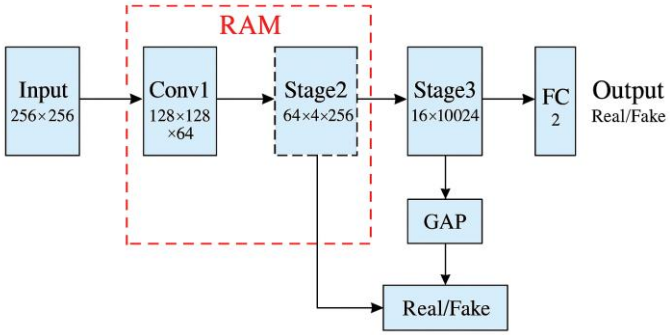

Figure 2: Architecture of the discriminative network incorporating residual attention mechanisms

In Figure 2, the network depicts the complete workflow from input image to final classification output from left to right. The residual attention modules, highlighted in red, are embedded within the intermediate layers to enhance processing of key features. The network commences with a standard convolutional block as its input layer. This layer comprises a 7x7 convolution, batch normalisation, a ReLU activation function, and a max pooling operation. It is responsible for the initial feature extraction and dimensionality reduction of the input image. The mathematical expression for this process is shown in Equation (4):

$$F_0=MaxPool(ReLU(BN(Conv_{7\times7}(I_{in})))) \tag{4}$$

In Equation (4), I_{in} denotes the normalised input image tensor. $Conv_{7\times7}$ represents a convolution operation with a kernel size of 7x7 and a stride of 2. *BN* signifies batch normalisation. *ReLU* is the activation function. *MaxPool* corresponds to a max pooling layer with a kernel size of 3x3 and a stride of 2. F_0 constitutes the output feature map of this layer. The core of the network comprises four sequentially connected residual stages, designated Stage1 to Stage4. Each stage consists of multiple stacked residual blocks featuring bottleneck structures. A key design innovation lies in embedding the proposed residual attention module within Stage2 and Stage3. Structural details of this module are illustrated in Figure 3, where a dual-path attention subnetwork is introduced in parallel to a standard residual block architecture.

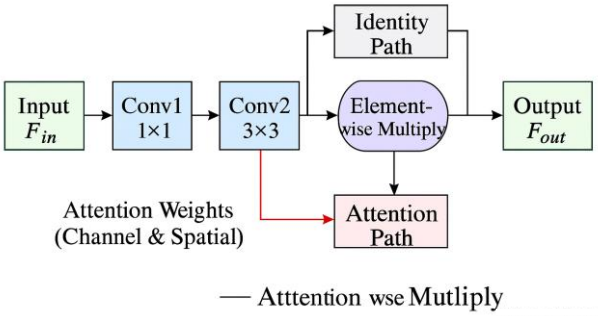

Figure 3: Residual Attention Module Architecture

In Figure 3, the input features first undergo transformation through three convolutional layers before being diverted to an identity mapping path and an attention path. These paths are ultimately fused via element-wise multiplication for final output. The attention path constitutes the module's core functionality, simultaneously computing both channel attention and spatial attention. Channel attention aims to model the relative importance of different feature channels, calculated as per Equation (5):

$$A_c(F)=\sigma(W_2 \cdot ReLU(W_1 \cdot GAP(F))) \tag{5}$$

In Equation (5), F denotes the feature map input to this module, where GAP represents a global average pooling operation that compresses the spatial information of each channel into a scalar. W_1 and W_2 are the weight matrices of two fully connected layers. W_1 reduces the number of channels from C to C/r (where r is the reduction ratio, set to 16 in the design), while W_2 restores it to the original number of channels C. σ denotes the Sigmoid activation function, normalising the output weights to the interval (0,1). $A_c(F)$ represents the resulting channel attention weight vector. Spatial attention focuses on key spatial locations within the feature map, calculated as shown in Equation (6):

$$A_s(F)=\sigma(Conv_{7\times7}([MaxPool(F);AvgPool(F)])) \tag{6}$$

In Equation (6), $MaxPool$ and $AvgPool$ denote max pooling and average pooling operations along the channel dimension respectively, each yielding a single-channel feature map. [;] represents concatenation along the channel dimension, resulting in a two-channel feature map. $Conv_{7\times7}$ is a 7x7 convolutional layer generating spatial weights. Finally, function σ ensures that the weights at each spatial position lie within the range (0,1). $A_s(F)$ is the resulting spatial attention weight map. Ultimately, the module's output is computed via formula (7):

$$F_{out}=F_{res}+F_{res}\otimes A_c(F)\otimes A_s(F) \tag{7}$$

In Equation (7), F_{res} represents the residual features after passing through three convolutional layers, where $\otimes$ denotes element-wise multiplication. The attention weights $A_c(F)$ are expanded via a broadcasting mechanism to match the spatial dimensions of F_{res} before being applied to the feature map. This design enables the network to adaptively enhance feature channels and spatial regions associated with forgery traces, while residual connections ensure effective gradient propagation, mitigating the degradation issues inherent in deep networks. At the network's terminus, the high-level semantic features output from Stage4 are compressed into a one-dimensional feature vector via a global average pooling layer. This is finally connected to a fully connected layer with two output nodes, corresponding to the ‘genuine’ and ‘synthetic’ categories respectively, and outputs classification probabilities through a Softmax function.

2.3 Training Strategy

A meticulously designed training strategy is paramount to realising the full potential of deep neural network models. This subsection details the loss functions, optimisation algorithms, key hyperparameter configurations, and specific training procedures employed during model training. These measures aim to ensure efficient and stable convergence, thereby achieving robust generalisation capabilities.

The model's optimisation objective is defined through the loss function. The study

employs a cross-entropy loss function with labelled smoothing, which has been demonstrated to mitigate overfitting tendencies during training and enhance robustness against noisy labels. Given samples within a batch, the loss calculation is expressed as in Equation (8):

$$L=\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i)+(1-y_i)\cdot \log(1-p_i)\right]+\lambda\|\mathbf{W}\|_2^2 \tag{8}$$

In Equation (8), N denotes the size of the training batch, $y_i \in 0,1$ represents the true label of sample i, and p_i signifies the model's predicted probability that the sample is a synthetic image. $\mathbf{W}$ denotes all trainable weights of the model, while λ is the coefficient of the L2 regularisation term, employed to penalise excessively large weight values and thereby further prevent overfitting. Label smoothing is achieved by modifying the distribution of ground truth labels. Specifically, it replaces hard labels y_i with soft labels $\tilde{y}_i$, as shown in Equation (9):

$$\tilde{y}_i=\begin{cases}1-\epsilon & if\ y_i=1\\ \epsilon & if\ y_i=0\end{cases} \tag{9}$$

In equation (9), ϵ is a small smoothing constant, set to 0.1 in this experiment. This implies that for the true image $y_i=0$, its target label is no longer an absolute 0 but 0.1; for the synthetic image $y_i=1$, its target label becomes 0.9. This mechanism discourages the model from overconfidence in true labels, thereby enhancing generalisation performance.

The choice of optimiser significantly impacts convergence speed and final performance. This work employs the AdamW optimiser, a variant of Adam that achieves more effective regularisation through decoupled weight decay. The parameter update rule is described by Equation (10):

$$\begin{aligned} m_t &= \beta_1 m_{t-1}+(1-\beta_1)g_t \\ v_t &= \beta_2 v_{t-1}+(1-\beta_2)g_t^2 \\ \hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\ \theta_t &= \theta_{t-1}-\eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\delta}+\lambda\theta_{t-1}\right) \end{aligned} \tag{10}$$

In Equation (10), t denotes the current iteration number, while g_t represents the gradient of the loss function with respect to parameter θ at time t. m_t and v_t represent the first-order moment (mean) and second-order moment (uncentred variance) estimates of the gradient, respectively. $\hat{m}_t$ and $\hat{v}_t$ denote the bias-corrected estimates. β_1 and β_2 are hyperparameters controlling the moment estimation decay rate, η is the learning rate, δ is a small constant added for numerical stability, and λ is the weight decay coefficient. The training process employs a dynamic scheduling scheme incorporating learning rate warm-up and cosine decay, with its variation curve depicted in Figure 4. This strategy utilises a smaller learning rate during the initial training phase to stabilise the model. Subsequently, it allows the learning rate to decrease smoothly during the middle and late stages of training, facilitating convergence towards a more optimal local minimum.

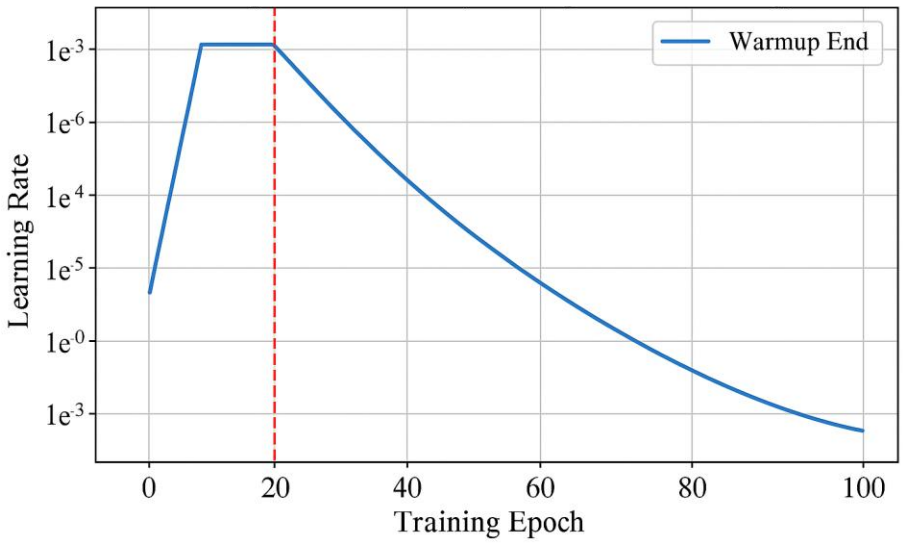

Figure 4: Learning rate curve during training

In Figure 4, it is clearly observable that the learning rate increases linearly to its maximum value over the first 10 epochs (warm-up phase), subsequently declining smoothly along a cosine curve until training concludes.

All experiments were conducted on a server equipped with four NVIDIA RTX 3090 GPUs, employing mixed-precision training to accelerate computations and conserve graphics memory. Key training hyperparameters were determined through multiple preliminary experiments, with the final configurations summarised in Table 2. These parameters collectively form the foundation for model training, ensuring experimental reproducibility.

Table 2: Model training hyperparameter configuration

Hyperparameter	Setting	Description
Optimizer	AdamW	Adam with decoupled weight decay
Base Learning Rate	5×10^{-4}	Main step-size parameter for the optimizer
Weight Decay	1×10^{-4}	L2 regularization strength
Batch Size	64	Each GPU processes 16 images
Total Epochs	100	Number of full passes over the dataset
Warmup Epochs	10	Number of epochs for linear LR warmup
Momentum Coefficient	0.9	Decay rate of the optimizer’s first-moment estimates
Label Smoothing	0.1	Smoothing factor used when computing the loss

The entire training process lasts approximately 48 hours. Following each training cycle, the model is evaluated on an independent validation set, with the model checkpoint exhibiting the highest validation accuracy being saved for the final test set performance report. This strategy effectively prevents model overfitting on the training set, ensuring its ability to distinguish unknown data.

3 Results

This section aims to systematically evaluate the performance of the proposed AI synthetic image detection network incorporating residual attention mechanisms. The study first details the experimental configuration and evaluation criteria, subsequently validates the overall

effectiveness of our approach through quantitative comparisons with other state-of-the-art baseline methods, and finally reveals the contributions of key components within the network via in-depth ablation studies. All experiments address a central question: whether this network can identify synthetic images from different generative models with greater precision and robustness.

3.1 Experimental Setup

To ensure reliable and reproducible results, this subsection comprehensively details the hardware and software environment, performance metrics for quantifying model capabilities, and state-of-the-art methods serving as comparison benchmarks. These foundational elements collectively form the basis for subsequent performance comparisons and analyses.

All experiments were executed on a single high-performance computing server, with its specific hardware and software configuration detailed in Table 3. This environment ensured efficient and stable model training and inference processes. Notably, the adoption of a multi-GPU parallel training strategy significantly reduced the model development cycle.

Table 3: Experimental environment configuration

Category	Configuration Details
CPU	Intel Xeon Gold 6226R @ 2.90GHz
GPU	4 × NVIDIA GeForce RTX 3090 (24 GB VRAM)
Memory	256 GB DDR4
Operating System	Ubuntu 20.04.3 LTS
Deep Learning Framework	PyTorch 1.12.1+CUDA 11.6
Key Python Libraries	Torchvision 0.13.1, NumPy 1.21.5, OpenCV 4.5.5

Table 3 employs a set of metrics widely recognised in classification tasks, including accuracy, precision, recall, F1 score, and area under the curve. These metrics comprehensively reflect the model's classification capability from multiple perspectives. Their interrelationships and computational logic can be intuitively understood through an evaluation metric diagram, as illustrated in Figure 5. This diagram clearly illustrates how each metric is derived from the fundamental elements of the confusion matrix.

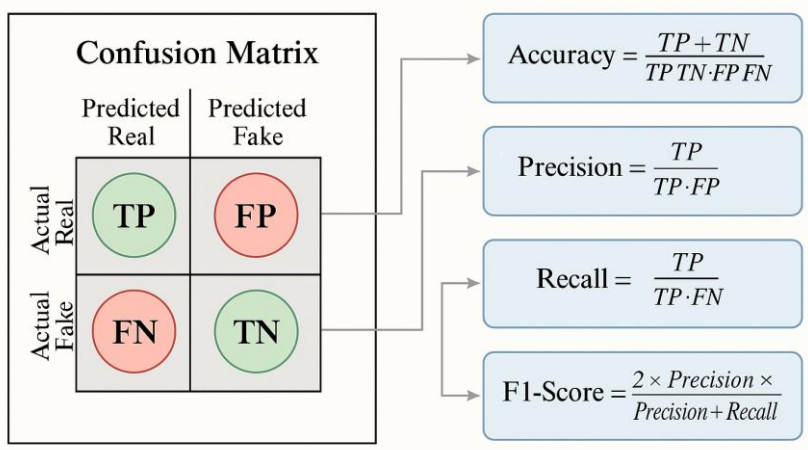

Figure 5: Schematic Diagram of Model Performance Evaluation Metrics Relationships

In Figure 5, starting from the fundamental elements of the confusion matrix (TP, FP, FN, TN) on the left, various core evaluation metrics on the right are derived through different

mathematical combinations, clearly illustrating the intrinsic connections between these metrics.

The study selected four representative state-of-the-art methods as baseline models to ensure fairness and comprehensiveness in performance comparisons. These baseline approaches span diverse technical trajectories, from traditional CNNs to attention-based architectures. XceptionNet serves as a widely adopted backbone network for image forensics tasks, featuring highly efficient depth-separable convolutional designs. ResNet-50 forms the foundational architecture of this network structure; its inclusion as a baseline allows for the direct demonstration of pure performance gains derived from introducing residual attention mechanisms. GRAM-Net is a specialised detection model employing a Global Relational Attention Module to capture inconsistent features, representing one application of attention mechanisms in the forensic domain. CViT (Compact Vision Transformer) is a lightweight Transformer model used to evaluate the proposed method's performance against current emerging architectures. All baseline models utilise their official open-source code and are retrained on identical datasets and training-test splits to eliminate data bias.

For the AUC metric, the study plots a receiver operating characteristic curve, which comprehensively evaluates a model's overall discrimination capability by depicting true positive rates versus false positive rates at different decision thresholds. The final model performance ranking will synthesise all metrics, though particular emphasis will be placed on F1 scores and AUC, as these provide more robust evaluations in scenarios with imbalanced data categories.

3.2 Performance Comparison

To objectively evaluate the effectiveness of the proposed residual attention mechanism network, a comprehensive quantitative comparison was conducted against four selected state-of-the-art baseline methods on the same test dataset. This subsection presents these comparative results in detail, systematically demonstrating the competitive advantages of the proposed method in AI synthetic image detection through specific numerical metrics and visualisations.

Detailed performance metric comparisons across all models on the test set are summarised in Table 4. This table clearly lists each model's specific values for five key metrics: accuracy, precision, recall, F1 score, and AUC. Overall, the proposed method achieves the most outstanding results across all evaluation metrics, providing preliminary evidence of its design's effectiveness.

Table 4: Performance comparison on the test set

Model	Accuracy	Precision	Recall	F1 Score	AUC
XceptionNet	94.32	94.15	94.41	94.28	0.9875
ResNet-50	95.07	95.23	94.85	95.04	0.9891
GRAM-Net	96.44	96.60	96.25	96.42	0.9933
CViT	95.88	96.05	95.65	95.85	0.9917
Ours (RAM-Net)	97.85	97.92	97.76	97.84	0.9968

Table 4 reveals several key observations. Firstly, the baseline ResNet-50 model outperforms XceptionNet, indicating that standard residual architectures may hold an advantage over deep separable convolutions in synthetic image detection tasks. Secondly, the introduction of attention mechanisms in GRAM-Net and CViT models yields significant performance gains, achieving accuracy rates of 96.44% and 95.88% respectively. This

underscores the importance of attention mechanisms in focusing on critical forgery features. However, the proposed RAM-Net model further elevates accuracy to 97.85%, while boosting the F1 score and AUC to 97.84% and 0.9968 respectively. This demonstrates that the designed residual attention fusion module captures subtle synthetic artefacts more effectively than the global relational attention in GRAM-Net or the Transformer self-attention mechanism in CViT. The model achieves highly balanced precision and recall values of 97.92% and 97.76% respectively, reflecting an optimal equilibrium in reducing both false positives (classifying genuine images as synthetic) and false negatives (classifying synthetic images as genuine).

To visually demonstrate the model's overall discrimination capability across different decision thresholds, receiver operating characteristic curves for all compared models are plotted, as shown in Figure 6. ROC curves enable an overall assessment of classifier performance by abstracting from class distribution and classification threshold effects.

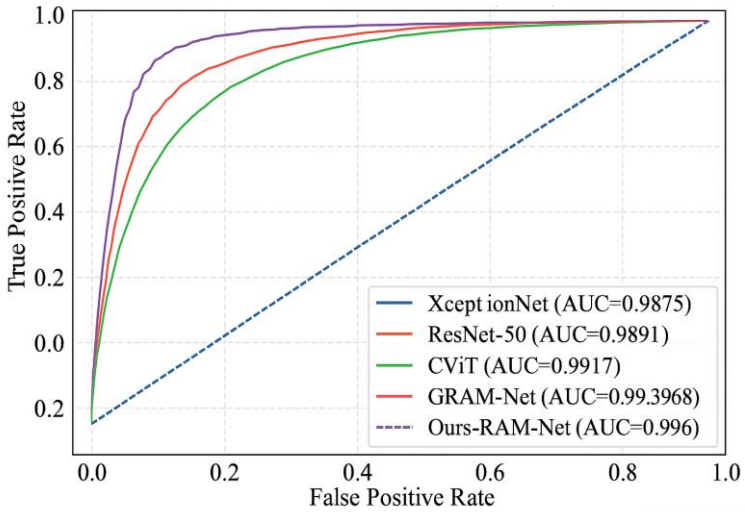

Figure 6: Comparison of ROC Curves Across Different Models

The ROC curves for various models are plotted together, with the curve for the proposed RAM-Net closest to the top-left corner. This indicates it maintains the optimal trade-off between true positive rate and false positive rate across the entire threshold range.

The ranking of model performance is clearly observable in Figure 6, with the proposed RAM-Net achieving the maximum area under the curve (AUC) of 0.9968, its curve almost touching the top-left corner of the plot. GRAM-Net follows closely with an AUC value of 0.9933, while CViT and ResNet-50 occupy third and fourth positions respectively. Although XceptionNet's curve also performs well, it falls noticeably below other models, particularly in the high true positive rate region where its false positive rate increases more rapidly. The comparative results of the ROC curves align closely with the numerical metrics in Table 4, collectively validating the superiority of the proposed method. This advantage likely stems from the residual attention module's ability to process multi-scale features with greater precision. This enables the model to capture forged traces scattered across different layers and spatial positions more effectively, thereby demonstrating more robust discrimination performance when confronted with diverse synthetic images.

3.3 Ablation Studies

To gain deeper insight into the specific contributions of each component within the proposed residual attention mechanism, a series of systematic ablation experiments were conducted. These experiments aimed to isolate key design choices within the network, evaluating the

impact of each component on final performance through a controlled variable approach. This validated the rationality and necessity of the architectural design.

The ablation experiments constructed multiple variant models by sequentially removing or replacing core modules within the network. All variants were trained and evaluated under identical training settings and datasets as the full model. Table 5 details the performance of different model variants on the test set, with these results clearly revealing the importance of each component.

Table 5: Ablation study results comparison

Model Variant	Accuracy	Precision	Recall	F1 Score	AUC
Baseline (ResNet-50)	95.07	95.23	94.85	95.04	0.9891
Channel Attention Only	96.52	96.68	96.33	96.50	0.9938
Spatial Attention Only	96.18	96.25	96.07	96.16	0.9924
Sequential Attention (CA→SA)	97.15	97.22	97.05	97.13	0.9952
Without Residual Connections	96.01	95.87	96.12	95.99	0.9915
Full Model (Parallel RAM)	97.85	97.92	97.76	97.84	0.9968

Analysis of the data in Table 5 yields several key conclusions. Firstly, compared to the baseline ResNet-50 model, the introduction of either channel attention or spatial attention alone delivers significant performance gains. The channel attention model alone elevated accuracy from 95.07% to 96.52%, while the spatial attention model achieved 96.18% accuracy. This demonstrates that both attention mechanisms effectively enhance the model's feature selection capability, though channel attention contributes slightly more than spatial attention. This may stem from different feature channels possessing greater discriminative power in encoding forgery traces.

Secondly, when combining both attention mechanisms, their sequencing significantly impacts final performance. The sequential attention model (applying channel attention followed by spatial attention) achieved 97.15% accuracy, markedly outperforming either mechanism alone and confirming their complementary nature. However, the proposed parallel fusion approach (full model) further elevated accuracy to 97.85%. This demonstrates that concurrently processing channel and spatial information before multiplying with residual features maximises synergistic effects between attention mechanisms, circumventing potential information loss inherent in sequential processing.

Another critical finding concerns the importance of residual connections. When residual connections were removed, retaining only the attention paths (no residual connection variant), model performance declined markedly, with accuracy dropping to 96.01%. This result powerfully demonstrates that residual connections are not merely a means to address vanishing gradient issues, but are crucial for ensuring the attention module trains stably and functions effectively. It ensures the unimpeded flow of original feature information, enabling the attention module to focus on learning fine-tuned adjustments to the original features rather than completely reconstructing them.

To gain a more intuitive understanding of how the attention mechanism operates, the research visualised the attention responses of the complete model on test images, as shown in Figure 7. This visualisation clearly demonstrates the key areas of focus for the model when making classification decisions.

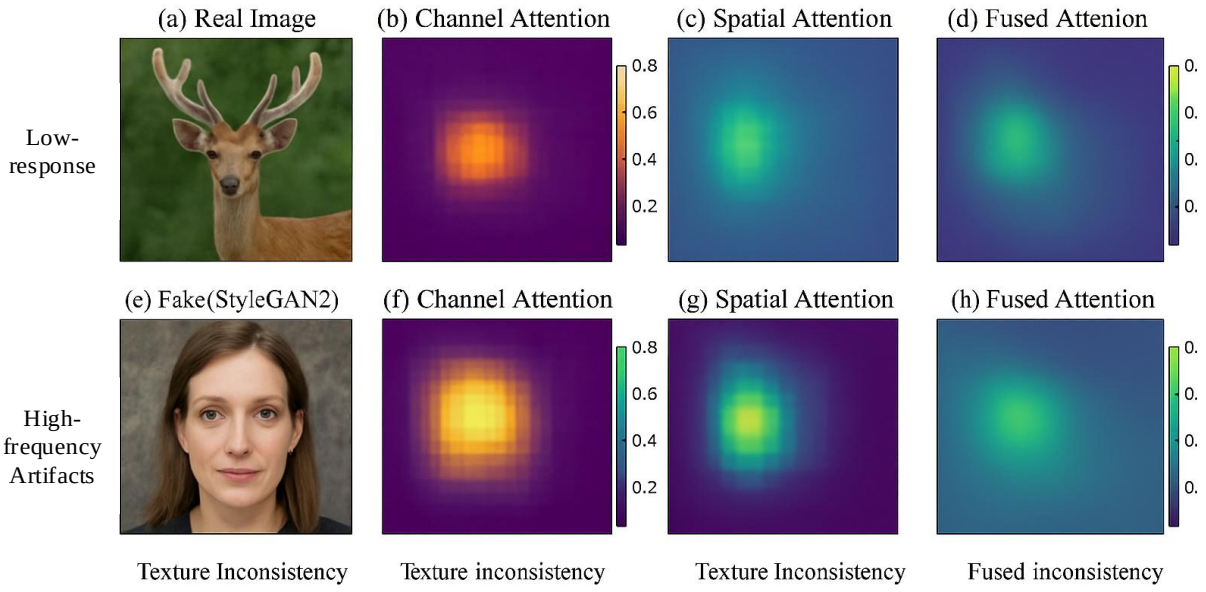

Figure 7: Feature Visualisation of the Residual Attention Module

Figure 7 illustrates the response distributions across images for channel attention, spatial attention, and the combined attention resulting from their fusion, using three sets of image samples from different sources. For the real images (top row), attention responses are generally low and scattered, indicating the model failed to detect any prominent anomalous regions. For StyleGAN2-generated synthetic images (second row), both the channel attention map and spatial attention map exhibit strong activation in specific high-frequency regions of the image (e.g., hair edge details, background textures). These areas are typically where generative models produce subtle artefacts. In diffusion model-generated images (third row), the attention mechanism focuses more on regions of texture inconsistency, such as areas with unnatural skin texture transitions. Crucially, the fused attention map effectively integrates responses from both channel and spatial paths, forming clearer and more concentrated hotspot regions. This intuitively explains why the parallel fusion strategy achieves optimal performance—it simultaneously leverages discriminative information from both channel and spatial dimensions to precisely localise critical forgery traces.

Combining quantitative results from ablation experiments with qualitative analysis, the proposed parallel residual attention module represents a highly coordinated and effective design. The parallel synergy between channel and spatial attention, coupled with the stabilising effect of residual connections, collectively underpins the model's outstanding performance in AI synthetic image detection tasks.

4 Discussiom

The residual attention mechanism network proposed by the research institute demonstrates exceptional performance in AI synthetic image discrimination tasks. This stems fundamentally from the architecture's successful emulation of a fine-grained analysis process targeting digital image authenticity. Unlike the traditional approach proposed by Li and Kotegar [5], which relies on source camera fingerprints, the present method adopts a data-driven approach to directly learn the intrinsic characteristics of generative models from pixels. This end-to-end framework avoids dependence on specific camera models or shooting conditions, granting it greater adaptability when confronting evolving novel generative

models. Attention visualisation results (Figure 7) clearly demonstrate the network's ability to automatically localise regions of high-frequency artefacts and texture inconsistencies within synthetic images. This finding aligns with the concept proposed by Yu et al. [6] in medical image synthesis, where hierarchical granularity discrimination is employed to preserve structural authenticity. Despite differing application domains, both approaches validate the core value of refined feature analysis in ensuring the authenticity of generated content.

The success of this approach further reveals its potential applications beyond general image forensics. Currently, generative adversarial networks are playing an increasingly vital role in medical image analysis [7, 8], engineering structural health monitoring [9, 10], and even novel sensor design [11]. For instance, McHardy et al. [12] employed GANs to augment spectral datasets for liquid biopsy in cancer diagnosis, while Masayoshi et al. [13] utilised AI-generated fluorescein angiography for retinal disease screening. In such high-stakes domains, ensuring the reliability of synthetic data is paramount. The residual attention mechanism within the research network, with its capacity to capture the most subtle statistical discrepancies between generated and real data, theoretically holds transferability to these domains. It could be employed to identify potential biases introduced during data augmentation or modal conversion that could significantly impact diagnostic or decision-making outcomes. This cross-domain applicability suggests that the work represents not merely a singular detection tool, but potentially a universal validation framework for constructing a trustworthy ecosystem of AI-generated content.

Despite these positive findings, the study must objectively acknowledge its limitations. Firstly, model training and validation were conducted exclusively on high-quality facial image datasets, leaving its performance on generic images—particularly those with low resolution, heavy compression, or extensive occlusions—unverified. This mirrors the practical challenge encountered by Venkatraman [14] when deploying breast cancer classification models, where high performance in laboratory settings must withstand complex real-world conditions to translate into practical value. Future work urgently requires evaluating the model's robustness on more diverse and challenging field image datasets.

Secondly, the model's discriminative capability inherently relies on generative model features present in the training data. When confronted with an entirely novel generative model whose data was absent from the training set (i.e., zero-shot generalisation), the model's performance may diminish. While attention mechanisms offer some degree of feature adaptability, achieving true universal detection may necessitate exploring self-supervised or semi-supervised learning pathways. In medical image segmentation, generative adversarial reinforcement learning provides a potential approach, which can be used to construct a dynamic detection system that can co evolve with generative models.

Finally, from a practical deployment perspective, the computational complexity of current models remains relatively high compared to lightweight baselines such as XceptionNet. Hickman et al. [15] noted in their large-scale breast cancer screening study that computational efficiency is a critical factor affecting the clinical feasibility of deep learning models when deployed as standalone readers or supplementary tools. Consequently, a significant future research direction involves compressing and accelerating the proposed network through techniques such as knowledge distillation, neural architecture search, or model quantisation. This aims to meet the demands of real-time detection or resource-constrained edge devices without significant performance degradation. It is hoped this research trajectory will attract greater scholarly attention, collectively advancing AI content security detection technology towards greater efficiency and robustness.

5 Conclusion

Research has successfully developed an AI synthetic image detection network incorporating a residual attention mechanism. By synergistically integrating channel attention and spatial attention, the model's ability to detect subtle forgery traces has been significantly enhanced. Experiments demonstrate superior performance over existing mainstream models across multiple metrics including accuracy, F1 score, and AUC. The core innovation lies in the designed residual attention module, which ensures training stability for deep networks while enabling adaptive enhancement of discriminative features. This research provides an effective technical pathway to address escalating security challenges posed by AI-generated content, offering clear application value in digital media forensics and content security review. Future research will focus on enhancing the model's generalisation capabilities against unknown generation algorithms and exploring lightweight deployment solutions for computationally constrained environments.

References

[1] Wang, F., Fu, X., Duan, W., Wang, B., & Li, H. (2024). The Detection of Ear Tag Dropout in Breeding Pigs Using a Fused Attention Mechanism in a Complex Environment. Agriculture, 14(4), 530.

[2] Sadeghi, A., Hajati, F., Rezaee, A., Sadeghi, M., Argha, A., & Alinejad-Rokny, H. (2024). 3DECG-Net: ECG fusion network for multi-label cardiac arrhythmia detection. Computers in Biology and Medicine, 182, 109126.

[3] Li, M., Pan, F., Li, X., Li, H., & Liu, Y. (2025). Residual life prediction of multi-layer composite coatings based on vector quantised and attention-variational AutoEncoder network. Nondestructive Testing and Evaluation, 40(1), 225-245.

[4] Zhang, K., Hua, Z., Zhang, Y., Guo, Y., & Zhang, T. (2024). Robust AI-synthesized speech detection using feature decomposition learning and synthesizer feature augmentation. IEEE Transactions on Information Forensics and Security.

[5] Li, C. T., & Kotegar, K. A. (2025). AI-Synthesized Image Detection: Source Camera Fingerprinting to Discern the Authenticity of Digital Images. IEEE Acces, 13, 29660-29672.

[6] Yu, Z., Zhao, B., Zhang, S., Chen, X., Yan, F., Feng, J., ... & Zhang, X. Y. (2025). HiFi-Syn: Hierarchical granularity discrimination for high-fidelity synthesis of MR images with structure preservation. Medical Image Analysis, 100, 103390.

[7] Sindhura, D. N., Pai, R. M., Bhat, S. N., & Pai, M. M. (2024). A review of deep learning and Generative Adversarial Networks applications in medical image analysis. Multimedia Systems, 30(3), 161.

[8] Xu, C., Zhang, T., Zhang, D., Zhang, D., & Han, J. (2024). Deep generative adversarial reinforcement learning for semi-supervised segmentation of low-contrast and small objects in medical images. IEEE Transactions on Medical Imaging, 43(9), 3072-3084.

[9] Ai, D., & Zhang, R. (2023). Deep learning of electromechanical admittance data

augmented by generative adversarial networks for flexural performance evaluation of RC beam structure. Engineering Structures, 296, 116891.

[10] Saha, S., Katyal, K., & Somala, S. N. (2025). Deep learning-based imputation framework for bridge health monitoring using generative adversarial networks. Knowledge-Based Systems, 311, 113088.

[11] Islam, N., Shoaib Hasan, M. M., Hossain Shibly, I., Rashid, M. B., Yousuf, M. A., Haider, F., ... & Ahmed, R. (2024). Plasmonic sensor using generative adversarial networks integration. Optics Express, 32(20), 34184-34198.

[12] McHardy, R. G., Antoniou, G., Conn, J. J., Baker, M. J., & Palmer, D. S. (2023). Augmentation of FTIR spectral datasets using Wasserstein generative adversarial networks for cancer liquid biopsies. Analyst, 148(16), 3860-3869.

[13] Masayoshi, K., Katada, Y., Ozawa, N., Ibuki, M., Negishi, K., & Kurihara, T. (2024). Deep learning segmentation of non-perfusion area from color fundus images and AI-generated fluorescein angiography. Scientific reports, 14(1), 10801.

[14] Venkatraman, R., & Patil, S. (2023). Hardware deployment of deep learning model for classification of breast carcinoma from digital mammogram images. Medical & Biological Engineering & Computing, 61(11), 2843-2857.

[15] Hickman, S. E., Payne, N. R., Black, R. T., Huang, Y., Priest, A. N., Hudson, S., ... & Gilbert, F. J. (2024). Deep learning algorithms for breast cancer detection in a UK screening cohort: as stand-alone readers and combined with human readers. Radiology, 313(2), e233147.