Low-Level and Attention-Enhanced GAN Framework for Facial Forgery Detection and Forensics

Mingzhen Zhang
School of Artificial Intelligence, Zhengzhou Railway Vocational & Technical College, Zhengzhou, 451460, China
E-mail: zhangmingzhenzhang@163.com

Keywords: facial information, forgery, forensics, GAN; attention mechanism

Received: April 10, 2025

With the rise of deepfake technologies, detecting fake facial images has become more difficult. Therefore, a forensic algorithm based on color and noise features is developed using generative adversarial networks for single facial forgery images to optimize extraction accuracy and efficiency. The multi-prediction partition spatial attention mechanism is simultaneously fused, and a complex processing facial forgery image forensics model is designed for multi-image processing, which improves the model's attention to forgery areas. The experimental results showed that the model could detect F1 scores of up to 94.21% for a single image, which was improved by 5.97% and 9.03% on the Celeb-DF dataset compared with Xception-DeepLab and DenseNet, respectively. The F1 score on the DFDC dataset was 93.02%, which was also 11.4% and 14.68% higher than the two mentioned above. The average forensic time was 0.29 seconds, which was significantly better than EfficientNet (0.51 seconds) and DenseNet (0.65 seconds). In the multi-image forensics task, the Area under the Curve (AUC) was the highest at 85.74% and the model complexity was the lowest at 80.54%, and the forensics latency was the shortest at 0.28 seconds, which was comprehensively better than the three mainstream comparison methods. This indicates that the proposed model can provide higher detection performance in fake images with different qualities and noise interference, and can provide an effective solution for the security verification and protection of facial information in future networks.

Povzetek: Članek predstavi LLF-MPPSA-GAN, dvo-vejični forenzični model za prepoznavanje ponarejenih obrazov. Združuje nizkonivojsko barvno-šumno analizo in večnapovedno prostorsko pozornost ter dosega odlične rezultate z latenco 0,28 s in visoko robustnostjo na šum.

1 Introduction

In recent years, technologies such as facial generation, face swapping, and enhancement have been widely used in film and television production, virtual reality, intelligent interaction, and other fields, bringing many conveniences to related industries. However, these technologies are also abused by criminals for malicious purposes such as creating false information, identity impersonation, and fraud, posing serious challenges to social public safety and personal privacy [1-2]. Especially, with the promotion of deep forgery technology, the generated fake facial images and videos are becoming increasingly realistic, making it difficult for traditional manual identification methods and low-level feature-based detection methods to effectively recognize, which poses new challenges to digital media forensics and information security. Zhu et al. designed a method based on 3D decomposition to highlight hidden forgery details to improve the effectiveness of existing facial digital information forgery detection. This method was more robust than traditional methods and had higher detection accuracy for fake facial images [3]. Ding et al. found that the deepfake technology of forged faces has posed a threat to electronic payments and identity verification. A countermeasure against deep forgery anti-fingerprint attacks was built. The faces under this strategy had high distinguishability from real faces [4]. Lan et al. adopted discrete cosine transform to perceive forgery trace features in the frequency domain to improve the detection level of facial forgery image information. A deep facial forgery forensics model with frequency domain and noise features was constructed. The model exhibited high forensic accuracy in multiple databases [5]. Liu et al. built a trajectory removal network based on adversarial learning to enhance the effectiveness of facial forgery forensics in deep forgery technology. The proposed trace removal method could reduce the detection accuracy of six state-of-the-art deep forgery detectors, thereby achieving efficient forensic results [6].

El-Shafai et al. proposed an adaptive unsupervised forgery image forensics algorithm by combining recurrent neural networks and multi-scale convolutional networks. The new method had higher accuracy and robustness compared with traditional methods in image and video forgery forensics [7]. Lai et al. proposed a new active forensics method that utilized pseudo-Zernike moment robust watermarking to embed

information into non-facial regions of video frames to enhance the facial swapping detection. This method had superior robustness to standard signal processing operations and excellent performance in detecting deep forgery operations [8]. Sharma et al. proposed a novel verification method to improve the authenticity and consistency judgment level of existing digital image tampering detection in digital photos. After combining the dataset standardization, the Generative Adversarial Network (GAN) was optimized. The experimental results showed that this method exhibited excellent processing accuracy and efficiency in verifying multiple facial digital photo information in forensic investigations, criminal investigations, and intelligence systems [9]. Video stitching forgery is an object-based intra frame forgery operation. Li et al. believed that stitched videos typically contained two different types of camera sensor mode noise. Accordingly, a video stitching detection and localization strategy based on camera fingerprints was proposed to address these two types of noise. This scheme could locate the tampered area and had high detection accuracy [10]. The summarized results for each method are shown in Table 1.

Table 1: Summary table of different methodologies

Method/Model	Description	Metrics/Advantages	Limitations
Zhu X et al. (3D Decomposition)	3D decomposition highlights forgery details	Acc≈91%, robust	Not noise-tolerant
Ding F et al. (Anti-fingerprint)	Countermeasure against fingerprint attacks	Recognition ↑ to 88%	Weak detail detection
Lan G et al. (Freq+Noise)	Frequency-domain forgery feature extraction	AUC≈90%, multi-dataset	Sensitive to low-freq
Liu C et al. (Trace Removal)	Trace removal to degrade detectors	Accuracy ↓ by 15%	Not a detection method
El-Shafai W et al. (RNN+CNN)	Unsupervised fusion of RNN and CNN	Image/video accuracy >89%	High training cost
Lai Z et al. (Watermarking)	Non-face watermark for swap detection	Deepfake detection ↑ to 91%	Requires watermark embedding
Sharma P et al. (Improved GAN)	Standardization + improved GAN	Forensic F1 score≈90%	Poor generalization
Li Q et al. (Camera Fingerprint)	Camera nose for splicing detection	Localization accuracy >92%	Limited applicability

In summary, some progress has been made in deep forgery forensics, with some methods improving detection accuracy and robustness through frequency domain feature extraction, adversarial learning, and watermark embedding. However, these methods still have certain limitations when facing complex forgery techniques, lighting, and resolution changes, especially on low-level feature extraction and multi-region fusion. Therefore, an improved GAN facial forgery forensics method that combines low-level feature extraction and partition space attention mechanism is proposed, aiming to further enhance the practical application value of facial forgery forensics and provide an effective auxiliary means for subsequent forensic work. The innovation of the research lies in optimizing color and noise feature extraction in single facial forgery detection, and introducing a multi-prediction partition spatial attention mechanism in multi-facial forgery detection, which improves the model's attention to forgery areas. In addition, the study adopts an efficient feature fusion strategy to optimize the accuracy and computational efficiency of evidence collection in complex environments. Compared with existing methods, this model performs stably under different levels of noise and image quality. The study proposes a GAN-based dual-architecture model that can handle single- and multi-sided forgery problems under different noise and quality conditions, utilizing underlying features and spatial attention to improve detection performance.

2 Methods and materials

2.1 Single facial forgery image forensics algorithm based on low-level features

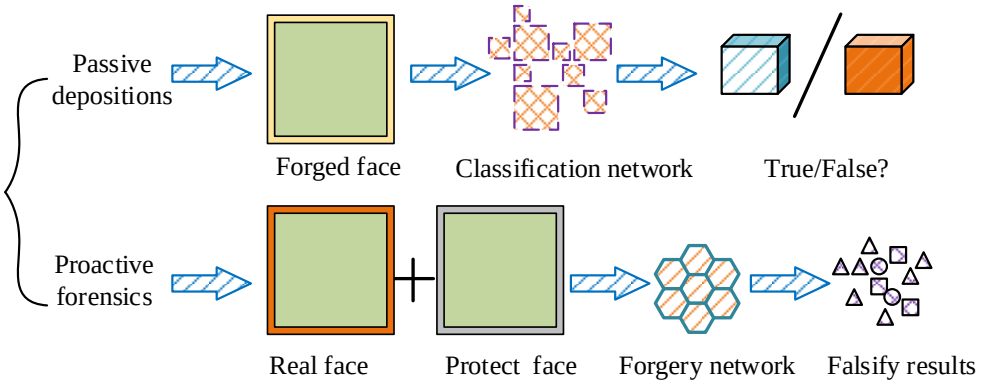


Figure 1: Forged facial image forensics technology principle.

In Figure 1, the basic framework of facial forgery image forensics has passive and active forensics. Passive forensics mainly takes a hierarchical network to classify input images, determine whether they are fake faces, and analyze them based on subtle differences in the images. Active forensics collection involves verifying the authenticity of input images, identifying forged images by comparing stored real facial images, and further detecting them through a forged network [15]. However, in cases where the image quality is high or there are minimal traces of forgery, traditional forensic algorithms may encounter recognition difficulties. In addition, the subtle changes in low-level features such as color and noise features in forged images are often overlooked, resulting in less-than-ideal detection performance of forged images [16]. Therefore, based on the GAN framework and optimized color and noise features as key features, a Low-level Feature-Generative Adversarial Network (LLF-GAN) based on GAN for facial forgery image forensics is proposed, as shown in Figure 2.

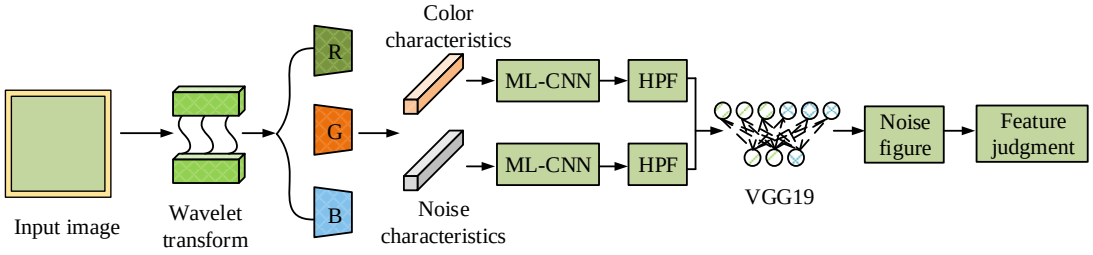


Figure 2: LLF-GAN algorithm framework (Discriminator-based model inspired by GAN structure).

In Figure 2, the LLF-GAN framework mainly consists of three core parts, i.e., the feature extraction module, the classifier module, and the final discriminant module. First, the input image is preprocessed by wavelet transformer and decomposed into three color channels, R, G, and B, respectively. On this basis, color features and noise features are extracted for each channel, respectively. Subsequently, the extracted color and noise features are jointly input into Multi-layer Convolutional Neural Network (ML-CNN), and the embedded High-Pass Filter (HPF) is used to further enhance the detailed features and edge texture, and eliminate the low-frequency background interference. In other words, ML-CNN and HPF are not directly applied to the original image, but are used to jointly process and enhance the extracted color and noise features. Then, these processed fused features are fed into a Visual Geometry Group 19-layer network (VGG19)-based classifier for deep feature learning and forgery discrimination. Finally, the classifier outputs the forgery probability of the image to determine the authenticity of the facial image. Assuming the image is in RGB format, color features can be extracted by converting it to HSV or YCbCr color space. The image color feature extraction is shown in equation (1).

$$C_{color} = \sum_{i=1}^{M}\sum_{j=1}^{N}\left| R_{i,j} - \frac{R_{avg}}{\sum_{k=1}^{M}\sum_{l=1}^{N} R_{k,l}} \right| \quad (1)$$

In equation (1), C_{color} represents the color feature of the image. $R_{i,j}$ represents the red channel value of the i-th and j-th pixels. R_{avg} represents the average value of the red channel in the image. M and N signify the width and height of the image. The noise capture is performed through local contrast and local noise, as displayed in equation (2).

$$C_{noise} = \sum_{i=1}^{M}\sum_{j=1}^{N}\left| \frac{I_{i,j} - I_{avg}}{I_{avg} + \varepsilon} \right| \quad (2)$$

In equation (2), C_{noise} represents the noise feature. $I_{i,j}$ represents the intensity values of the i-th and j-th pixels. I_{avg} represents the average intensity of the image. ε represents a small constant term. The fused low-level features is shown in equation (3).

$$F_{fusion} = \frac{\alpha C_{color} + \beta C_{noise}}{\alpha + \beta} \quad (3)$$

In equation (3), F_{fusion} represents the fused feature. α and β respectively represent the weight factors of

color features and noise features. In addition, as an important part of the entire algorithm framework, ML-CNN is shown in Figure 3.

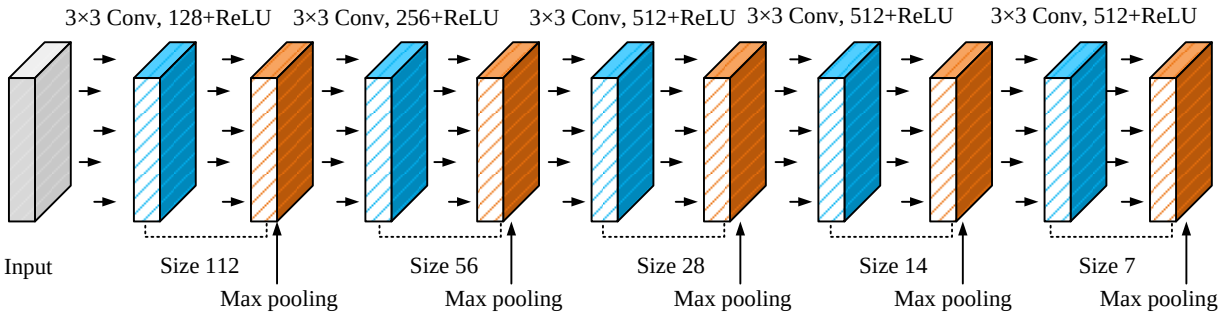


Figure 3: ML-CNN structure.

In Figure 3, the ML-CNN structure includes a combination of multiple convolutional layers and pooling layers, with each convolutional layer using a 3×3 convolution kernel and non-linear mapping processing through ReLU activation function. Each convolutional layer is followed by a 2×2 max pooling layer to lower the feature map size, and reduce computational complexity, and preserve important spatial information. After the input layer, ML-CNN performs a series of convolution and pooling operations on images with a size of 112×112, gradually extracting image features to more abstract levels, and ultimately obtaining high-dimensional features that can be used for classification. The ML-CNN feature extraction is shown in equation (4).

$$C_{conv} = \sum_{I=1}^{H}\sum_{j=1}^{W} W_{i,j} \cdot F_{i,j} + b \qquad (4)$$

In equation (4), C_{conv} signifies the feature after convolution operation. $W_{i,j}$ represents the convolutional kernel. $F_{i,j}$ represents the color and noise features after fusion processing. H and W signify the height and width of the input image. b signifies the bias term. The classification calculation for forged images in the classifier is shown in equation (5).

$$P(y=c|\ x) = \frac{e^{W_c^T F_{fusion} + b_c}}{\sum_{c'} e^{W_{c'}^T F_{fusion} + b_{c'}}} \qquad (5)$$

In equation (5), $P(y=c|\ x)$ signifies the probability that the image belongs to category c. W_c and b_c signify the weights and bias terms of the corresponding category. The final formula for determining the output face image at this point is shown in equation (6).

$$Output = \arg\max(P(y=c|\ x)) \qquad (6)$$

In equation (6), *Output* represents the output of the classifier. If the probability of $P(y=c|\ x)$ is high, it indicates that the type of image is forged.

2.2 Construction of a forensic detection model for multi-facial forgery images in complex scenarios

After constructing the forensics algorithm design for single facial forgery image, the research found that when the complexity of forgery image increases or in different environmental conditions, such as lighting changes, posture changes and image resolution, the traditional single feature and single model methods have certain challenges [17-18]. Specifically, a single prediction method based on low-level features may lead to misjudgments when processing high-quality fake images due to small differences in color and noise features [19]. To enhance the performance of the single forensic algorithm, a facial forgery image forensics method based on Multi-Prediction Partitioned Spatial Attention-Generative Adversarial Network (MPPSA-GAN) is proposed. This method introduces multiple sub-models for multi-angle prediction and combines partition spatial attention mechanism to better focus on forgery areas in the image. The framework structure of MPPSA-GAN is presented in Figure 4.

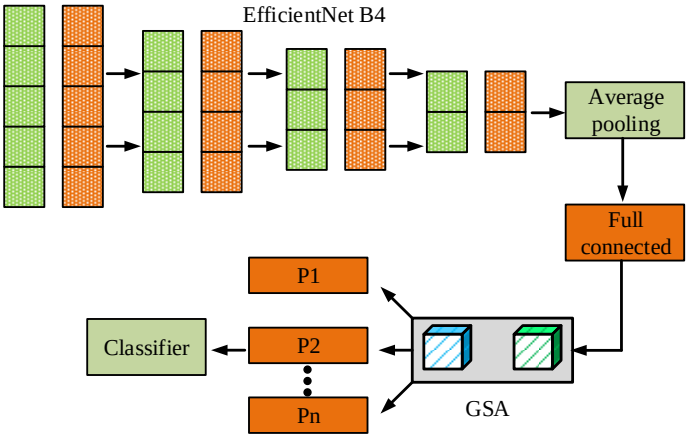


Figure 4: Frame structure of MPPSA-GAN (Discriminator-based model with partitioned attention, inspired by adversarial feature modeling).

In Figure 4, MPPSA-GAN has three main components: feature extraction module, multi-stage prediction module, and partition space attention module. Firstly, the input image undergoes feature extraction through the backbone network of Efficient Neural Network-B4 (EfficientNet-B4) to obtain preliminary image features. Assuming that the input image is I and the preliminary features obtained from feature extraction are F, the r predicted by each sub-model is presented in equation (7).

$$P_r = \sigma\left|\sum_{i=1}^{n} \omega_i \cdot f_i(r)\right| \tag{7}$$

In equation (7), ω_i represents the weight coefficient of each sub-model. $f_i(r)$ signifies the feature output of the i-th sub model on region r. σ represents the sigmoid activation function. P_r signifies the predicted probability of forgery in the region. To further enhance the spatial attention ability to the forged region, the Grouped Spatial Attention (GSA) mechanism is introduced to assign spatial features to each channel separately. The output of the j-th channel in region r is $f_j(r)$ and the spatial attention coefficient is α_j. The spatial attention aggregated feature value of region r is shown in equation (8).

$$A_r = \frac{\sum_{j=1}^{m} \alpha_j \cdot f_j(r)}{\sum_{j=1}^{m} \alpha_j} \tag{8}$$

In equation (8), m represents the number of features. The local feature weighting process fed to each submodel is used to enhance the information representation in the region of interest of the forgery by calculating the attention map A_r, i.e., the attention value A_r outputted by the GSA is used as a feature channel weighting factor embedded in the prediction paths of all sub-models to update the feature representations in their regions. Finally, all predicted results are fused and finally judged by a classifier to output the authenticity of the image. The module structure of GSA is shown in Figure 5.

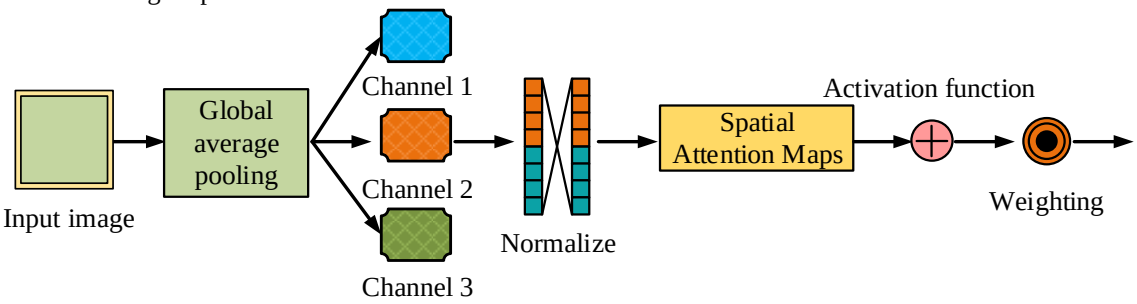


Figure 5: Module structure of GSA.

In Figure 5, the GSA module structure mainly consists of multiple processing units. Firstly, the input feature map is subjected to global average pooling to obtain the global information of each channel. Then, normalization is performed to adjust the scale of the feature map. Next, the spatial attention map is calculated to weight the features of different regions. The weighted feature map is used for subsequent processing. The entire process effectively captures important spatial regions through spatial attention mechanisms and enhances the expressive ability and performance in feature fusion. To enhance the fused region confidence calculation, the study introduces the intra-region feature scoring mechanism. The feature responses $f_k(r)$ of all channels in region r are combined with the scoring coefficients β_k to calculate their aggregation scores S_r. The calculation is shown in equation (9).

$$S_r = \sum_{k=1}^{m} \beta_k \cdot \left(\frac{f_k(r)}{\max_{r'} f_k(r')} \right) \quad (9)$$

In equation (9), r' represents the set of all regions. Finally, to realize the overall determination, all the region prediction results P_r are fused with the region weighting coefficients γ_r to output the forgery probability P -rate of the overall image, as shown in equation (10).

$$P = \frac{\sum_{r=1}^{N} \gamma_r \cdot P_r}{\sum_{r=1}^{N} \gamma_r} \quad (10)$$

In equation (10), γ_r denotes the global importance weighting factor of region r, which is usually calculated by combining r with region scoring S_r in the GSA module. The process embodies a step-by-step weighting mechanism from feature channel to region prediction, with α_j for spatial attention, β_k for channel scoring, and γ_r for region fusion, overall forming a hierarchical and clear chain of attention determination. By integrating the misclassification of forged images with the spatial attention weighting mechanism of the model, a composite loss function is constructed, as defined in equation (11).

$$L_{ce} = -\sum_{r=1}^{N} \left(y_r \cdot \log(P_r) + (1 - y_r) \cdot \log(1 - P) \right) \quad (11)$$

In equation (11), y_r represents the true label of region r, where forged is 1 and true is 0. L_{ce} represents cross entropy loss. The research combines the LLF-GAN forensic algorithm for single facial forgery images and the MPPSA-GAN algorithm for multiple images to propose an improved GAN-based complex facial forgery image forensic algorithm. The algorithm flow is shown in Figure 6.

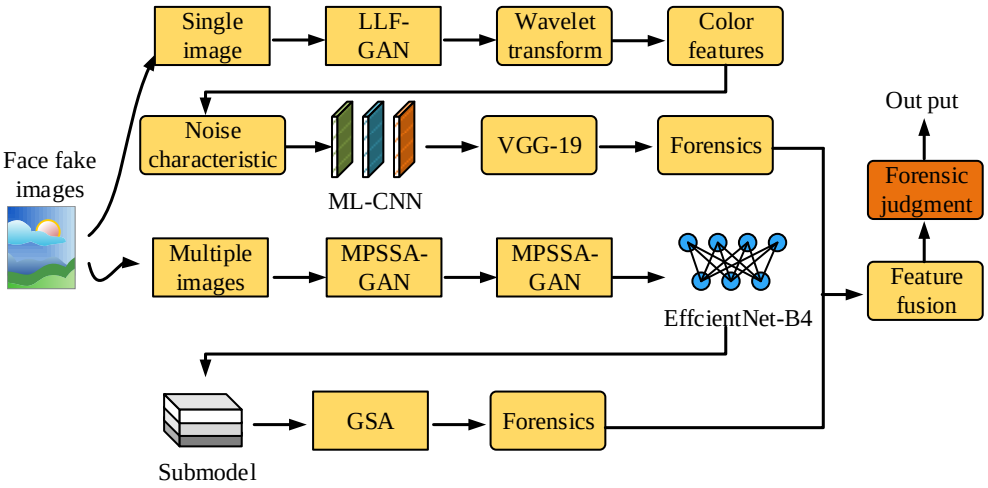


Figure 6: The unified architecture of the LLF-MPPSA-GAN model combining low-level and spatial-attention-based forensic branches.

As illustrated in Figure 6, the algorithm first conducts category-specific feature routing on the input image. Then, they are fed into two feature extraction branches, respectively. If the image is a single facial type, it will be processed through the LLF-GAN path, and low-level features such as color and noise will be extracted by wavelet transform. The feature reinforcement will be carried out by ML-CNN, and fed into the VGG19 classifier to complete the preliminary discrimination. If the image is a multi-facial or a more complex structure type, it is fed into the MPPSA-GAN path, and global semantic features will be extracted by the EfficientNet-B4. Multiple sub-models will process the image partition independently. EfficientNet-B4 is used to extract global semantic features, and image partitions are processed independently by multiple sub-models, and the forged regions are weighted and focused through the GSA module. The output features of the two branches are spliced in the fusion module and the final discrimination is performed by a unified classifier. For the forgery probability of the final output, the study sets a threshold of 0.5. If the probability is greater than 0.5, it is judged as a forged image. Otherwise, it is judged as a real image. The pseudo-code of the LLF-MPPSA-GAN algorithm is shown in Figure. 7.

```
# Input: facial image I
# Output: final forgery probability P_final

# Step 1: Low-level feature extraction via LLF-GAN
I_wavelet = WaveletTransform(I)
[R, G, B] = SplitChannels(I_wavelet)
color_feat = ExtractColorFeatures(R, G, B)
noise_feat = ExtractNoiseFeatures(R, G, B)
low_feat = Concatenate(color_feat, noise_feat)
LLF_feat = ML_CNN(low_feat)
LLF_enhanced = HighPassFilter(LLF_feat)
LLF_output = VGG19Classifier(LLF_enhanced)

# Step 2: Multi-region attention-based inference via MPPSA-GAN
F_init = EfficientNetB4(I)
region_preds = []
region_weights = []

for region r in Regions(F_init):
    # Multi-submodel prediction (Eq. 7)
    P_r = Sigmoid(Sum(w_i * f_i(r) for i in submodels))

    # Spatial attention weighting (Eq. 8, 9)
    A_r = ComputeGSA(r)          # attention map
    S_r = RegionScore(A_r, r)    # weighted region score

    region_preds.append(P_r)
    region_weights.append(S_r)

# Step 3: Region-level prediction fusion (Eq. 10)
P_MPPSA = WeightedAverage(region_preds, region_weights)

# Step 4: Final feature fusion and classification
P_final = FusionClassifier(LLF_output, P_MPPSA)

return P_final
```

Figure 7: Pseudo-code for the LLF-MPPSA-GAN algorithm.

3 Results

3.1 Performance testing of a new facial forgery image forensics model

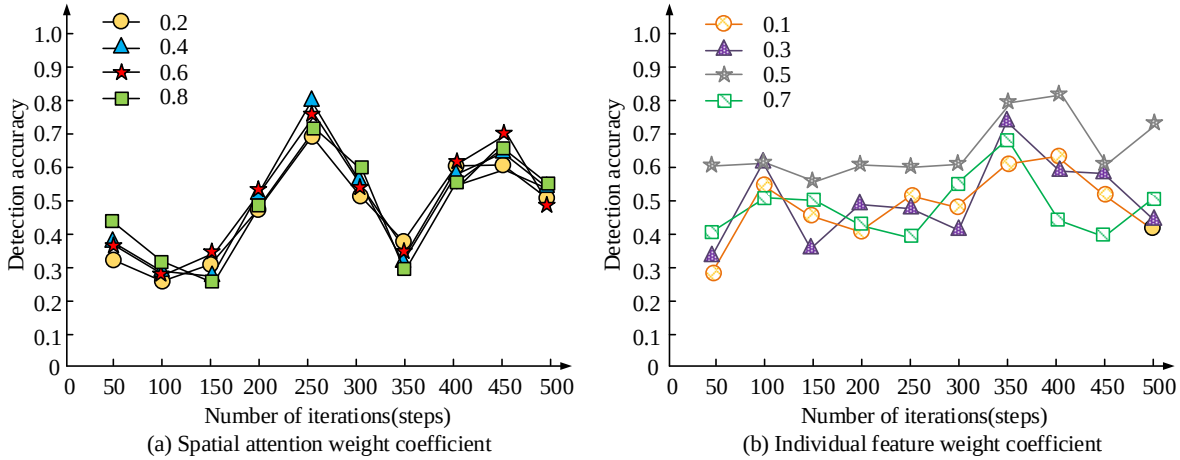
The research sets the CPU to Intel Core i7 3.6GHz, GPU to Nvidia GeForce GTX 1080 Ti, memory to 32GB, and uses Python 3.7 and TensorFlow 2.4 frameworks for model training and testing. The pre-training weights used in the modules are all obtained based on training on publicly available datasets and are fine-tuned in this study to fit the forgery image detection task. The pre-trained model of VGG19 is trained on ImageNet with about 143.7M parameters. EfficientNet-B4 is trained on ImageNet with about 19M parameters. In LLF-GAN, the classifier adopts the classical VGG19 network structure, which contains 16 convolutional layers and 3 fully connected layers. In MPPSA-GAN, the EfficientNet-B4 network, which consists of composite scaled convolutional modules with strong expressive power, is used as the feature extractor. Both are loaded with weights pre-trained on ImageNet and fine-tuned for this research task. For both LLF-GAN and MPPSA-GAN, cross-entropy loss is used as the main training objective function. In the overall integration model, the output losses of the LLF path and MPPSA path are each given the same weight, i.e., $\lambda_1 = \lambda_2 = 0.5$, and the final loss is the weighted sum of the two.

The experiments are evaluated based on two mainstream facial forgery datasets: Celeb-DeepFake Dataset (Celeb-DF) (a total of 5,639 videos with about 590,000 images extracted) and DeepFake Detection Challenge Dataset (DFDC) (19,000 images selected from it). The data is divided into 70% training set, 15% validation set, and 15% test set. The training process uses random level flipping and luminance adjustment for data enhancement, the total number of training rounds is 80, and the optimizer uses Adam (with an initial learning rate of $1e^{-4}$). Five-fold cross-validation is adopted. The experimental results are shown in Table 2. The mean ± standard deviation (std) of each metric is used to evaluate the generalization ability of the method. Both are loaded pre-trained weights on ImageNet and fine tuned them for this research task.

Table 2: Performance metrics results under five-fold cross-validation

Fold	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
Fold 1	93.82	91.02	92.41	92.63
Fold 2	94.36	92.85	93.22	93.17
Fold 3	94.12	90.98	92.51	92.69
Fold 4	93.74	92.41	93.06	93.08
Fold 5	94.01	91.66	92.71	92.78
Mean ± standard deviation	94.01 ± 0.22	91.78 ± 0.67	92.78 ± 0.31	92.84 ± 0.25

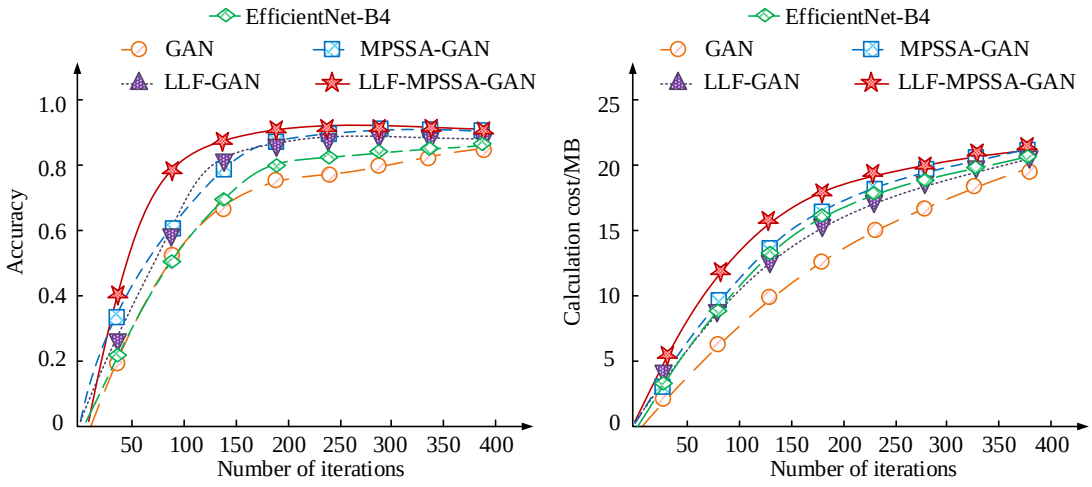
From Table 2, the proposed model showed good stability and robustness in five-fold cross-validation on the DFDC dataset. The fluctuations of each index in different folds were small, with the average precision reaching 94.01%, recall 91.78%, F1 score 92.78%, and accuracy 92.84%. The standard deviations were all controlled within 1%, showing that the model had consistent and excellent detection performance under different data divisions. This further validated the generalization ability of the proposed model, indicating its good adaptability and reliability in real complex environments. The study first conducts value validation on the two types of hyperparameters that have the greatest impact on model performance, namely the spatial attention weight coefficient α_j and the weight coefficient of individual features β_k. The test results are shown in Figure 8.



(a) Spatial attention weight coefficient (b) Individual feature weight coefficient

Figure 8: Hyperparameter selection test result.

Figure 8 (a) displays the spatial attention weight coefficient selection test. Figure 8 (b) displays the weight coefficient selection test for a single feature. From Figure 8 (a), as the spatial attention weight coefficient increased from 0.2 to 0.8, the detection accuracy fluctuated. The coefficients of 0.6 and 0.8 could achieve an accuracy of 0.7 at 250 iterations, while the highest accuracy was 0.8 at 0.4. In Figure 8 (b), when the weight coefficient of a single feature was 0.7, the accuracy reached 0.6 after 300 iterations, while it was only 0.55 when it was 0.1. The accuracy at 0.3 and 0.5 was 0.75 and 0.8, respectively. Higher or lower spatial attention weight coefficients and individual feature weight coefficients can lead to poor detection accuracy. When the spatial attention weight coefficient was 0.4 and the weight coefficient of a single feature was set to 0.5, the detection accuracy is significantly improved. In addition, the research conducts ablation tests on the combined model, as displayed in Figure 9.



(a) Ablation aCcuracy testing (b) Ablation calculation cost test

Figure 9: Ablation test results.

Figure 9 (a) displays the ablation test results under the DFDC. Figure 9 (b) displays the ablation test results under the Celeb-DF dataset. In Figure 9, the LLF-MPPSA-GAN model had the fastest convergence speed

in terms of accuracy improvement, reaching about 0.9 in about 150 iterations, and continued to lead the other models in the subsequent stages. In contrast, even after 400 iterations of the standard GAN model, its accuracy still did not exceed 0.85 and its convergence was significantly lagging. Meanwhile, in terms of computational cost changes, although the resource consumption of LLF-MPPSA-GAN was slightly higher than that of a single model, it consistently maintained a controllable growth during the iteration process and had a higher cost-effectiveness in terms of accuracy improvement, reflecting a better efficiency-performance balance. Combining the results of the two figures, it is verified that the proposed fusion model has strong convergence stability and resource utilization advantages while improving detection performance. The research introduces advanced forensic algorithms for comparison, such as Xception-DeepLab Network (Xception-DeepLab), EfficientNet, and Densely Connected Convolutional Network (DenseNet), Swin Transformers, Two-stream Convolutional and Long Short-Term Memory Networks (Two-stream CNN+LSTM). Precision, Recall, F1 score, and average forensic time are taken as indexes. Table 3 displays the results.

Table 3: Performance comparison of different facial forgery forensics algorithms on benchmark datasets.

Dataset	**Model**	**Precision / %**	**Recall / %**	**F1 score / %**	**Average time spent on depositions / s**	*p*
DFDC	Xception-DeepLab	87.98 ± 0.3	83.86 ± 0.4	81.62 ± 0.5	0.36	<0.01
	EfficientNet	86.33 ± 0.3	82.91 ± 0.3	80.88 ± 0.4	0.51	<0.01
	DenseNet	81.82 ± 0.4	77.25 ± 0.5	78.34 ± 0.5	0.65	<0.01
	Swin Transformer	88.94 ± 0.3	85.13 ± 0.4	84.71 ± 0.4	0.47	<0.01
	Two-stream CNN+LSTM	90.03 ± 0.2	86.57 ± 0.3	85.34 ± 0.3	0.52	<0.01
	Our model	94.36 ± 0.2	91.68 ± 0.2	93.02 ± 0.2	0.28	/
Celeb-DF	Xception-DeepLab	82.17 ± 0.4	81.07 ± 0.5	88.24 ± 0.5	0.35	<0.01
	EfficientNet	82.93 ± 0.4	83.39 ± 0.3	81.88 ± 0.4	0.48	<0.01
	DenseNet	83.98 ± 0.3	83.72 ± 0.4	85.18 ± 0.4	0.59	<0.01
	Swin Transformer	85.11 ± 0.2	84.02 ± 0.3	86.64 ± 0.3	0.43	<0.01
	Two-stream CNN+LSTM	86.85 ± 0.3	85.91 ± 0.3	88.42 ± 0.3	0.46	<0.01
	Our model	95.21 ± 0.2	93.02 ± 0.2	94.21 ± 0.2	0.29	/

According to Table 3, both Swin Transformer and Two-stream CNN+LSTM showed superior detection performance among the selected comparison methods on both DFDC and Celeb-DF datasets. Specifically, the F1 score of Two-stream CNN+LSTM on the DFDC dataset reached 85.34%, which was slightly higher than that of Swin Transformer (84.71%), and both of them were significantly better than traditional methods such as DenseNet and Xception-DeepLab. Meanwhile, the proposed LLF-MPPSA-GAN still had the most outstanding performance on the two datasets, with F1 scores of 93.02% and 94.21% on DFDC and Celeb-DF, respectively, which were significantly higher than all the comparison models. In addition, the computational efficiency of the model also had an obvious advantage. In terms of average forensic time, the proposed model achieved the fastest inference speed with 0.28s and 0.29s, which was better than Swin Transformer (0.47s / 0.43s) and Two-stream CNN+LSTM (0.52s / 0.46s), indicating that the proposed method maintained the high accuracy and has strong real-time performance and deployment potential. All results pass the two-tailed t-test with *p*-values less than 0.01, indicating that the performance improvement was statistically significant.

3.2 Simulation testing of a new facial forgery image forensics model

To evaluate the performance of a new facial forgery image forensics model, six types of facial forgery images are randomly obtained from the DFDC and pre-processed

to ensure the validity of the image data. The research compares forensic detection of six types of images with different qualities, taking Area under Curve (AUC) as the indicator, as displayed in Figure 10.

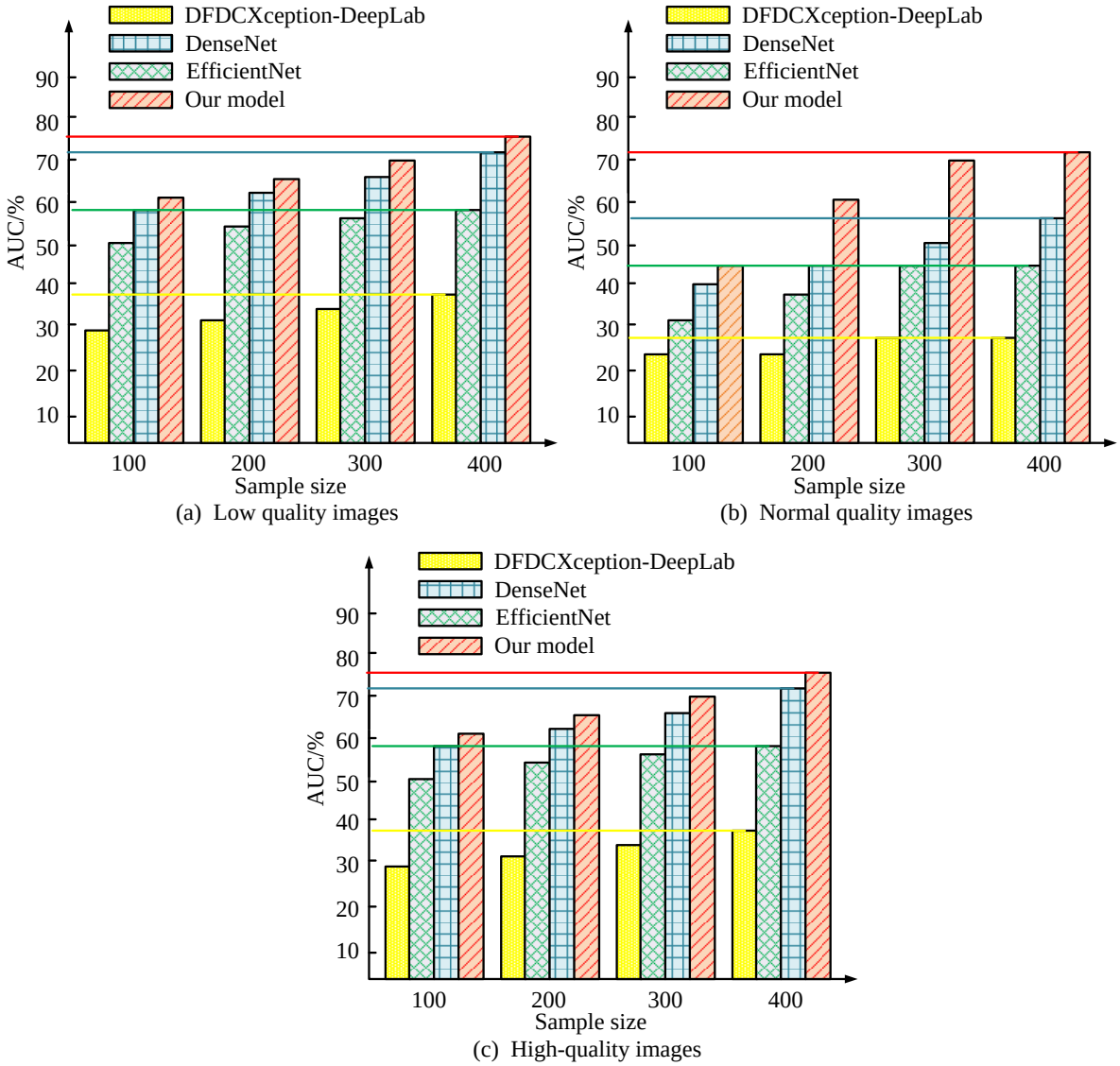


Figure 10: Forensic detection AUC results of forged images of different qualities.

Figure 10 displays the AUC results of forensic detection for low-quality forged images, normal quality forged images, and high-quality forged images. In Figure 10, the proposed model performed the best in low-quality forged images, with an AUC value of 68.34%, which was superior to Xception-DeepLab, EfficientNet, and DenseNet, with improvements of 28.23%, 23.77%, and 10.95%, respectively. In normal quality forged images, the AUC value of the proposed model was 75.56%, once again surpassing other models, especially when dealing with small samples, with an improvement of 28.32%. In high-quality forged images, the AUC of the proposed model reached the highest, at 85.74%, proving the high accuracy in dealing with high-quality forged images. Compared with Xception-DeepLab, EfficientNet, and DenseNet, the AUC values increased by 15.81%, 11.76%, and 9.87%. The proposed model has obvious advantages in various quality forged images, especially in detecting high-quality images, where the improvement in AUC value reflects its strong adaptability and robustness. The study conducts confusion tests on four types of forgery: emotion exchange, identity exchange, attribute editing, and global facial generation. The results are shown in Figure 11.

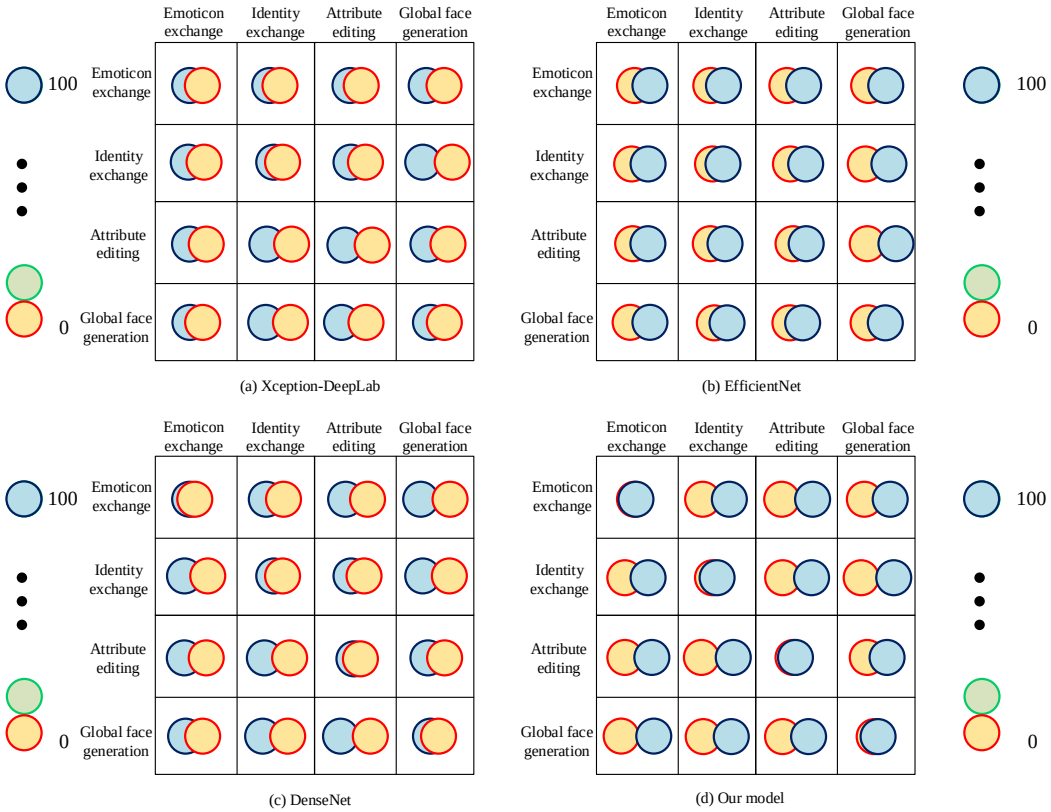


Figure 11: Confusion test results.

Figure 11(a) shows the Xception-DeepLab test results for the four types of forgery image confusion. Figure 11(b) shows the EfficientNet test results for the four types of forgery image confusion. Figure. 11(c) shows the DenseNet test results for the four types of forgery image confusion. Figure 11(d) shows the research method test results for the four types of forgery image confusion. From Figure. 11, the ability of the proposed model to distinguish four types of forgery types (expression exchange, identity replacement, attribute editing, and full-facial generation) was significantly better than the other models. In contrast, Xception-DeepLab, EfficientNet, and DenseNet had significant confusion between identity substitution and attribute editing with high error rates, especially in the full-facial generation task where the confounding judgment was particularly prominent. In addition, the proposed model maintained high accuracy on cross-recognition in all categories, especially showing clearer boundaries between expression swapping and attribute editing, which significantly reduced type confusion. This indicates that the proposed fusion model has stronger fine-grained recognition ability and structural discrimination, and can effectively deal with complex and diverse counterfeiting techniques. Taking the Receiver Operating Characteristic curve (ROC) as an indicator, the results are shown in Figure 12.

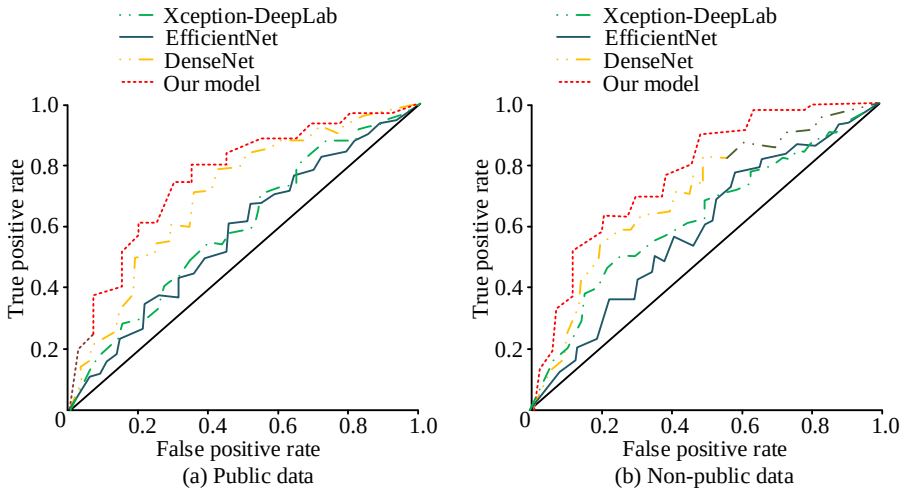


Figure 12: Statistical results of AUC indicators.

Figure 12(a) shows the ROC curves of different models in the public data, and Figure 12(b) shows the ROC curves of different models in the unpublished data. The horizontal axis represents the false-positive rate and the vertical axis represents the true-positive rate. The larger the AUC, which is enclosed by the ROC curve and the horizontal and vertical coordinates, the better the model performance. From Figure 12, the proposed model achieved optimal performance in both data conditions, with AUC values of 0.91 and 0.87, which were significantly higher than Xception-DeepLab (0.82 / 0.77), EfficientNet (0.84 / 0.79) and DenseNet (0.85 / 0.80). Especially in the non-public data test, the proposed model still maintained a large advantage, indicating its stronger robustness and generalization ability. Overall, the fusion structure not only improves the recognition accuracy on public forged images, but also significantly enhances its adaptability when facing unknown forged samples. The forged facial images with low noise, normal noise and high noise are detected using forensic Mean Average Precision (mAP), model complexity, and forensic delay as metrics. The results are shown in Table 4. The detection performance on low-noise, moderate-noise, and high-noise forged facial images is evaluated using mAP, model complexity, and processing latency, as presented in Table 4. All noises were added using a Gaussian distribution simulation, with different standard deviations set to correspond to low ($\sigma = 5$), medium ($\sigma = 15$), and high ($\sigma = 30$) noise intensities, respectively, and superimposed on the RGB channel of the image to generate interference samples.

Table 4: Robustness evaluation of forensic algorithms under varying noise conditions.

Type of noise	Model	mAP/%	Model complexity/%	Delayed depositions/%	FLOPs(G)	Memory (MB)	*p*
Low noise	Xception-DeepLab	89.34 ± 0.4	88.67	0.34	12.4	698	<0.01
	EfficientNet	90.12 ± 0.3	89.34	0.37	10.8	645	<0.01
	DenseNet	91.44 ± 0.3	90.22	0.32	13.5	732	<0.01
	Our model	92.98 ± 0.2	80.54	0.28	9.3	528	/
Normal noise	Xception-DeepLab	88.23 ± 0.4	87.46	0.35	12.4	698	<0.01
	EfficientNet	89.56 ± 0.3	88.97	0.31	10.8	645	<0.01
	DenseNet	90.87 ± 0.3	89.12	0.33	13.5	732	<0.01
	Our model	93.12 ± 0.2	83.88	0.29	9.3	528	/
High noise	Xception-DeepLab	84.56 ± 0.5	83.12	0.36	12.4	698	<0.01
	EfficientNet	85.12 ± 0.4	84.78	0.32	10.8	645	<0.01
	DenseNet	86.34 ± 0.4	85.54	0.34	13.5	732	<0.01
	Our model	89.23 ± 0.3	81.67	0.35	9.3	528	/

According to Table 4, under different types of noise interference, the proposed LLF-MPPSA-GAN model showed strong stability and advantages in terms of mAP value, model complexity, and delayed forensic performance. Taking high noise environment as an example, the proposed model still achieved 89.23% mAP, which was better than Xception-DeepLab (84.56%), EfficientNet (85.12%), and DenseNet (86.34%), and the complexity of the model stayed at 81.67%, which was much lower than the average of other models at about 85%-90%, verifying its lightweight and low-cost performance. This indicates the effectiveness of the lightweight design strategy. In terms of delayed forensics, the proposed model achieved the shortest forensics time under all types of noise conditions, with a minimum of only 0.28s, which further highlighted its real-time response capability. In addition, the research method is statistically examined on the mAP results of all the comparison models under three noise levels. The *p*-values obtained from two-tailed independent samples t-tests were less than 0.05, which indicated that the advantages of the research model on noise robustness are statistically significant.

4 Discussion

Aiming at the current facial forgery detection problems of insufficient multi-region sensing ability, low feature detail extraction efficiency, and poor robustness in complex environments, the study proposes a two-branch improved GAN forensic model, LLF-MPPSA-GAN, which integrates low-level feature extraction and multi-prediction partition spatial attention mechanism. The experimental results showed that on two mainstream datasets, DFDC and Celeb-DF, both achieved F1 scores of over 93% and mAP of over 89%, significantly outperforming DenseNet and Swin Transformer. Both achieved F1 scores over 93%, mAP stayed above 89% in multiple noisy environments, and the average inference time was as low as 0.28 seconds, which was significantly better than methods such as DenseNet, EfficientNet, and Swin Transformer. In a single image path, LLF-GAN, which fuses color and noise, enhances fine-grained feature perception and effectively locates low-frequency residual forgery traces. MPPSA-GAN combines the global semantic understanding of EfficientNet with the local weighting mechanism of GSA, enhancing the accuracy of multi region forgery recognition and improving the ability to capture edge contours and microstructural changes. Compared with the single-branch forgery recognition framework using attention convolution proposed by Lin K et al., the two-way parallel mechanism proposed in this paper significantly mitigates the ambiguous model recognition and weak local response when oriented to multi-class forgery scenarios [20]. Meanwhile, the feature scoring mechanism based on multi-stage fusion improves the accuracy of determining the forgery in different regions, further verifying the adaptability of the weight allocation strategy on complex samples. Despite the multi-module combination, the overall complexity of the model is still controlled at about 81%, and the inference latency is no more than 0.35 seconds, which possesses strong deployment efficiency and edge device adaptability. Especially under non-public data and high noise conditions, the AUC remains above 0.87, indicating its good generalization ability. Subsequently, the model can be further compressed and distillation or quantization strategies can be introduced to adapt to the real-time forensic task of embedded platforms, In addition, it can also enhance the recognition stability of dynamic video forgery and occlusion interference scenes.

5 Conclusion

A two-branch facial forgery forensic model LLF-MPPSA-GAN fusing low-level feature extraction and multi-prediction partitioned spatial attention mechanism was proposed to construct discriminative paths for single facial images and multi-region complex forgery scenarios, respectively and determine image authenticity through branch fusion. Experimental results showed that the method achieved better detection performance than existing methods on multiple benchmark datasets, and exhibited good stability and robustness in terms of noise interference, forgery type differentiation, and computational efficiency. The model can be widely used in the fields of social platform content censorship, judicial image appraisal, identity verification security, etc. It provides a feasible technical basis for the practical deployment of forgery forensics system while improving the practicality of deep forgery detection technology.

6 Future work and limitations

Although the proposed LLF-MPPSA-GAN model performs well on multiple datasets and test tasks, there are still several limitations that need to be emphasized. First, the current method performs forgery identification based on single-frame images, which is difficult to effectively capture cross-frame information changes, limiting the ability to detect dynamic video forgery. Second, the experiments are mainly based on two public datasets, Celeb-DF and DFDC, which cover common types of forgeries, but there are some limitations in the sample distribution and forging means. This may lead to model overfitting in real complex environments, and the generalization ability still needs to be further verified. In addition, for anomalous forgery images generated by other novel generative models (e.g., StyleGAN3 or Diffusion model), the detection robustness of the model has not been fully evaluated, and there may be misjudgments. Future work can consider introducing a multi-scale attention mechanism based on the Transformer structure to enhance cross-region feature interaction capability. A video level forensic framework can be constructed by combining multi-modal information such as audio synchronization, speech consistency, etc., to improve the perception depth of deepfake scenes. Zero sample or small sample forgery recognition methods can be explored to enhance the adaptability of the model to unknown forgery samples, thereby expanding its deployment value and practicality in real security scenarios.

References

[1] M. Tampubolon, (2024) "Digital face forgery and the role of digital forensics," International Journal for the Semiotics of Law-Revue Internationale De SÉMiotique Juridique, vol. 37, no. 3, pp. 753-767, https://doi.org/10.1007/s11196-023-10030-1

[2] M. T. Pham, T. T. Huynh, T. T. Nguyen, J. Jo, H. Yin, and Q. V. H. Nguyen, (2024) "A dual benchmarking study of facial forgery and facial forensics," CAAI Transactions on Intelligence Technology, vol. 9, no. 6, pp. 1377-1397, https://doi.org/10.1049/cit2.12362

[3] X. Zhu, H. Fei, B. Zhang, T. Zhang, X. Zhang, and S. Li, (2023) "Face forgery detection by 3d decomposition and composition search," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 7, pp. 8342-8357, https://doi.org/10.1109/TPAMI.2022.3233586

[4] F. Ding, B. Fan, Z. Shen, K. Yu, G. Srivastava, and K. Dev, (2022) "Securing facial bioinformation by

eliminating adversarial perturbations," IEEE Transactions on Industrial Informatics, vol. 19, no. 5, pp. 6682-6691, https://doi.org/10.1109/TII.2022.3201572

[5] G. Lan, S. Xiao, J. Wen, D. Chen, and Y. Zhu, (2022) "Data-driven deepfake forensics model based on large-scale frequency and noise features," IEEE Intelligent Systems, vol. 39, no. 1, pp. 29-35, https://doi.org/10.1109/MIS.2022.3217391

[6] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, (2023) "Making DeepFakes more spurious: Evading deep face forgery detection via trace removal attack," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 6, pp. 5182-5196, https://doi.org/10.1109/TDSC.2023.3241604

[7] W. El-Shafai, M. A. Fouda, E. S. M. El-Rabaie, and N. A. EI-Salam, (2024) "A comprehensive taxonomy on multimedia video forgery detection techniques: Challenges and novel trends," Multimedia Tools and Applications, vol. 83, no. 2, pp. 4241-4307, https://doi.org/10.1007/s11042-023-15609-1

[8] Z. Lai, Z. Yao, G. Lai, C. Wang, and R. Feng, (2024) "A novel face swapping detection scheme using the pseudo zernike transform based robust watermarking," Electronics, vol. 13, no. 24, pp. 4955-4963, https://doi.org/10.3390/electronics13244955

[9] P. Sharma, M. Kumar, and H. Sharma, (2023) "Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: An evaluation," Multimedia Tools and Applications, vol. 82, no. 12, pp. 18117-18150, https://doi.org/10.1007/s11042-022-13808-w

[10] Q. Li, R. Wang, and D. Xu, (2023) "A video splicing forgery detection and localization algorithm based on sensor pattern noise," Electronics, vol. 12, no. 6, pp. 1362-1367, https://doi.org/10.3390/electronics12061362

[11] L. Guarnera, O. Giudice, F. Guarnera, A. Ortis, G. Puglisi, A. Paratore, L. M. Q. Bui, M. Fontani, D. A. Coccomini, R. Caldelli, F. Falchi, C. Gennaro, N. Messina, G. Amato, G. Perelli, S. Concas, C. Cuccu, G. Orrù, G. L. Marcialis, and S. Battiato, (2022) "The face deepfake detection challenge," Journal of Imaging, vol. 8, no. 10, pp. 263-267, https://doi.org/10.3390/jimaging8100263

[12] H. Zhang, (2022) "A survey of anti-forensic for face image forgery," Journal of Information Hiding and Privacy Protection, vol. 4, no. 1, pp. 41-44, https://doi.org/10.32604/jihpp.2022.031707

[13] Y. Li, L. Ye, H. Cao, W. Wang, and Z. Hua, (2024) "Cascaded adaptive graph representation learning for image copy-move forgery detection," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 21, no. 2, pp. 1-24, https://doi.org/10.1145/3669905

[14] S. Tyagi, and D. Yadav, (2023) "A detailed analysis of image and video forgery detection techniques," The Visual Computer, vol. 39, no. 3, pp. 813-833, https://doi.org/10.1007/s00371-021-02347-4

[15] H. Zhang, B. Chen, J. Wang, and G. Zhao, (2022) "A local perturbation generation method for GAN-generated face anti-forensics," IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 2, pp. 661-676, https://doi.org/10.1109/TCSVT.2022.3207310

[16] P. Sharma, M. Kumar, and H. Sharma, (2023) "Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: An evaluation," Multimedia Tools and Applications, vol. 82, no. 12, pp. 18117-18150, https://doi.org/10.1007/s11042-022-13808-w

[17] S. Cherian, J. Joseph, and B. Thomas, (2024) "Navigating the new normal: A bibliometric analysis of masked face recognition research using VOSviewer and Biblioshiny," Informatica (Slovenia), vol. 48, no. 22, pp. 193–212, December, DOI: 10.31449/inf. v48i22.6342.

[18] Y. Li, T. Xie, and D. Mei, (2025) "Using DTL-MD with GANs and ResNet for malicious code detection," Informatica (Slovenia), vol. 49, no. 14, pp. 63–78, March, DOI: 10.31449/inf. v49i14.7937.

[19] Sandhya, and A. Kashyap, (2024) "A comprehensive analysis of digital video forensics techniques and challenges," Iran Journal of Computer Science, vol. 7, no. 2, pp. 359-380, https://doi.org/10.1007/s42044-023-00165-6

[20] K. Lin, W. Han, S. Li, Z. Gu, H. Zhao, J. Ren, L. Zhu, and J. Lv, (2023) "IR-capsule: Two-stream network for face forgery detection," Cognitive Computation, vol. 15, no. 1, pp. 13-22, https://doi.org/10.1007/s12559-022-10008-4