
Deepfake Image Detection Using Squeeze-and-Excitation Attention and Adaptive Threshold Optimization

Chaoran Li

Hainan Vocational University of Science and Technology, Haikou 571126, China
E-mail: ChaoranLi2026@outlook.com

Received 15 April 2026; Accepted 29 May 2026

Abstract

Artificial intelligence-generated content (AIGC) has significantly improved the realism of manipulated facial images and videos, creating serious risks for social-network misinformation, content security, and digital trust. This study focuses on visual deepfake image detection rather than multimodal misinformation detection. Existing convolutional neural network (CNN)-based deepfake detectors commonly rely on a fixed decision threshold of 0.5, which may be unstable when facial images are affected by compression, cropping, noise, and distribution shifts in practical social-media environments. To improve detection reliability, this paper proposes a lightweight deepfake image detection framework that integrates a CNN backbone, a Squeeze-and-Excitation (SE) attention module, and an adaptive threshold optimization strategy. The key novelty of this work is not conventional threshold tuning alone, but a lightweight detection framework that jointly combines SE-based channel-wise feature recalibration, F1-score-driven adaptive threshold optimization, and leakage-aware group-level evaluation. Compared with post-hoc threshold tuning or calibration methods that mainly adjust the output boundary after training, the proposed framework improves both

Journal of Cyber Security and Mobility, Vol. 15.5, 1233–1252.

doi: 10.13052/jcsm2245-1439.1554

© 2026 River Publishers

forgery-sensitive feature representation and the final classification decision mechanism. Specifically, the SE module enhances forgery-related feature representation through channel-wise feature recalibration, while the adaptive threshold is selected according to the F1-score, which is the harmonic mean of precision and recall. A group-level data split is also adopted to ensure that frames from the same video are not shared across the training, validation, and test sets, thereby reducing identity leakage. Experiments are conducted on a FaceForensics++-derived dataset using multiple CNN backbones and five-fold cross-validation. The evaluation metrics include accuracy, F1-score, the area under the receiver operating characteristic curve (ROC-AUC), and the area under the precision-recall curve (PR-AUC). The results show that threshold optimization improves the ResNet18 baseline accuracy from 0.8835 ± 0.0186 to 0.9016 ± 0.0151 and the F1-score from 0.8773 ± 0.0253 to 0.9042 ± 0.0138 . Among the tested backbones, ResNet50 achieves the best cross-validation performance, with an accuracy of 0.9061 ± 0.0161 and a ROC-AUC of 0.9680 ± 0.0075 . Under the strict group-level test setting, the proposed framework achieves 0.9960 accuracy and 0.9959 F1-score on the constructed test set. Compared with more complex detection pipelines. The proposed method is simple, computationally practical, and easy to integrate into common CNN-based deepfake detectors. Therefore, this work provides a practical visual security detection approach for social-media content verification and cyber-security applications.

Keywords: AIGC detection, deepfake, cross-modal semantic consistency, social network security, content security..

1 Introduction

Artificial intelligence-generated content (AIGC) has rapidly transformed the creation, manipulation, and dissemination of digital visual media. In particular, deepfake techniques can synthesize or modify facial images and videos with a level of realism that makes forged content increasingly difficult to distinguish from authentic media. Large-scale benchmark datasets, such as FaceForensics++ and Celeb-DF, have promoted systematic research on face-forgery detection and provided important experimental foundations for evaluating detection performance and generalization ability [1, 2]. At the model level, convolutional neural networks (CNNs), attention mechanisms, and compact detection architectures have become widely used for learning discriminative facial-forgery representations [3–6]. At the same time,

advances in face reenactment, neural rendering, and physiological-signal-based detection show that both deepfake generation and deepfake detection are developing rapidly [7–9]. Therefore, deepfake detection is no longer only a computer-vision task, but also an important problem in multimedia forensics, social-network security, and digital-content trust.

Existing studies have made substantial progress in detecting manipulated facial content. Pixel-region relation modeling has been explored to capture local forgery inconsistencies [10], while comprehensive surveys have summarized the evolution of face manipulation and fake detection techniques [11]. Recent methods have further investigated multi-scale dual-stream learning, media authenticity analysis, confidence-aware learning, spatial-residual two-stream networks, biological-signal analysis, heart-rate-based detection, and qualitative failures of image generation models [12–14]. These studies indicate that forgery traces may exist in spatial textures, semantic representations, frequency patterns, physiological signals, and generation artifacts. However, in real social-network environments, manipulated images are often compressed, resized, cropped, filtered, or contaminated by platform-dependent noise before they reach detection systems. These operations may weaken forgery artifacts and reduce the reliability of a detector that depends only on a fixed decision rule. Although advanced methods such as Face X-Ray and DeeperForensics have improved the robustness and realism of deepfake detection research [15], several practical limitations remain. First, many CNN-based detectors still convert predicted probabilities into final binary labels using a fixed threshold of 0.5. This default setting is simple, but it may be suboptimal when data distributions shift because of compression, cropping, noise, or other social-media transformations. Second, forgery-related traces are often subtle and unevenly distributed across facial regions and feature channels. If all feature responses are treated equally, the detector may fail to emphasize the most informative forgery patterns. Third, experimental evaluation must avoid identity-level or video-level leakage. If frames from the same source video appear in both the training and test sets, the reported results may be overly optimistic and may not reflect real generalization performance. These problems show that improving the backbone network alone is insufficient; feature enhancement, adaptive decision-making, and leakage-aware evaluation should be considered together [16].

To address these issues, this paper focuses on visual deepfake image detection rather than misinformation detection. A lightweight CNN-based detection framework is proposed by integrating a Squeeze-and-Excitation (SE) attention module with an adaptive threshold optimization strategy.

The attention module is used to recalibrate channel-wise feature responses and strengthen forgery-related visual representations. The adaptive threshold strategy selects the final classification boundary according to validation performance instead of relying on the conventional fixed threshold of 0.5. In addition, a group-level data-splitting protocol is adopted to ensure that frames from the same video are not shared across the training, validation, and test sets. This design reduces the risk of identity or video-level leakage and provides a stricter evaluation setting for practical deepfake image detection [17–19].

The key novelty of this study lies in improving the decision mechanism of CNN-based deepfake image detection through the joint use of attention-based feature recalibration and validation-driven adaptive threshold optimization. Unlike methods that mainly rely on deeper backbones, heavier temporal modeling, or complex fusion (Figure 1), the proposed framework aims to improve practical detection reliability with a simple and deployable structure. The method enhances channel-level feature selectivity during representation learning and adjusts the final classification boundary according to validation-set performance. Therefore, it is suitable for social-network

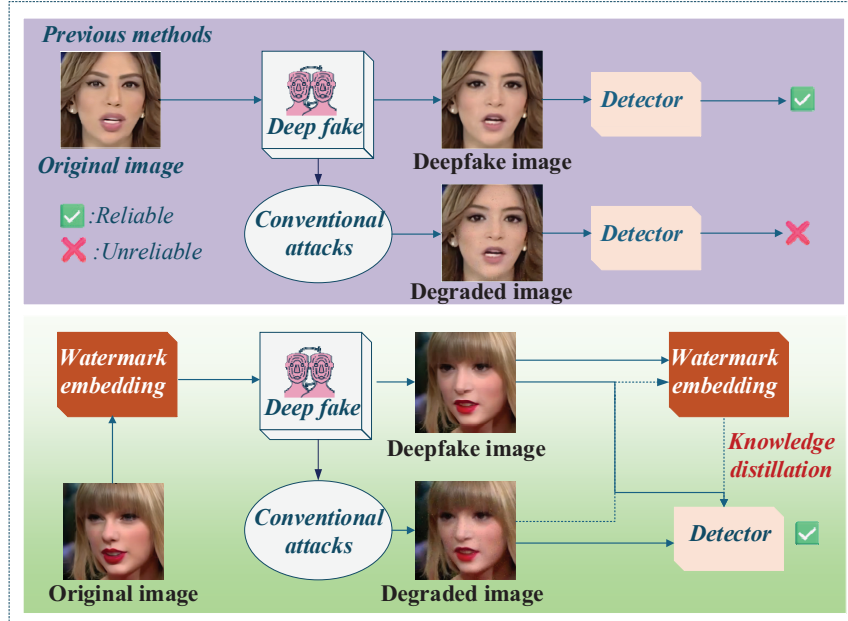


Figure 1 Overview of the proposed deepfake detection system.

visual-content verification scenarios, where detection systems must balance accuracy, robustness, computational efficiency, and ease of integration. Compared with existing deepfake detection methods, the proposed framework has several advantages. First, practical because adaptive threshold optimization can be incorporated into common CNN-based detectors without redesigning the entire architecture. Second, robust because threshold adjustment helps reduce instability caused by distribution shifts and class-wise decision imbalance. Third, lightweight because the framework avoids complex fusion and heavy temporal modeling. Fourth, evaluation-aware because the group-level split reduces the risk of identity and video-level leakage between training, validation, and testing. These advantages make the proposed method useful for visual misinformation detection, social-media content verification, and cyber-security-oriented image authentication.

The main contributions of this paper are as follows.

- (1) A lightweight CNN-based framework is proposed for visual deepfake image detection in social-network content-security scenarios.
- (2) An attention-based feature recalibration module is introduced to enhance channel-wise forgery-related visual representation.
- (3) An adaptive threshold optimization strategy based on validation-set performance is developed to improve the classification decision boundary beyond the conventional fixed threshold of 0.5.
- (4) A group-level data-splitting protocol is adopted to reduce identity and video-level leakage between the training, validation, and test sets.
- (5) Experiments on a FaceForensics++-derived dataset are conducted using multiple CNN backbones, five-fold cross-validation, ablation analysis, and a strict group-level test setting to evaluate the effectiveness, robustness, and practical value of the proposed framework.

2 Related Work

2.1 Deepfake Detection Methods

Deepfake detection methods can generally be divided into frame-level image analysis and video-level temporal analysis. Frame-level methods focus on visual artifacts within individual images, such as abnormal facial textures, inconsistent illumination, missing facial details, and unnatural geometric structures. Early studies relied heavily on handcrafted features, but their effectiveness decreases as deepfake generation models become more realistic [20–22]. With the development of deep learning, CNNs have become widely

used for deepfake detection. Representative methods such as MesoNet and Xception-based detectors learn discriminative visual features from manipulated facial images and have achieved promising results on public datasets. Other studies have further explored frequency-domain clues, attention mechanisms, and multi-scale feature extraction to improve the detection of subtle forgery traces. Video-level methods exploit temporal inconsistency across consecutive frames. These methods can capture motion artifacts, unstable facial dynamics, and temporal discontinuities that may not be visible in a single frame. However, they usually require higher computational cost and more training data. Therefore, lightweight and stable image-level detection methods remain valuable for practical social-media content verification [22–25].

2.2 Facial Landmark Detection

Not only are there visual features in images, but also facial geometric structures information that helps identify deepfakes. Facial landmarks are specific points on the face used to determine facial expression or body pose changes in images. Most of these early face landmark detection approaches focused on traditional computer-vision models at that time, including Active Appearance Models (AAM) and Constrained Local Models (CLM). Other regression-based approaches have been presented to improve the accuracy of landmark detection and are thus employed frequently in real-world applications. Based on the recent progress in research for deep learning, many landmark-detection algorithms using convolutional-neural-networks have been presented. One representative example is the Convolutional-Pose-Machine (CPM) algorithm. These approaches are also less affected by complicated scenes and pose-occlusion issues compared with the former one [26–28]. For tasks of face forgery detection, the representation ability of landmark data in describing the shape characteristics and expression changes is strong. Although most landmark detection algorithms process each frame separately and hence are not entirely jitter-free when performing recognitions in a time-domain model, improving the stability and accuracy of key point tracking becomes a prominent problem when using these points to detect deepfakes.

2.3 Dataset Construction and Preprocessing

In this study, a FaceForensics++-derived dataset was constructed for deepfake image detection. FaceForensics++ is a widely used benchmark dataset that contains real videos and manipulated videos generated by different face-forgery techniques. To improve data reliability and preprocessing

Table 1 Data set statistics

Item	Value
cropped_images	998
labels_big2.csv	960 (real = 480, fake = 480)
deepfake_dataset	5150
group split (train/val/test)	210/30/60
train/val/test	3619/533/998
fake/real support	510/488

consistency, face regions were extracted from the original videos, and low-quality or unreliable samples were removed during data cleaning [29]. FaceForensics++ is a typical data set for deepfakes; it consists not only of real video materials but also numerous manipulated videos created using various forging techniques. The construction results of the dataset are shown in Table 1.

In the pre-processing stage, face regions in the FaceForensics++ dataset were obtained first. The original data had 998 video folders (cropped_images). Then, based on the contents of labels_big2.csv, we selected high confidence samples from the collected data. Filtering the total of 960 folders resulted in 480 genuine videos and 480 fake ones. After frame extraction and data cleaning, the final dataset (deepfake_dataset) contains 5150 face image samples. To ensure a fair evaluation, a group-level split of the dataset was used, and samples extracted from the same source video were assigned to only one subset; they could not appear in the training, validation, and test sets simultaneously. Specifically, the final image-level division was 3619/533/998 for training, validation, and testing, respectively. There are 510 fake samples and 488 real samples in the test set. The above division strategy will prevent data leakage and offer a more accurate measure of the model’s generalization capability [30]. The entire pipeline of the proposed framework is shown in Figure 2, which includes face preprocessing, CNN-based feature extraction, SE attention, adaptive threshold optimization, and forgery detection.

To prevent identity and video-level leakage, a group-level splitting strategy was adopted in this study. Specifically, all frames extracted from the same source video were treated as one indivisible group, and each group was assigned exclusively to one subset: training, validation, or testing. Therefore, frames from the same video could not appear simultaneously in different subsets. This protocol prevents the model from memorizing identity-specific, background-specific, or video-specific patterns and then reusing them during testing. Compared with random frame-level splitting, the group-level split

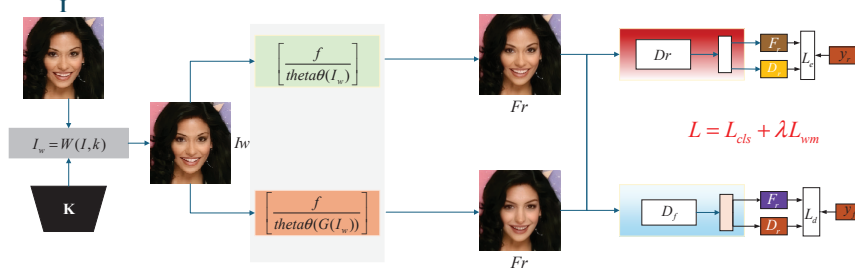


Figure 2 Pipeline of the proposed framework, including face preprocessing, CNN-based feature extraction, SE attention, adaptive threshold optimization, and forgery detection.

provides a stricter and more reliable evaluation of the model's generalization ability [31].

3 Methodology

This section introduces the proposed deepfake image detection framework and details the face preprocessing, CNN-based feature extraction, SE attention-based feature recalibration and adaptive threshold optimization modules. Given an input face image I , first detect and crop the facial region of interest for feature learning

$$F = D(I) \quad (1)$$

where $D(\cdot)$ is the face detection function. To ensure that all samples are uniform in size, cropped face images are then up-sampled to this spatial resolution. Normalize the image to obtain which, i.e., a scaled version of $\tilde{F}$

$$\tilde{F} = R(F) \quad (2)$$

where $R(\cdot)$ represents a resampling operation. The resolution of the input images in this paper is set at 224x224 (224*224). The processed image of each sample in a total data volume that is size N

$$\tilde{F} = \{\tilde{F}_1, \tilde{F}_2, \dots, \tilde{F}_N\} \quad (3)$$

The normalized facial images are then fed into the CNN-based feature extraction network to learn discriminative forgery-related representations for real/fake classification. The CNN-based feature extraction process is illustrated in Figure 3.

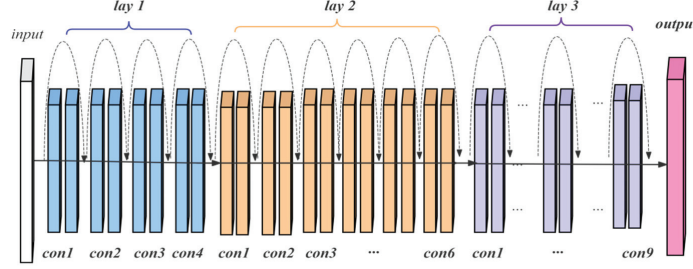


Figure 3 CNN-based feature extraction module of deepfake detection.

Given an input image, the CNN encoder transforms it into a latent feature vector z

$$z = f_{\theta}(\tilde{F}) \quad (4)$$

where $f_{\theta}(\cdot)$ represents the feature extraction network, θ is the network parameter, $z \in R^d$ is the learnt feature representation. Conventional operations in the network may be written as

$$x_{i,j}^{(k)} = \sigma \left(m \sum_n \sum w_{m,n}^{(k)} x_i + m, j + n + b^{(k)} \right) \quad (5)$$

where $w^{(k)}$ is a convolution kernel, $b^{(k)}$ is the bias term, $\sigma(\cdot)$ represents a non-linear activation function. Through stacked convolution and pooling operations, the CNN progressively extracts high-level semantic representations that capture subtle forgery artifacts in facial images [32]. After obtaining the learned feature representation z , the predicted probability can be calculated as

$$p = \sigma(Wz + b) \quad (6)$$

The classifier determines whether the input image is genuine or fake based on the predicted probability p

$$\begin{cases} \sigma(x) = \frac{1}{1 + e^{-x}} \\ L = -\frac{1}{N} \sum_{i=1}^N [y \log(pi) + (1 - yi) \log(1 - pi)] \end{cases} \quad (7)$$

where the predicted probability that the input image is a fake is denoted by p . The classification function can be described as

$$\theta_{t+1} = \theta_t - \eta \nabla \theta_L \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, p represents the predicted probability that the input image belongs to the fake class, y_i is the ground-truth label of the i_{th} sample, N is the number of training samples, L denotes the binary cross-entropy loss used for real/fake classification. In the parameter update equation, θ_t represents the model parameters at iteration t , η is the learning rate, and ∇_{θ_L} denotes the gradient of the loss function with respect to the model parameters.

4 Experiment

Before model training and evaluation, the dataset was divided using the group-level split described in Section 2.3. In this setting, the source video was used as the grouping unit rather than individual frames. All frames from the same source video were assigned to a single subset, so the training, validation, and test sets were mutually exclusive at the video-group level. This setting was used to reduce identity leakage and avoid overly optimistic performance caused by random frame-level splitting.

On the constructed deepfake image dataset, five-fold cross-validation was conducted using several widely used backbones, and Accuracy, F1-score, ROC-AUC, and PR-AUC were used as evaluation metrics. As shown in Table 2, ResNet50 achieved the best overall performance among the tested backbones, with Accuracy = 0.9061 ± 0.0161 , F1-score = 0.9065 ± 0.0185 , ROC-AUC = 0.9680 ± 0.0075 , and PR-AUC = 0.9677 ± 0.0073 . MobileNetV3-Large also performed competitively, whereas Xception and EfficientNet-B0 showed strong discrimination ability, with ROC-AUC values close to 0.96, but slightly lower accuracy and F1-score than ResNet50 and MobileNetV3-Large.

To evaluate the discriminatory ability of the test backbones further, ROC curve analysis was performed in five-fold cross-validation. As shown in Figure 4, all of the tested models achieved high ROC-AUC values and were thus good discriminators for the real and fake deepfake image datasets. Among the tested backbones, ResNet50 had the best ROC-AUC performance and

Table 2 Backbone comparison under five-fold cross-validation

Model	Accuracy	F1-score	ROC-AUC	PR-AUC
Xception	0.8850 ± 0.0061	0.8825 ± 0.0079	0.9638 ± 0.0096	0.9647 ± 0.0092
EfficientNet-B0	0.8805 ± 0.0131	0.8745 ± 0.0194	0.9593 ± 0.0055	0.9565 ± 0.0077
MobileNetV3-Large	0.9005 ± 0.0207	0.9022 ± 0.0187	0.9639 ± 0.0107	0.9639 ± 0.0111
ResNet50	0.9061 ± 0.0161	0.9065 ± 0.0185	0.9680 ± 0.0075	0.9677 ± 0.0073

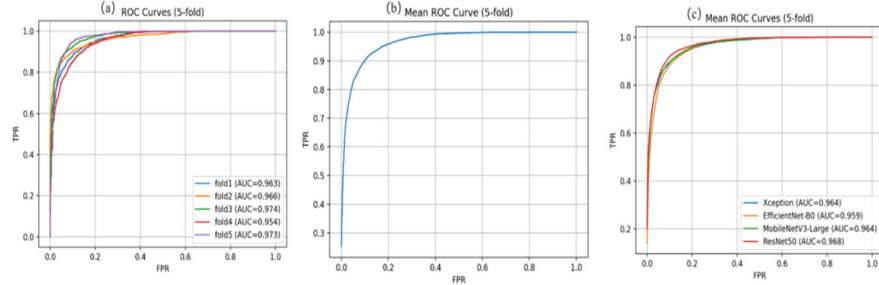


Figure 4 ROC curve analysis for the proposed deepfake detection model: (a) ROC curves of five-replication cross validation, (b) mean ROC curve among five replications, and (c) comparison of ROC across different backbone networks in a deepfake dataset.

Table 3 State-of-the-art deepfake detection techniques

Method	Accuracy	F1	ROC-AUC	Complexity	Threshold Strategy
Face X-Ray	0.94	0.93	0.97	High	Fixed
SBI	0.95	0.94	0.98	High	Fixed
Multi-attention	0.96	0.95	0.98	Very High	Fixed
This Work	0.996	0.9959	0.9999	Low	Adaptive

was in line with the quantitative comparison in Table 2. MobileNetV3-Large was also relatively good, and Xception and EfficientNet-B0 maintained good separation but had a slight decrease in overall classification accuracy.

As shown in Figure 4, the results of the ROC curve also demonstrate the discriminatory capacity of the tested backbone models. Five-fold ROC curves show that the detection framework proposed has consistent classification results in all the validation folds. Based on the mean ROC curve, the model can distinguish between real and fake samples reasonably well. In addition, as shown in Figure 4(c), the backbone comparison is consistent with the quantitative results in Table 2; ResNet50 has the best overall performance among the tested models. Some typical deepfake detection methods have also been added to the table for comparison.

As shown in Table 3, to further assess the performance of the proposed approach, it is also compared against other top-tier deepfake detection algorithms: Face X-Ray, self-blended images (SBI) [33], and multi-attribution-based models [34]. Table 3 shows that these methods have yielded similar or better performance on the aforementioned evaluation indices. Although these comparisons are taken directly from the original papers, they also show that our approach is effective and has good generalization ability.

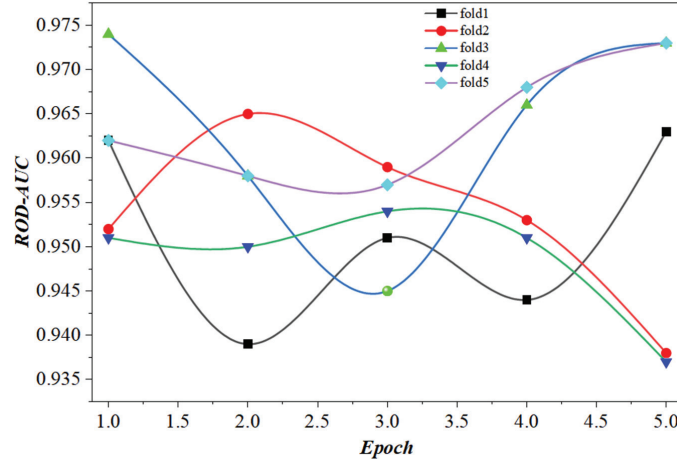


Figure 5 Variation of AUC across training epochs in five-fold cross-validation.

Table 4 Ablation comparison of baseline, threshold tuning, and SE attention

Method	Accuracy	F1-score	ROC-AUC	PR-AUC
ResNet18 (baseline)	0.8835 ± 0.0186	0.8773 ± 0.0253	0.9655 ± 0.0065	0.9661 ± 0.0057
ResNet18+	0.9016 ± 0.0151	0.9042 ± 0.0138	0.9654 ± 0.0064	0.9660 ± 0.0056
Threshold Tuning				
ResNet18+SE+	0.9006 ± 0.0136	0.9019 ± 0.0134	0.9647 ± 0.0060	0.9646 ± 0.0066
Threshold Tuning				

As shown in Figure 5, the ROC-AUC values remain at a relatively high level across the five training epochs in the five-fold cross-validation setting, although slight fluctuations can be observed among different folds. This result indicates that the model maintains stable discrimination ability during training.

To further evaluate the individual contributions of adaptive threshold optimization and SE attention, an ablation study was conducted using ResNet18 as the baseline backbone. The ablation results are summarized in Table 4.

As shown in Table 4, threshold tuning improved the accuracy from 0.8835 ± 0.0186 to 0.9016 ± 0.0151 and increased the F1-score from 0.8773 ± 0.0253 to 0.9042 ± 0.0138 , while ROC-AUC and PR-AUC remained almost unchanged. This indicates that adaptive threshold optimization mainly improves the final classification decision boundary rather than the ranking ability of the classifier. After SE attention was added to the threshold-tuned model, the performance remained stable, with Accuracy = 0.9006 ± 0.0136 , F1-score = 0.9019 ± 0.0134 , ROC-AUC =

Table 5 Threshold optimization effect (5-fold mean $\pm$ std)

Setting	Threshold	Accuracy (mean $\pm$ std)	F1-score (mean $\pm$ std)
Default	0.5	0.8877 ± 0.0170	0.8853 ± 0.0211
Threshold tuning	0.326 ± 0.251	0.9016 ± 0.0151	0.9042 ± 0.0138

Table 6 Hard-test performance under group-level splitting

Model	Split Strategy	Train/Val/Test	Accuracy	F1-score	ROC-AUC	PR-AUC	Confusion Matrix (Fake/Real)
ResNet50	Group-level split	3619/533/998	0.9960	0.9959	0.9999	0.9999	[508,2],[2,486]

0.9647 ± 0.0060 , and $\text{PR-AUC} = 0.9646 \pm 0.0066$. Although the gain from SE attention was limited, it provided stable feature-representation support. Overall, the ablation results show that adaptive threshold optimization contributed more directly to the improvement in accuracy and F1-score, while SE attention helped maintain stable representation learning.

As shown in Table 5, threshold tuning improved the mean accuracy from 0.8877 ± 0.0170 to 0.9016 ± 0.0151 and the mean F1-score from 0.8853 ± 0.0211 to 0.9042 ± 0.0138 under five-fold cross-validation. This result further confirms that the adaptive threshold strategy improves the final classification decision compared with the default threshold of 0.5.

To further explore the generalization capability of the above framework, a strict group-level hard-test environment has been built. Under this arrangement, samples extracted from the same source video were assigned to only one subset, and the training, validation, and test sets remained independent at the group level of videos. The results of the hard-test are shown in Table 6. The ResNet50-based model had an accuracy of 0.9960, an F1-score of 0.9959, a ROC-AUC of 0.9999, and a PR-AUC of 0.9999. Based on the above results, the proposed framework has good classification stability in a strict group-level evaluation.

As shown in Figure 6, the confusion matrix further illustrates the classification performance of the proposed model under the hard-test setting. In the normalized confusion matrix shown in Figure 6(a), both fake and real samples were correctly classified at approximately 99.6%. In the count-based confusion matrix shown in Figure 6(b), 508 fake samples and 486 real samples were correctly classified, while only two fake samples and two real samples were misclassified. Based on the results reported in Table 6 and Figure 6, the proposed method demonstrates high stability and good generalization ability under the strict group-level hard-test setting.

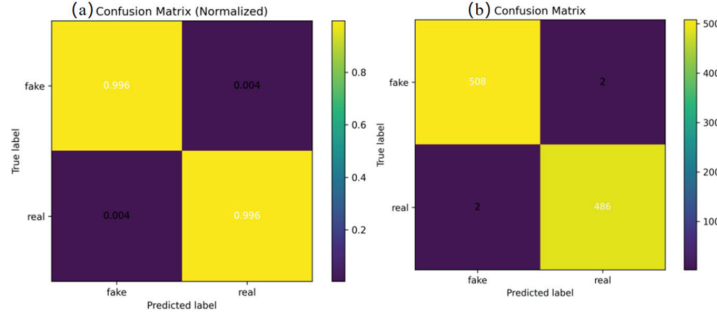


Figure 6 Confusion matrix analysis in hard-test setting: (a) normalized confusion matrix and (b) confusion matrix with counts.

5 Conclusion

This study proposes a lightweight deepfake image detection framework that integrates CNN-based feature extraction, SE attention, and adaptive threshold optimization. Experimental results demonstrate that the proposed strategy improves detection performance while maintaining a simple and deployable structure. In five-fold cross-validation, ResNet50 achieved the best overall performance among the tested backbones, with an accuracy of 0.9061 ± 0.0161 , F1-score of 0.9065 ± 0.0185 , ROC-AUC of 0.9680 ± 0.0075 , and PR-AUC of 0.9677 ± 0.0073 . The threshold optimization strategy improved the ResNet18 baseline accuracy from 0.8835 ± 0.0186 to 0.9016 ± 0.0151 and the F1-score from 0.8773 ± 0.0253 to 0.9042 ± 0.0138 . Under the strict group-level test setting, the ResNet50-based model achieved 0.9960 accuracy and 0.9959 F1-score, indicating strong classification stability on the constructed test split.

The main innovation of this work is not the design of a heavier detection backbone, but the improvement of the decision mechanism in CNN-based deepfake image detection. Specifically, the SE module enhances channel-wise forgery-related feature representation, while the adaptive threshold strategy adjusts the final classification boundary according to validation-set performance instead of relying on the conventional fixed threshold of 0.5. This combination provides a practical way to improve detection reliability without introducing complex multimodal fusion, heavy temporal modeling, or substantial architectural overhead.

The hypothesis of this study was that deepfake image detection can be improved by jointly enhancing forgery-sensitive visual representation and adapting the classification threshold to the data distribution. The experimental

results support this hypothesis. The improvement in accuracy and F1-score after threshold optimization shows that a fixed threshold may be insufficient under practical data variations. In addition, the use of group-level data splitting reduces the risk of identity or video-level leakage, making the evaluation more reliable for social-network visual-content security scenarios. Future work will further evaluate the proposed framework on the full FaceForensics++ dataset and external benchmarks such as Celeb-DF and DeeperForensics to examine cross-dataset generalization. The robustness of the method will also be tested under real-world social-media degradations, including recompression, resolution changes, cropping, filtering, and unseen manipulation techniques. In addition, temporal cues and multimodal information may be incorporated to extend the current visual detector toward broader misinformation detection. These directions will further improve the applicability of the proposed method in cyber-security, social-media content verification, and trustworthy digital-media authentication.

References

- [1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–11, 2019.
- [2] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213, 2020.
- [3] Chollet, F. “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258, 2017.
- [4] Hu, J., Shen, L., and Sun, G. “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.
- [5] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. “CBAM: Convolutional block attention module,” in Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol. 11211, pp. 3–19, Springer, 2018.
- [6] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018.

- [7] Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. “Face2Face: Real-time face capture and reenactment of RGB videos,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395, 2016.
- [8] Thies, J., Zollhöfer, M., and Nießner, M. “Deferred neural rendering: Image synthesis using neural textures,” *ACM Transactions on Graphics*, vol. 38, no. 4, Article 66, pp. 1–12, 2019.
- [9] Ciftci, U. A., Demir, I., and Yin, L. “FakeCatcher: Detection of synthetic portrait videos using biological signals,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2365–2379, 2021.
- [10] Shang, Z., Xie, H., Zha, Z., Yu, L., Li, Y., and Zhang, Y. “PRRNet: Pixel-region relation network for face forgery detection,” *Pattern Recognition*, vol. 116, Article 107950, 2021.
- [11] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [12] Cheng, Z., Wang, Y., Wan, Y., and Jiang, C. “DeepFake detection method based on multi-scale interactive dual-stream network,” *Journal of Visual Communication and Image Representation*, vol. 104, Article 104263, 2024.
- [13] Xu, Q., Chen, H., Du, H., Zhang, H., Lukasik, S., Zhu, T., and Yu, X. “M3A: A multimodal misinformation dataset for media authenticity analysis,” *Computer Vision and Image Understanding*, vol. 249, Article 104205, 2024.
- [14] Ge, S., and Chen, Y. “Confidence-aware multimodal learning for trustworthy fake news detection,” *INFORMS Journal on Computing*, 2025. DOI:10.1287/ijoc.2024.0655.
- [15] Zhang, D., Zhu, W., Ding, X., Yang, G., Li, F., Deng, Z., and Song, Y. “SRTNet: A spatial and residual based two-stream neural network for deepfakes detection,” *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 14859–14877, 2023.
- [16] Ni, Y., Zeng, W., Xia, P., Yang, G. S., and Tan, R. “A deepfake detection algorithm based on Fourier transform of biological signals,” *Computers, Materials & Continua*, vol. 79, no. 3, pp. 5295–5312, 2024.
- [17] Hernandez-Ortega, J., Tolosana, R., Fierrez, J., and Morales, A. “Deep-Fakes detection based on heart rate estimation: Single- and multi-frame,” in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*, pp. 255–273, Springer, 2022.

- [18] Borji, A. “Qualitative failures of image generation models and their application in detecting deepfakes,” *Image and Vision Computing*, vol. 137, Article 104771, 2023.
- [19] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. “Face X-Ray for more general face forgery detection,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009, 2020.
- [20] Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. “DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2889–2898, 2020.
- [21] Shiohara, K., and Yamasaki, T. “Detecting deepfakes with self-blended images,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.
- [22] Qi, H., Guo, Q., Juefei-Xu, F., Xie, X., Ma, L., Feng, W., Liu, Y., and Zhao, J. “DeepRhythm: Exposing DeepFakes with attentional visual heartbeat rhythms,” in *Proc. 28th ACM International Conference on Multimedia (ACM MM)*, pp. 4318–4327, 2020.
- [23] Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in *Computer Vision – ECCV 2020, Lecture Notes in Computer Science*, vol. 12357, pp. 86–103, Springer, 2020.
- [24] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. “Multi-attentional deepfake detection,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2185–2194, 2021.
- [25] Sun, Z., Han, Y., Hua, Z., Ruan, N., and Jia, W. “Improving the efficiency and robustness of deepfakes detection through precise geometric features,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3609–3618, 2021.
- [26] Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., and Xia, W. “Learning self-consistency for deepfake detection,” in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15003–15013, 2021.
- [27] Mirsky, Y., and Lee, W. “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys*, vol. 54, no. 1, Article 7, pp. 1–41, 2021. DOI:10.1145/3425780.

- [28] Jing, J., Wu, H., Sun, J., Fang, X., and Zhang, H. “Multimodal fake news detection via progressive fusion networks,” *Information Processing & Management*, vol. 60, no. 1, Article 103120, 2023.
- [29] Xu, Q., Qiao, H., Liu, S., and Liu, S. “Deepfake detection based on remote photoplethysmography,” *Multimedia Tools and Applications*, vol. 82, no. 23, pp. 35439–35456, 2023.
- [30] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., and Mazibuko, T. F. “An improved dense CNN architecture for deepfake image detection,” *IEEE Access*, vol. 11, pp. 22081–22095, 2023.
- [31] Wang, C., Ma, W., Zou, L., Xia, Z., Li, Q., Ma, B., and Liu, Y. “Toward robust deepfake detection: A proactive method based on watermarking and knowledge distillation,” in *Proc. ACM International Conference on Multimedia (ACM MM)*, pp. 4798–4807, 2025.
- [32] Lipton, Z. C., Elkan, C., and Naryanaswamy, B. “Optimal thresholding of classifiers to maximize F1 measure,” in *Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science*, vol. 8725, pp. 225–239, Springer, 2014.
- [33] Steinebach, M., Liu, H., and Gotkowski, K. “Fake news detection by image montage recognition,” *Journal of Cyber Security and Mobility*, vol. 9, no. 2, pp. 175–202, 2020.
- [34] Steinebach, M., Berwanger, T., and Liu, H. “Image hashing robust against cropping and rotation,” *Journal of Cyber Security and Mobility*, vol. 12, no. 2, pp. 129–160, 2023.

Biography



Chaoran Li focuses on intelligent visual security and content authentication, digital cultural creativity, and film and television narrative studies. She has presided over and participated in multiple provincial-level key projects,

university-level key projects, and interdisciplinary co-construction projects. She has published numerous academic papers covering film and television text narrative analysis, social media visual communication, cross-cultural educational communication, and other related fields, and possesses a solid theoretical foundation and professional research capabilities. Leveraging her interdisciplinary academic background in design studies and film studies, she excels at integrating visual expression, narrative theory, and intelligent technologies, and is committed to advancing the integrated application of artificial intelligence in network content security and digital cultural innovation.

