
Robust Multimodal Deepfake Forensics for Digital Human Identity Protection in Mobile Multimedia Security

Xiao Han

Guilin Institute of Information Technology, Guilin 541004, Guangxi, China
E-mail: xiaohan998@outlook.com

Received 22 May 2026; Accepted 14 July 2026

Abstract

Deepfake forgeries pose increasing risks to digital identity protection, media integrity, and forensic reliability in mobile multimedia environments. Although recent detection methods have shown promising results on benchmark datasets, their performance often degrades under cross-domain distribution shifts and real-world perturbations, particularly in digital-human scenarios with complex appearance and motion patterns. To address this issue, this paper proposes a robust multimodal deepfake forensics framework for digital human identity protection. The proposed method jointly exploits visual, audio, and motion cues and further enhances deployment robustness through perturbation-aware learning, adversarial robustness enhancement, and domain-specific adaptation. Experimental results show that the EfficientNet-based implementation achieves ACC/F1-score/AUC values of 0.9245/0.9246/0.9728 on the benchmark setting. Under cross-dataset evaluation, the proposed framework obtains ACC/F1-score/AUC values of 0.8612/0.8729/0.9285 on Celeb-DFv2 and 0.8346/0.8481/0.9043 on WildDeepfake. After opera-domain fine-tuning, performance on OperaDeepfake improves from 0.8415/0.8568/0.9017 to 0.9317/0.9385/0.9781.

Under five perturbation settings, the proposed robustness-oriented training strategy improves the F1-score by 0.0662-0.0909 compared with the baseline. These quantitative results demonstrate that the proposed framework provides a practical and robust solution for deepfake forensics, digital human identity protection, and mobile multimedia security applications.

Keywords: Deepfake forensics, digital identity protection, mobile multimedia security, multimodal detection, adversarial robustness, domain adaptation.

1 Introduction

Deepfake generation has rapidly evolved from early face-swapping pipelines to highly realistic neural synthesis systems, making forged audiovisual content increasingly difficult to distinguish from authentic media in open and mobile network environments [1]. This trend creates serious risks for digital identity protection, multimedia integrity, remote communication, and forensic verification. In mobile multimedia systems [2], forged videos can be rapidly disseminated after compression, retransmission, resizing, and illumination variation, which further obscures manipulation traces [3]. Accordingly, robust deepfake forensics should be viewed not only as a multimedia analysis task but also as an important problem in digital forensics [4], privacy protection, and security-oriented multimedia applications. Although existing detectors often report promising results on public benchmarks, their performance drops noticeably under cross-dataset deployment, realistic perturbations, and domain shifts [5–8]. Recent studies have attempted to improve deepfake forensics from different directions, including transformer-based representation learning, frequency-domain artifact mining, multimodal audio-visual consistency modeling, adversarial robustness training, and domain-generalization strategies [9]. These methods have advanced the detection of general facial forgeries, but most of them are still designed for conventional benchmark datasets and do not explicitly address opera-oriented digital humans [10], where appearance, motion, illumination, and audio-visual patterns differ substantially from ordinary face videos. Most current methods rely primarily on visual artifacts learned from common datasets and therefore remain vulnerable to compression [11], blur, noise, contrast variation, and other degradations frequently encountered in practical transmission environments. This limitation becomes more pronounced in opera-oriented digital-human scenarios, where heavy makeup [12], facial decorations,

exaggerated expressions, and complex stage lighting substantially alter facial appearance statistics, temporal motion patterns, and audio-visual consistency. As a result, models trained on conventional benchmarks often suffer from severe generalization failure when transferred to this domain [13].

This issue is very serious from a security standpoint because forged opera-style digital humans can be used to impersonate real people, hold fake performances, mislead the public, and circumvent content verification systems in network media. However, the current deepfake forensic studies still have three main deficiencies [14]. First, most public datasets and evaluation protocols are designed for general facial forgeries and provide limited coverage of opera-specific visual features, such as heavy makeup, facial decoration, exaggerated expressions, complex lighting, and stylized motion. Second, many robustness-oriented detectors improve performance under general perturbations but rarely consider the combined effects of mobile multimedia degradation, adversarial-like disturbances, and operational-domain distribution shifts [15]. Third, single-modality visual detectors may fail if appearance artifacts have been reduced by compression or if forged content maintains plausible facial texture but disrupts audio-visual or motion consistency. Therefore, it can be seen that strengthening the backbone network alone is insufficient; a deployment-oriented forensic framework should incorporate domain-specific data, multimodal inconsistency modeling, robustness enhancement, and adaptive decision calibration simultaneously [16].

As shown in Figure 1, this paper presents a robust multimodal deepfake forensics framework for the protection of digital human identities in mobile multimedia security. The main new feature of this paper is not the independent application of multimodal fusion, adversarial training, and domain adaptation, but rather their combined use for opera-style digital-human identity protection. First, a dataset of OperaDeepfake is built in this work, containing genuine and forged Chinese opera digital-human samples generated by representative talking-face synthesis tools such as Wav2Lip and SadTalker [17]. Then, it collects corresponding evidence from the visual, audio, and motion branches to learn facial artifact patterns, voice-consistency cues, and temporal motion inconsistencies [18]. To enhance the robustness of the deployment, perturbation-aware augmentation and PGD-based adversarial training are introduced to improve resistance to compression, blur, noise, illumination variation, and adversarial-like disturbances [19]. Opera-domain fine-tuning is used to reduce the discrepancy between general deepfake data and opera-specific facial distributions further [20]. Finally, adaptively calibrated thresholds and Grad-CAM-based interpretability analyses are used

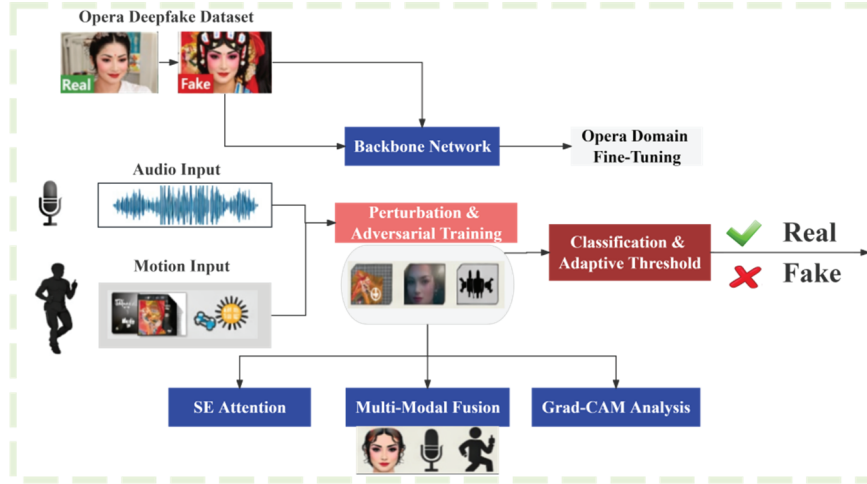


Figure 1 Overall framework of the proposed robust multimodal deepfake forensics method for digital human identity protection.

to enhance the robustness and forensic transparency of the decisions [21]. Thus, the proposed framework is not the same as the previous ones; it jointly addresses domain specificity, multimodal inconsistency, robustness to mobile multimedia degradation, and identity-protection-oriented forensic decision-making [22, 23].

The main contributions of this paper are as follows:

- (1) We explain that opera-oriented digital-human deepfake forensics is a mobile multimedia security problem, and we clarify its relationship to digital identity protection, content authentication, and networked media verification.
- (2) We construct and characterize an OperaDeepfake dataset for domain-specific evaluation, including both authentic and forged Chinese opera digital-human samples with opera-style appearance, motion, and illumination features.
- (3) We propose a deployment-oriented multimodal forensic framework that integrates visual artifact learning, audio-consistency modeling, motion-dynamics analysis, perturbation-aware augmentation, projected gradient descent (PGD)-based robustness enhancement, domain adaptation, and adaptive threshold calibration.
- (4) We establish an all-encompassing evaluation system in benchmark, cross-dataset, perturbation, in-the-wild, and opera-domain environments

to verify the effectiveness, robustness, and practical value of the proposed framework for mobile multimedia security.

2 Related Work

2.1 Deepfake Detection Methods

Early deepfake detection approaches are mainly composed of visual backbones based on convolutional neural networks (CNNs); for example, Xception and EfficientNet that can detect manipulation artefacts through spatial texture inconsistency [1, 2].

Given an input face frame, the basic CNN-based detection process is formulated in Equation (1).

$$\begin{cases} z = f_{\theta}(x) \\ p = \sigma(z) \end{cases} \quad (1)$$

where f_{θ} denotes the backbone network, $\sigma(\cdot)$ denotes the sigmoid activation function, z represents the extracted feature/logit representation, and p denotes the predicted probability of the input being fake. The binary cross-entropy loss used for detector optimization is given in Equation (2).

$$L_{ce} = -y \log p - (1 - y) \log(1 - p) \quad (2)$$

where y denotes the ground-truth label, and p denotes the predicted fake probability. This loss encourages the detector to assign higher probabilities to forged samples and lower probabilities to authentic samples. Recent studies have further explored long-range dependency modeling through transformer-based architectures such as ViT, where the transformer attention operation is expressed in Equation (3).

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (3)$$

These models can capture richer global representations; however, they still remain susceptible to domain shifts and dataset biases. Besides spatial-domain learning, frequency-domain and physiological-cue methods have also been explored, e.g., spectrum inconsistency and rPPG-related artifacts [24]. A common limitation is that most of these approaches rely predominantly on visual data and may degrade under substantial changes in facial appearance and structure.

2.2 Generalization and Robustness

Cross-dataset generalization remains a major challenge because models trained in one environment often perform poorly when tested in another. Prior studies have therefore introduced perturbation augmentation, such as compression, blur, noise injection, and illumination variation, as well as adversarial training, to improve robustness [25].

PGD-based adversarial optimization is widely adopted to enhance robustness against worst-case perturbations during training, and the PGD-based adversarial update is shown in Equation (4).

$$x^{(t+1)} = \Pi_{B_{\epsilon}(x)}(x^{(t)} + \alpha \text{sign}(\nabla_x L_{ce}(f\theta^{(x^{(t)})}, y))) \quad (4)$$

where $x^{(t)}$ denotes the adversarial sample at the t -th iteration, α is the step size, $\nabla_x L_{ce}$ represents the gradient of the classification loss with respect to the input, and $\Pi_{B(x)}$ denotes the projection operation that constrains the perturbation within a predefined admissible region. Based on this adversarial update, the robust training objective is defined in Equation (5).

$$\min_{\theta} E_{(x,y)}[L_{ce}(f_{\theta}^{(x)}, y) + \lambda L_{ce}(f\theta(x_{adv}), y)] \quad (5)$$

Although these approaches reduce perturbation sensitivity, their effectiveness under domain-specific generalization settings remains limited.

2.3 Multimodal Deepfake Detection

Recent studies have explored combinations of facial, audio, and motion cues to capture inconsistencies across multiple modalities in forged content [17]. Let the visual, audio, and motion branches denote three complementary feature representations for multimodal forgery analysis. The modality-specific representations are defined in Equation (6).

$$\begin{cases} h_v = \phi_v(v) \\ h_a = \phi_a(a) \\ h_m = \phi_m(m) \end{cases} \quad (6)$$

where h_v , h_a , and h_m denote the feature representations extracted from the visual, audio, and motion branches, respectively. These representations provide complementary evidence for detecting facial artifacts, voice-consistency abnormalities, and temporal motion inconsistencies. A common strategy is

late fusion, where the modality-specific features or predictions are aggregated at the decision level. The late-fusion strategy is formulated in Equation (7).

$$\begin{cases} h = [h_v; h_a; h_m] \\ p = \sigma(W_h + b) \end{cases} \quad (7)$$

where the multimodal representation h is obtained by concatenating the visual, audio, and motion features, and the fused prediction is produced through a sigmoid classifier.

Another strategy is score-level fusion, which combines the confidence scores from different modalities. The score-level fusion rule is given in Equation (8).

$$\begin{cases} s = \alpha s_v + \beta s_a + \gamma s_m \\ \alpha + \beta + \gamma = 1 \end{cases} \quad (8)$$

where s_v , s_a , and s_m denote the visual, audio, and motion confidence scores, respectively, while α , β , and γ are the corresponding modality weights. Cross-attention mechanisms have also been introduced to improve modality alignment and multimodal interaction. The cross-attention-based multimodal interaction is described in Equation (9).

$$\begin{cases} h_v = \text{Attn}(h_v, h_a, h_m) \\ h_a = \text{Attn}(h_a, h_v, h_m) \\ h_m = \text{Attn}(h_m, h_v, h_a) \end{cases} \quad (9)$$

In Equation (9), each modality is updated by attending to complementary information from the other modalities, thereby improving cross-modal consistency modeling. Despite recent progress, multimodal deepfake detection in opera-oriented digital-human scenarios remains underexplored.

2.4 Limitations of Existing Studies

In summary, prior work has advanced visual detection, robust optimization, and multimodal fusion, but several gaps still remain for opera-oriented digital humans.

- (1) Existing studies lack domain-specific opera deepfake datasets and dedicated evaluation protocols.
- (2) Robustness-oriented training strategies that explicitly adapt to opera-like perturbations remain insufficiently explored.

- (3) Multimodal consistency modeling under makeup-heavy, occlusion-rich, and lighting-variant conditions is still limited.

These gaps motivate the development of an opera-oriented robust multimodal deepfake forensics framework.

3 Methodology

3.1 Framework Overview

The proposed framework processes both images and videos and produces a binary prediction indicating whether the input content is authentic or forged.

As illustrated in Figure 2, the framework consists of three main components:

- (1) Visual branch; face-level forgery representation learning using CNN- or transformer-based backbones.
- (2) Audio branch; spectral and voice-consistency cue extraction from synchronized audio signals.
- (3) Motion branch; temporal dynamics modeling of mouth and facial movement trajectories.

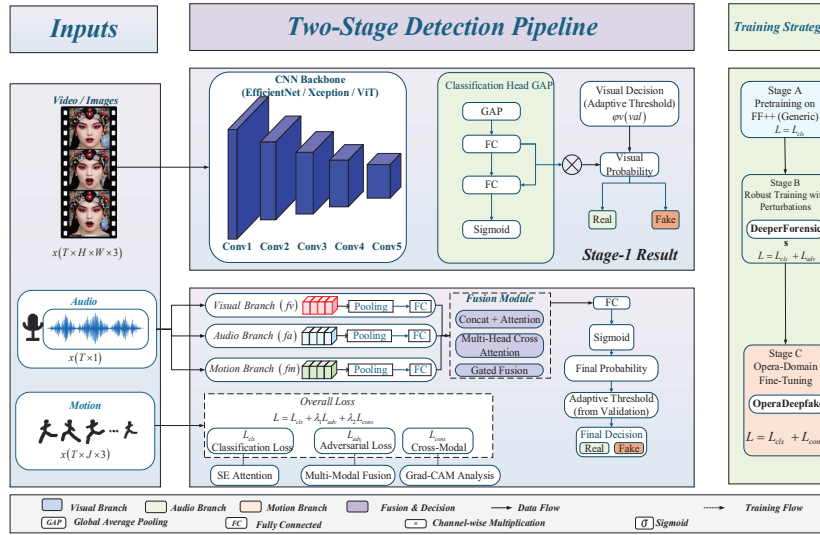


Figure 2 Overall architecture of the proposed robust multimodal deepfake forensics framework for opera-oriented digital humans.

The final prediction is obtained through multimodal fusion together with an adaptive threshold calibration strategy.

3.2 Dataset and Training Protocol

We adopt a staged training-and-evaluation protocol to assess generic detection capability, cross-domain transferability, robustness, and domain adaptation:

- (1) Pretraining on FaceForensics++ (FF++);
- (2) Cross-dataset testing on Celeb-DFv2;
- (3) Robustness evaluation under DeeperForensics-style perturbation settings;
- (4) In-the-wild generalization testing on WildDeepfake;
- (5) Opera-domain adaptation and evaluation on OperaDeepfake.

An OperaDeepfake dataset was also built to assess the performance of the proposed framework in opera-oriented digital-human situations. The set of data has 720 face images, 240 of which are real and 480 of which are fake. The real subset has 8 opera-style digital-human identities, and the fake subset contains 16 generated identity/video groups. For each identity or generated group, after face detection, quality filtering, and identity-level checking, 30 representative face frames were extracted. Representative talking-face synthesis tools were used to create the forged samples, including Wav2Lip and SadTalker, and both lip-synchronization manipulation and portrait animation were covered.

For annotation, each sample was labelled as real or fake according to its generation source, and these labels were further verified manually to exclude failed face crops, severe occlusions, and visually corrupted frames. To avoid identity leakage, the division of the dataset was carried out at the level of identity/video group rather than at the frame level. Therefore, frames from the same identity or generated video groups were assigned to only one subset. In addition to the dataset of face images, it also contains synchronous video and audio files, metadata records, and identity-level split files for training, validation, and testing. The above design supports the assessment of visual artefacts, audio-consistency cues, and motion inconsistencies in a single multimodal forensic environment. This design provides a unified protocol for evaluating base accuracy, cross-domain transfer, perturbation robustness, and domain-specific adaptability [26–28].

3.3 Framework Design

Let x denote a cropped face input, and let the backbone network, such as Xception, EfficientNet, or ViT, produce the corresponding visual feature representation. The visual feature extraction process is defined in Equation (10).

$$\begin{cases} s_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W U_c(i, j) \end{cases} \quad (10)$$

The visual branch outputs a confidence score indicating the probability that the input is forged. To emphasize manipulation-sensitive channels, we incorporate squeeze-and-excitation attention into the visual representation learning process. The squeeze-and-excitation attention operation is formulated in Equation (11).

$$\begin{cases} s_c = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W U_c(i, j) \\ \hat{s} = \sigma(W_2 \delta(W_1 s)) \\ U_c = \hat{s} \cdot U_c \end{cases} \quad (11)$$

Given a feature map, channel-wise reweighting is applied to enhance informative responses and suppress irrelevant activations. The detector is optimized using a binary classification objective function. The binary classification objective used for training is given in Equation (12).

$$L_{ce} = -y \log p - (1 - y) \log(1 - p) \quad (12)$$

3.4 Training Strategy

Stage I: initial visual decision. A preliminary visual prediction is obtained by thresholding the visual confidence score. The initial visual decision rule is shown in Equation (13).

$$y_1 = 1(p_v \geq \tau_1) \quad (13)$$

Stage II: multimodal re-decision. Audio and motion scores are fused with the visual score to obtain a refined prediction. The multimodal re-decision process is formulated in Equation (14).

$$\begin{cases} p_f = \alpha p_v + \beta p_a + \gamma p_m \\ \alpha + \beta + \gamma = 1 \end{cases} \quad (14)$$

The final decision is produced after multimodal score aggregation. The final multimodal decision rule is given in Equation (15).

$$y_2 = 1(p_f \geq \tau_2) \quad (15)$$

For adaptive threshold optimization, instead of using a fixed threshold of 0.5, we optimize the decision threshold on the validation set. The adaptive threshold selection rule is formulated in Equation (16).

$$\begin{cases} \tau^* = \arg \max F1(\tau) \\ \tau \in [0, 1] \end{cases} \quad (16)$$

The learned threshold is then applied during inference to adapt to domain-specific score distributions. The training pipeline combines perturbation-aware augmentation with PGD-based adversarial learning.

For perturbation-aware augmentation, training samples are randomly transformed using compression, blur, noise, brightness or contrast variation, and downsampling operations. This perturbation-aware data transformation process is defined in Equation (17).

$$\begin{cases} x' = T(x) \\ T \sim \text{Paug} \end{cases} \quad (17)$$

The PGD-based adversarial update is shown in Equation (18).

$$x^{t+1} = \Pi_{B_{\epsilon}(x)}(x^t + \eta \text{sign}(\nabla_x L_{ce}(f\theta(x^t), y))). \quad (18)$$

The final robust training objective jointly considers clean samples and adversarially perturbed samples, as defined in Equation (19).

$$L = L_{ce}(f\theta(x), y) + \lambda L_{ce}(f\theta(x_{adv}), y) \quad (19)$$

For multimodal features, the fusion objective is designed to encourage complementary information sharing across modalities. The feature-level fusion and final multimodal prediction are formulated in Equation (20).

$$\begin{cases} h = [h_v; h_a; h_m] \\ p_f = \sigma(W_f h + b_f) \end{cases} \quad (20)$$

Both score-level fusion and feature-level fusion are evaluated in practice. The multimodal branch improves reliability for challenging opera cases where visual-only cues are ambiguous, such as heavy makeup, occlusion, stage lighting variation, and exaggerated expressions [29].

Algorithm 1 Robust multimodal deepfake detection for Chinese opera digital humans.

Input: Training sets D_{FF++} , D_{Opera} ; robustness transforms T ; epochs E_1, E_2, E_3 ; fusion weights α, β, γ ; PGD params ϵ, η

Output: Trained model Θ^* , adaptive threshold τ^* Step Pseudo-code

```

1 Initialize backbone  $f_\theta$  (Xception/EfficientNet/ViT), SE module, visual/audio/moti on
  heads, fusion head.
2 Stage A: Generic Pretraining on FF++
3 for epoch=1 to E1 do
4   for  $(x, y)$  in  $D_{FF++}$  do
5     Extract face crop  $x_v$ ; compute visual score  $p_v = f_\theta(x_v)$ .
6     Compute  $L_{ce}$ ; update  $\theta$  by Adam  $W$ .
7   end for
8 end for
9 Stage B: Robust Optimization
10 for epoch=1 to E2 do
11   for  $(x, y)$  in  $D_{FF++}$  do
12     Generate perturbed sample  $x' = T(x)$  (compression/blurr/noise/brightness/
      downsample).
13     Generate adversarial sample  $x_{adv}$  by  $PGD(\epsilon, \eta, K)$ .
14     Compute  $L = L_{ce}(x', y) + \lambda L_{ce}(x_{adv}, y)$ .
15     Update model parameters.
16   end for
17 end for
18 Stage C: Opera Domain Fine-tuning
19 for epoch=1 to E3 do
20   for  $(v, a, m, y)$  in  $D_{Opera}$  do
21     Compute branch scores:  $p_v = \phi_v(v), p_a = \phi_a(a), p_m = \phi_m(m)$ 
22     Fuse:  $p_f = \alpha p_v + \beta p_a + \gamma p_m$ 
23     Compute fusion loss  $L_{fusion}$ ; update parameters.
24   end for
25 end for
26 Adaptive Threshold Selection
27 Search  $\tau \in [0, 1]$  on validation set, choose  $\tau^* = \arg \max_\tau F1(\tau)$ .
28 Inference
29 Stage 1:  $\hat{y}_1 = 1 (p_v \geq \tau_1)$ ;
   Stage 2:  $\hat{y} = 1 (p_f \geq \tau)$ .
30 Return  $\Theta, \tau^*$ .
```

The full optimization pipeline consists of the following stages:

- (1) FF++ pretraining for generic forgery representation learning;
- (2) Perturbation-aware augmentation and PGD-based adversarial training for robustness enhancement;
- (3) Opera-domain fine-tuning on OperaDeepfake;
- (4) Multimodal fusion training and adaptive threshold recalibration.

4 Experiments

4.1 Experimental Setup and Backbone Selection

Evaluations were conducted on five datasets: FaceForensics++, Celeb-DFv2, DeeperForensics-1.0 perturbation settings, WildDeepfake, and OperaDeepfake, as summarized in Table 1. Specifically, FF++ is used for benchmark training, Celeb-DFv2 for cross-dataset testing, DeeperForensics-style perturbations for robustness evaluation, WildDeepfake for in-the-wild generalization, and OperaDeepfake for domain adaptation and opera-specific validation.

We first compare Xception, EfficientNet, and ViT under the same training protocol. Table 2 shows that EfficientNet achieves the best overall accuracy ($\text{ACC} = 0.9245$) and F1-score ($\text{F1} = 0.9246$), outperforming both Xception and ViT under the benchmark setting.

The same trend is also reflected in the ROC curves in Figure 3, where EfficientNet consistently shows superior discrimination capability. Therefore, EfficientNet is selected as the default backbone in the subsequent experiments.

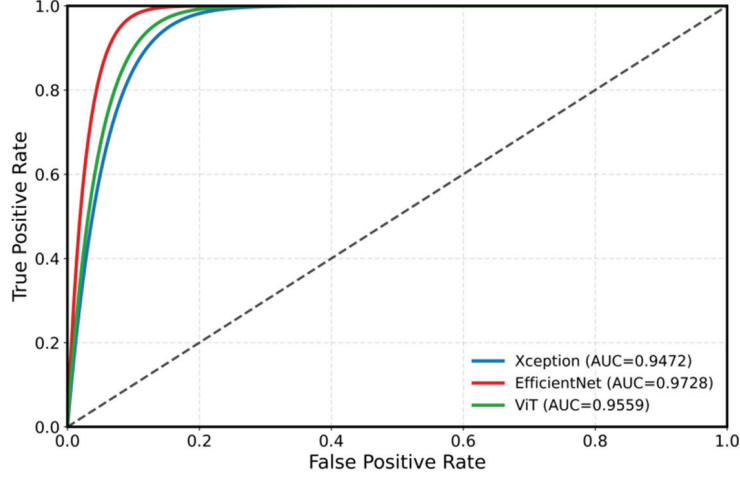
As shown in Table 3, the proposed improvements remain stable across three repeated runs. Opera-domain fine-tuning significantly improves ACC from 0.8415 ± 0.0058 to 0.9317 ± 0.0043 , F1-score from 0.8568 ± 0.0061 to 0.9385 ± 0.0039 , and AUC from 0.9017 ± 0.0049 to 0.9781 ± 0.0028 . Similarly, robustness-oriented training improves all three metrics compared with the plain training setting. The full multimodal setting also outperforms

Table 1 Overview of the datasets used in this study

Dataset	Type	Usage	Real	Fake	Scenario
FaceForensics++	Public	Benchmark	1000	1000	General
Celeb-DFv2	Public	Cross-dataset	590	5639	High-quality
DeeperForensics-1.0	Public	Robustness	1000	10,000	Perturbation-rich
WildDeepfake	Public	In-the-wild	3805	3509	Real-world
OperaDeepfake	Ours	Domain adaptation	240	480	Chinese opera

Table 2 Performance comparison of representative backbone networks

Model	ACC $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$
Xception	0.8920	0.8815	0.9050	0.8931	0.9472
EfficientNet	0.9245	0.9183	0.9310	0.9246	0.9728
ViT	0.9038	0.8971	0.9124	0.9047	0.9559

**Figure 3** ROC curves of representative backbone models.**Table 3** Repeated-run stability of key experimental settings

Comparison Setting	ACC	F1-score	AUC	P
EfficientNet baseline	0.8415 ± 0.0058	0.8568 ± 0.0061	0.9017 ± 0.0049	–
EfficientNet+OperaFT	0.9317 ± 0.0043	0.9385 ± 0.0039	0.9781 ± 0.0028	$P < 0.01$
Plain training	0.8513 ± 0.0072	0.8607 ± 0.0068	0.9128 ± 0.0055	–
Robust training	0.9245 ± 0.0048	0.9246 ± 0.0045	0.9728 ± 0.0031	$P < 0.01$
Visual only	0.9138 ± 0.0056	0.9215 ± 0.0051	0.9634 ± 0.0038	–
Visual+Audio+Motion	0.9317 ± 0.0043	0.9385 ± 0.0039	0.9781 ± 0.0028	$P < 0.05$

Table 4 Cross-dataset evaluation results

Train Set	Test Set	ACC $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$	Note
FF++	Celeb-DFv2	0.8612	0.8729	0.9285	Cross-dataset generalization
FF++	WildDeepfake	0.8346	0.8481	0.9043	In-the-wild evaluation
FF+++Opera FT	OperaDeepfake-Test	0.9317	0.9385	0.9781	Domain-adapted evaluation

the visual-only setting, indicating that audio and motion cues provide complementary forensic evidence for opera-domain samples.

4.2 Cross-Dataset Generalization and Domain Adaptation

As shown in Table 4, to assess cross-domain generalization, the model is trained on FF++ and then evaluated on Celeb-DFv2 and WildDeepfake. As reported in Table 3, the model achieves ACC/F1-score/AUC values of 0.8612/0.8729/0.9285 on Celeb-DFv2 and 0.8346/0.8481/0.9043 on WildDeepfake. These results indicate that the proposed detector retains

Table 5 Effect of opera-domain fine-tuning

Model	Setting	ACC $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$
EfficientNet	Baseline	0.8415	0.8568	0.9017
EfficientNet	+Opera FT	0.9317	0.9385	0.9781

Table 6 Comparison with representative baseline methods on FF++

Method	ACC $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$
MesoNet	0.8410	0.8550	0.9180
Xception	0.8920	0.8931	0.9472
EfficientNet	0.9040	0.9090	0.9560
ViT	0.8980	0.9010	0.9510
Ours (EfficientNet-based)	0.9162	0.9198	0.9685
Ours (EfficientNet + Opera FT)	0.9245	0.9246	0.9728

competitive generalization capability under unseen-domain conditions, although a clear gap remains relative to domain-specific evaluation.

To further address this gap, opera-domain fine-tuning is introduced. As shown in Table 5, ACC increases from 0.8415 to 0.9317, F1-score increases from 0.8568 to 0.9385, and AUC rises from 0.9017 to 0.9781. Figure 5 further illustrates that the fine-tuned model consistently outperforms the non-adapted counterpart, highlighting the importance of domain-specific adaptation for opera-oriented digital-human scenarios.

As reported in Table 6, the proposed method is compared with representative baseline models, including MesoNet, Xception, standard EfficientNet, and ViT. The EfficientNet-based version of our framework achieves the best overall performance with ACC = 0.9245, F1-score = 0.9246, and AUC = 0.9728. Together with the robustness results in Figure 4, these findings suggest that the proposed design improves both benchmark accuracy and deployment stability.

Compared with recent CNN-based, transformer-based, frequency-domain, multimodal, and domain-generalization deepfake detection methods, the proposed framework offers a more all-encompassing solution by jointly considering opera-domain specificity, visual-audio-motion multimodal evidence, perturbation robustness, adaptive threshold calibration, and mobile deployment suitability. This design is especially suitable for opera-oriented digital-human scenarios, as heavy makeup, facial decorations, stylized motion, and complex lighting may reduce the effectiveness of traditional visual-only detection methods.

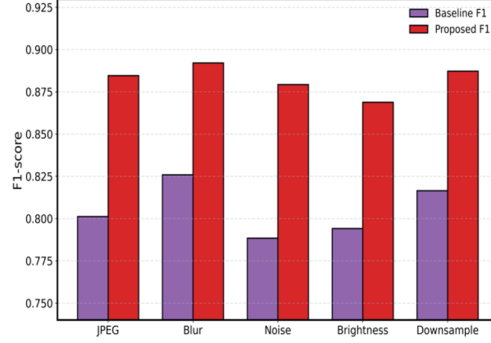


Figure 4 Robustness comparison under different perturbation settings.

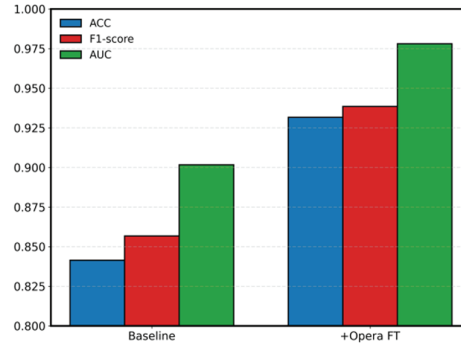


Figure 5 Effect of opera-domain fine-tuning on detection performance.

4.3 Robustness and Ablation Studies

We further evaluate robustness under five typical perturbation settings: JPEG compression, Gaussian blur, Gaussian noise, brightness or contrast variation, and resize downsampling. According to Table 7, the proposed approach consistently outperforms the baseline under all perturbation settings. The corresponding F1-scores gains are 0.0834, 0.0662, 0.0909, 0.0747, and 0.0707, respectively. These results demonstrate that the robustness-oriented training strategy effectively improves detection performance under realistic distortion conditions, with the largest gain observed under Gaussian noise.

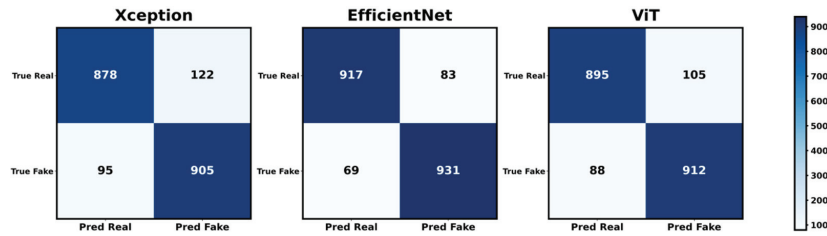
We also investigate the contribution of each training component through ablation analysis, as shown in Table 8. The plain setting (augmentation = none, adversarial training = none) yields the lowest performance, with ACC/F1-score/AUC = 0.8513/0.8607/0.9128. Adding perturbation-aware augmentation alone improves the metrics to 0.8748/0.8821/0.9316,

Table 7 Robustness evaluation under perturbation settings

Perturbation	Baseline F1	Proposed F1	Δ (Improvement)
JPEG Compression	0.8012	0.8846	+0.0834
Gaussian Blur	0.8259	0.8921	+0.0662
Gaussian Noise	0.7884	0.8793	+0.0909
Brightness/Contrast	0.7941	0.8688	+0.0747
Resize Downsample	0.8165	0.8872	+0.0707

Table 8 Ablation study of training strategies

Strategy	ACC $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$
aug=df, adv=None	0.8748	0.8821	0.9316
aug=None, adv=None	0.8513	0.8607	0.9128
aug=df, adv=pgd	0.9245	0.9246	0.9728
aug=None, adv=pgd	0.9031	0.9095	0.9517

**Figure 6** Confusion matrices of representative backbone models.

while using PGD-based adversarial training alone further raises them to 0.9031/0.9095/0.9517. The best result is obtained when both modules are enabled, reaching $\text{ACC}/\text{F1-score}/\text{AUC} = 0.9245/0.9246/0.9728$. This ablation study confirms that perturbation-aware augmentation and adversarial training are complementary rather than redundant.

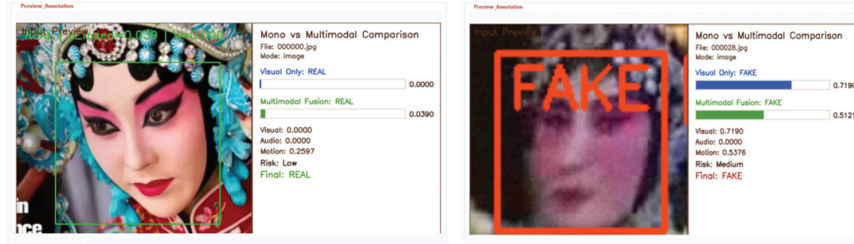
Figure 6 provides the confusion matrices of representative settings to further support the quantitative results. Compared with weaker settings, the stronger model exhibits more concentrated diagonal responses and fewer off-diagonal errors, indicating improved stability, discriminative capability, and practical reliability.

4.4 Domain-Specific Evaluation and Visualization

Domain-specific results on OperaDeepfake are reported in Table 9 to validate performance under opera-oriented conditions. Among the compared models, Efficient Net achieves the best domain-specific performance, with

Table 9 Domain-specific evaluation on OperaDeepfake

Model	ACC $\uparrow$	F1-score $\uparrow$	AUC $\uparrow$
Xception	0.8928	0.9014	0.9536
EfficientNet	0.9317	0.9385	0.9781
ViT	0.9046	0.9129	0.9615

**Figure 7** Visual comparison of prototype-level detection results for representative real and fake opera-domain samples.

ACC/F1-score/AUC = 0.9317/0.9385/0.9781, surpassing Xception (0.8928/0.9014/0.9536) and ViT (0.9046/0.9129/0.9615). These findings indicate that EfficientNet provides stronger feature discrimination under appearance-complex opera scenarios.

As shown in Figure 7, the proposed PyQt-based prototype system can visually distinguish representative real and fake opera-domain samples and provide the corresponding prediction results for practical forensic verification.

As shown in Table 10, adding audio and motion cues consistently improves detection performance compared with the visual-only setting. The visual-only model achieves ACC/F1-score/AUC values of 0.9138/0.9215/0.9634, whereas the full visual-audio-motion configuration improves the results to 0.9317/0.9385/0.9781. Audio cues provide voice-consistency evidence, while motion cues capture temporal mouth and facial movement inconsistencies. This improvement is particularly meaningful for opera-oriented samples because heavy makeup, facial decorations, exaggerated expressions, and stage lighting may weaken purely visual forgery traces. Therefore, the modality ablation results confirm that the proposed multi-modal design provides complementary forensic evidence rather than simply increasing model complexity.

As shown in Table 11, the EfficientNet-based visual branch achieves the best balance between performance and computational cost. Compared with the ViT-based configuration, it has fewer parameters, a smaller model

Table 10 Modality ablation study on OperaDeepfake

Modality Setting	Visual	Audio	Motion	ACC	F1-score	AUC
Visual only	✓	–	–	0.9138	0.9215	0.9634
Visual+Audio	✓	✓	–	0.9226	0.9298	0.9702
Visual+Motion	✓	–	✓	0.9254	0.9321	0.9724
Visual+Audio+Motion	✓	✓	✓	0.9317	0.9385	0.9781

Table 11 Computational efficiency comparison of representative configurations

Model	Params (M)	Size (MB)	Train (min/epoch)	Infer (ms/frame)	FPS	Mem (MB)	Suitability
Xception	22.9	91.6	4.8	13.8	72.5	986	Medium
EfficientNet-Visual	5.3	21.4	3.1	7.9	126.6	742	High
ViT-Visual	86.6	346.4	7.4	25.6	39.1	1638	Medium/Low
Ours-Visual	5.8	23.2	3.4	8.4	119.0	786	High
Ours-Multimodal	8.7	34.8	4.6	12.6	79.4	1035	Medium/High

size, lower memory consumption, and faster inference speed. Although the proposed multimodal framework introduces additional audio and motion branches, the overall overhead remains acceptable because the final decision mainly relies on lightweight score-level and feature-level fusion. The multimodal version still reaches 79.4 FPS, supporting its practical use in mobile multimedia security and digital-human identity verification scenarios.

These results suggest that the proposed framework improves robustness and domain adaptability without introducing prohibitive computational cost, which supports its potential use in mobile multimedia security and digital-human identity protection applications.

5 Conclusion

This paper proposes a robust multimodal deepfake forensics framework for digital human identity protection in mobile multimedia security scenarios. The experimental results show that the EfficientNet-based implementation achieves strong benchmark performance on FF++, competitive cross-dataset generalization on Celeb-DFv2 and WildDeepfake, and clear performance gains after domain-specific fine-tuning on OperaDeepfake. Robustness and ablation results further demonstrate that perturbation-aware augmentation, PGD-based adversarial training, and adaptive threshold calibration improve detection stability under compression, blur, noise, brightness variation, and downsampling.

The main innovation of this study lies in its deployment-oriented integration of opera-domain data construction, visual-audio-motion multimodal inconsistency modeling, robustness enhancement, domain-specific adaptation, and adaptive forensic decision calibration. Rather than treating multimodal fusion, adversarial training, or domain adaptation as isolated techniques, the proposed framework combines them for the specific task of opera-style digital-human identity protection, where heavy makeup, facial decorations, stylized motion, and complex lighting create substantial domain shifts.

The results support the central assumption of this work: single-modality visual detection alone is insufficient for challenging opera-oriented digital-human scenarios, whereas multimodal evidence, robustness-aware training, and opera-domain adaptation can jointly improve detection reliability and practical forensic applicability. The modality ablation, repeated-run stability analysis, and efficiency comparison further confirm that the proposed design improves robustness without introducing prohibitive computational cost.

Future work will focus on expanding OperaDeepfake with more identities, opera styles, generation methods, and real-world transmission conditions. We will also explore stronger cross-modal interaction mechanisms, more lightweight deployment architectures, and broader validation in online media verification and mobile multimedia security applications.

Funding

This work was supported by the 2025 Guangxi Universities Young and Middle-aged Teachers' Basic Research Ability Improvement Project under Grant No. 2025KY1051.

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. DOI: 10.1109/CVPR42600.2020.00327.

- [3] 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. 2886–2895, 2020. DOI: 10.1109/CVPR42600.2020.00296.
- [4] Zi, B., Chang, M., Chen, J., Ma, X., and Jiang, Y.-G. “WildDeepfake: A challenging real-world dataset for deepfake detection,” in Proc. ACM International Conference on Multimedia (ACM MM), pp. 2382–2390, 2020.
- [5] Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. “MesoNet: A compact facial video forgery detection network,” arXiv preprint arXiv:1809.00888, 2018.
- [6] Li, Y., and Lyu, S. “Exposing DeepFake videos by detecting face warping artifacts,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 46–52, 2019.
- [7] Li, Y., Chang, M.-C., and Lyu, S. “In ictu oculi: Exposing AI generated fake face videos by detecting eye blinking,” in Proc. IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018.
- [8] Nguyen, H. H., Yamagishi, J., and Echizen, I. “Capsule-Forensics: Using capsule networks to detect forged images and videos,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311, 2019.
- [9] Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K. “On the detection of digital face manipulation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5780–5789, 2020.
- [10] Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in Proc. European Conference on Computer Vision (ECCV), pp. 86–103, 2020.
- [11] 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. 15023–15033, 2021.
- [12] Luo, Y., Zhang, Y., Yan, J., and Liu, W. “Generalizing face forgery detection with high-frequency features,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16317–16326, 2021.

- [13] Cozzolino, D., Rössler, A., Thies, J., Nießner, M., and Verdoliva, L. “ID-Reveal: Identity-aware DeepFake video detection,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15108–15117, 2021.
- [14] Kwon, P., You, J., Nam, G., Park, S., and Chae, G. “KoDF: A large-scale Korean DeepFake detection dataset,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10744–10753, 2021.
- [15] Le, T.-N., Nguyen, H. H., Yamagishi, J., and Echizen, I. “OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10117–10127, 2021.
- [16] 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.
- [17] Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., and Guo, B. “Protecting celebrities from DeepFake with identity consistency transformer,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9468–9478, 2022.
- [18] 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.
- [19] Mirsky, Y., and Lee, W. “The creation and detection of deepfakes: A survey,” *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021.
- [20] Nanda, S. K., Ghai, D., and Ingole, P. “Analysis of video forensics system for detection of gun, mask and anomaly using soft computing techniques,” *Journal of Cyber Security and Mobility*, vol. 11, no. 4, pp. 549–574, 2022.
- [21] Haddad, N. M., Mustafa, M. S., Salih, H. S., Jaber, M. M., and Ali, M. H. “Analysis of the security of internet of multimedia things in wireless environment,” *Journal of Cyber Security and Mobility*, vol. 13, no. 1, pp. 161–192, 2024.
- [22] Xiao, P. “Network malware detection using deep learning network analysis,” *Journal of Cyber Security and Mobility*, vol. 13, no. 1, pp. 27–52, 2024.
- [23] Xiao, P. “Malware cyber threat intelligence system for Internet of Things (IoT) using machine learning,” *Journal of Cyber Security and Mobility*, vol. 13, no. 1, pp. 53–90, 2024.

- [24] Hussain, S. S., Razak, M. F. A., and Firdaus, A. “Deep learning based hybrid analysis of malware detection and classification: A recent review,” *Journal of Cyber Security and Mobility*, vol. 13, no. 1, pp. 91–134, 2024.
- [25] Singh, V. K., Sivashankar, D., Kundan, K., and Kumari, S. “An efficient intrusion detection and prevention system for DDOS attack in WSN using SS-LSACNN and TCSLR,” *Journal of Cyber Security and Mobility*, vol. 13, no. 1, pp. 135–160, 2024.
- [26] Deng, Y. “Design of industrial IoT intrusion security detection system based on LightGBM feature algorithm and multi-layer perception network,” *Journal of Cyber Security and Mobility*, vol. 13, no. 2, pp. 327–348, 2024.
- [27] Zhang, J., Zou, H., Zeng, Z., Xu, W., and Jiang, J. “Feasibility of using Seq-GAN model in vulnerability detection of industrial control protocols,” *Journal of Cyber Security and Mobility*, vol. 13, no. 3, pp. 393–416, 2024.
- [28] Li, W. “Construction and analysis of QPSO-LSTM model in network security situation prediction,” *Journal of Cyber Security and Mobility*, vol. 13, no. 3, pp. 417–438, 2024.
- [29] Lu, Y. “Security and privacy of Internet of Things: A review of challenges and solutions,” *Journal of Cyber Security and Mobility*, vol. 12, no. 6, pp. 813–844, 2023.

Biography



Xiao Han's primary research focuses on media communication and the protection of intangible cultural heritage. She has led and participated in multiple provincial and ministerial-level research projects, chaired a key interdisciplinary laboratory teaching project in educational planning, and directed several municipal-level projects on rural cultural brand communication, media

convergence, and cultural dissemination from the perspective of intelligent media. She has published numerous academic papers covering topics such as image reconstruction in virtual reality technology and immersive media experiences, demonstrating solid theoretical foundations and independent research capabilities. With her interdisciplinary academic background in communication studies, cultural studies, and digital technologies, Han excels at integrating media theory, cultural narrative, and intelligent systems. She is committed to advancing the innovative application of emerging technologies, including artificial intelligence, in intangible cultural heritage preservation, rural cultural revitalization, and digital content communication.