Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6, 1342-1359
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i6.8119

Deep fake video detection based on multimodal feature fusion: Cross-modal consistency and adversarial enhancement

Ruofan Wang[1,2*], Vladimir Y. Mariano[1]
[1]College of Computing and Information Technologies, National University, Manila 1008, Philippines; f346687454@gmail.com (R.W.).
[2]Department of Electronics Xinzhou Normal University, Xinzhou 034000, Shanxi, China; vymariano@national-u.edu.ph (V.Y.M.).

Abstract: This study proposes a deepfake video detection framework leveraging multimodal feature fusion and adversarial enhancement to address limitations in single-modality detectors for high-quality forgeries and noise interference, systematically integrating cross-modal consistency analysis and robustness training through a tri-modal architecture extracting spatio-temporal visual features via SlowFast-R50, audio context embeddings using VGGish-BiLSTM, and text semantics through Whisper-Transformer, dynamically fused via cross-modal self-attention with adaptive weight allocation, while a dual-branch discriminator jointly optimizes classification accuracy and cross-modal consistency losses; FGSM-based adversarial training injects perturbations in both RGB frame and audio spectrogram domains to enhance robustness against Gaussian/salt-and-pepper noise (σ=0.05/0.02), achieving state-of-the-art performance on FaceForensics++ with video-level accuracies of 98.9% (DeepFake), 98.8% (FaceSwap), 97.6% (Face2Face), and 92.8% (NeuralTextures), exceeding benchmarks like ResNet18 by 1.1–5.1%, maintaining ≥88.5% accuracy under noise and 0.893 ROC-AUC, where multimodal fusion captures subtle cross-modal contradictions while adversarial training ensures stable decision boundaries near perturbation thresholds.

Keywords: *Adversarial Enhancement, Cross-modal Consistency, Deep Fakes, Multimodal Features, Video Detection.*

1. Introduction

With the rapid development of deep generative algorithms such as generative adversarial networks and diffusion models, fake videos based on face replacement and video tampering [1, 2] are becoming increasingly rampant in social media and public opinion scenes. Existing single visual detection methods mostly rely on frame-level image features or frequency domain analysis. When the quality of fake videos [3, 4] continues to improve, the generalization performance and robustness of traditional detectors have significantly decreased. At the same time, videos not only contain static image information but also audio tracks and implied language semantics. The three have natural temporal and semantic consistency in real scenes. The use of cross-modal associations for fake discrimination has achieved initial results. Existing research is limited to shallow alignment in cross-modal consistency and has not explored deep fusion; adversarial enhancement is mostly focused on image classification and lacks a systematic method for video detection. Most studies only stay at the shallow alignment of vision and audio or simple strategies based on lip movement-speech synchronization [5, 6] failing to fully explore the potential of multimodal deep fusion. In addition, research on the robustness of adversarial samples is mostly focused on the field of image classification [7, 8] and there is a lack of systematic methods for adversarial enhancement of fake video detection. Therefore, how to balance multimodal consistency analysis and anti-adversarial attack capabilities while ensuring high detection accuracy has become a key issue that needs to be urgently addressed in the field of deep fake video detection. Although public datasets such as

History: Received: 26 March 2025; Revised: 2 May 2025; Accepted: 4 May 2025; Published: 17 June 2025
* Correspondence: f346687454@gmail.com

FaceForensics++ [9, 10] provide samples of various types of fakes, the performance of existing models on different fake techniques varies significantly, further highlighting the necessity of multimodal and adversarial research.

The core contribution of this study is to systematically construct a trinity framework of multimodal fusion-cross-modal consistency-adversarial enhancement for deep fake video detection. This paper innovatively uses a cross-modal self-attention mechanism to achieve deep interaction between vision, audio, and text, dynamically balances the contribution of the tri-modal through a gated weighted fusion module, and constructs a joint representation with temporal alignment capabilities in a 2048-dimensional feature space. A dual-branch discriminator architecture is proposed. Through the consistency discriminant branch, "highly consistent" and "inconsistent" samples are constrained to be classified into two categories, and cross-modal temporal correlation and semantic contradictions are quantitatively modeled, so that the model's detection accuracy for high-difficulty fake types such as NeuralTextures on the FaceForensics++ dataset is improved. The first multimodal domain adversarial training strategy is created, and FGSM perturbations are injected simultaneously in the visual frame RGB space and the audio spectrogram domain. By jointly optimizing the classification loss, consistency loss, and adversarial loss, the model can still maintain high accuracy under Gaussian noise and salt and pepper noise interference, providing a solution that combines accuracy and robustness for fake detection in complex scenes.

2. Related Work

In the field of deep fake video detection, single-modal visual methods have long dominated the mainstream. Early studies mainly relied on convolutional neural networks [11, 12] to extract features from frame-level images, such as classifiers based on ResNet [13] Xception [14] or Inception architectures, which learned facial texture, local distortion, and inter-frame differences to identify fake traces. With the rapid improvement of generative adversarial networks [15] and diffusion models [16, 17] in video generation quality, the detection accuracy and generalization ability of these single-visual solutions have significantly decreased. In addition, some studies have attempted to apply temporal models [18, 19] to capture continuous changes between frames, but they are still limited by the representation ability of visual information itself and are difficult to deal with the subtle disguises of high-quality fake videos. To supplement the lack of visual information, in recent years, there has also been work that incorporates audio signals [20] into the detection scope. By analyzing lip-speech synchronization, the detector can identify abnormal situations where the audio and video are out of sync. There are also studies that use the joint encoding of spectrograms and visual features to judge authenticity, but most of them stay at the shallow level of splicing or simple weighting, lacking systematic mining of the deep coupling relationship between modalities.

In contrast, multimodal and cross-modal detection methods have gradually emerged in recent years. Some scholars combine the three signals of vision, audio, and text into the classifier through early or late fusion strategies and verify the advantages of multimodal fusion [21, 22] on synthetic datasets or actual recorded videos. Other studies apply cross-modal consistency loss based on contrastive learning or two-stream networks [23, 24] to measure the correspondence between different modal features at the temporal and semantic levels. However, these methods usually ignore the impact of adversarial samples on the robustness of detectors and do not fully consider the interference caused by adversarial perturbations during the training stage. In addition, adversarial training has been widely studied in the field of image classification and object detection, but few works have applied its system into the framework of deep fake video detection. Therefore, combining multimodal fusion with adversarial training to enhance the model's dual defense capabilities against high-quality fakes and adversarial attacks has become a key direction of current research.

Adversarial training technology has made significant progress in the field of image classification [25, 26] but its application in deepfake video detection [27, 28] is still in its infancy. Existing research focuses on the generation and defense of unimodal adversarial samples, lacking a systematic solution for

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

multimodal joint attacks. Research on injecting perturbations in the visual frame or audio spectrum domain through the fast gradient sign method only stays at the data enhancement level, and fails to deeply explore the enhancement mechanism of adversarial training on the cross-modal feature alignment ability. In addition, traditional adversarial training strategies [29, 30] often ignore the particularity of temporal consistency constraints in video detection tasks, resulting in unstable decision boundaries when facing multimodal collaborative attacks. At present, it is urgent to establish a joint framework that integrates adversarial perturbation modeling and cross-modal consistency verification, and to simultaneously optimize the spatial alignment ability of visual-audio-text features during the training phase by designing a multimodal collaborative adversarial loss function [31, 32]. This dual enhancement strategy not only improves the model's sensitivity to high-precision forgery techniques [33, 34] but also significantly improves the model's robustness in noise interference scenarios by learning the smoothness of the decision boundaries of adversarial samples.

3. Methods

3.1. Overall Method Architecture

The method architecture of this paper is shown in Figure 1.

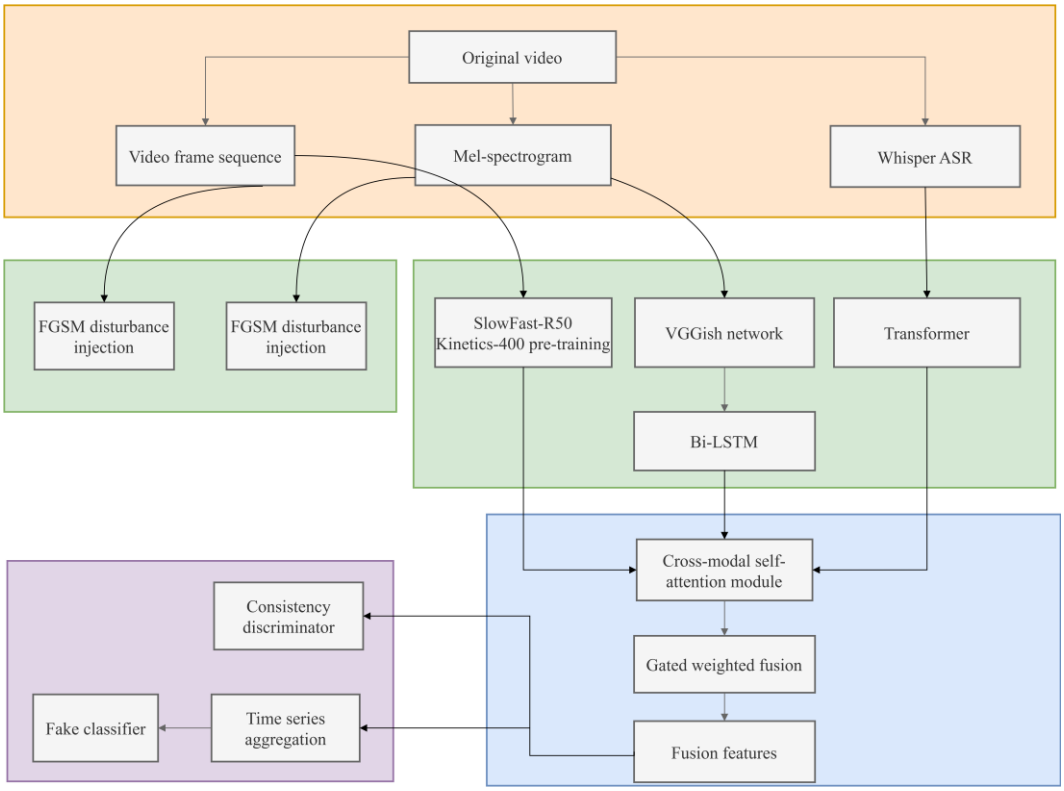

Figure 1.
Method architecture of this paper.

This study proposes a deep fake video detection framework that integrates multimodal features and adversarial enhancement, and achieves efficient detection through cross-modal consistency analysis and robustness training. Multimodal feature extraction is performed on the input video: the visual modality uses the SlowFast-R50 network pre-trained by Kinetics-400 to capture spatio-temporal dynamic features, and uses a dual-path structure to model appearance details and motion information

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

respectively. The audio modality extracts Mel spectrogram features through the VGGish network and then connects to Bi-LSTM to capture temporal context dependencies. The text modality uses Whisper ASR to generate sentence-by-sentence subtitles, and the Transformer encoder parses the semantic vector. Then, a cross-modal self-attention mechanism is designed to deeply interact with the tri-modal features, and the contribution weights of vision, audio, and text are dynamically adjusted through the gated weighted fusion module to construct a joint feature representation with temporal alignment capabilities. On this basis, a dual-branch discrimination system is constructed: the main classifier completes the true and false binary classification based on the fusion features, and the consistency discriminator mines the temporal correlation and semantic contradiction between cross-modal features through an adversarial learning strategy. The two share the backbone network and achieve feature complementary optimization through a joint loss function. To enhance the model's robustness, the FGSM adversarial training method is used to simultaneously inject controllable perturbations in the visual frame RGB space and the audio spectrogram domain, and the optimization direction of the original sample and the adversarial sample is balanced through a hybrid loss function.

3.2. Multimodal Feature Extraction

To achieve comprehensive perception of deep fake videos [35, 36]. This method uses modular and scalable deep encoders on three information streams: visual, audio, and text, and refines the temporal and spatial information.

For the visual stream, SlowFast-R50 spatio-temporal feature encoding is used, which can take into account both high-level semantics and low-level motion details. Given an input video $\{x_t\}_{t=1}^{T}$ of length T frames, each frame is center-cropped and scaled to 256×256, denoted as x_t'.

The sampling strategy is expressed as:

$$X_{\text{slow}} = \{x_t' \mid t = 1,1 + \alpha_s, 1 + 2\alpha_s, \dots\} \quad (1)$$

In Formula 1, $\alpha_s = 8$.

$$X_{\text{fast}} = \{x_t' \mid t = 1,1 + \alpha_f, 1 + 2\alpha_f, \dots\} \quad (2)$$

In Formula 2, $\alpha_f = 32$.

The Slow branch and the Fast branch each generate feature maps $F_{\text{slow}} \in \mathbb{R}^{\frac{T}{\alpha_s} \times 7 \times 7 \times 1024}$ and $F_{\text{fast}} \in \mathbb{R}^{\frac{T}{\alpha f} \times 7 \times 7 \times 1024}$ through a 5-layer ResNet-50 residual block (including bottleneck structure).

Vectors $\mathrm{f}_{\text{slow}}, \mathrm{f}_{\text{fast}} \in \mathbb{R}^{1024}$ are obtained by spatio-temporal global average pooling and concatenated by dimensions:

$F_v = [\mathrm{f}_{\text{slow}}; \mathrm{f}_{\text{fast}}] \in \mathbb{R}^{2048}$ (3)

This fusion retains the advantages of the slow branch in understanding the background and scene, while strengthening the fast branch's ability to capture facial micro-expressions and movements.

The parameters of SlowFast-R50 are shown in Table 1.

Table 1.
SlowFast-R50 parameters.

Parameters	Value	Function
Input resolution	256×256	Unify frame size, balance efficiency and details
Slow sampling rate	8	Capture global semantics and scene dynamics
Fast sampling rate	32	Capture fine-grained motion and micro-expressions
Number of residual block layers	5-layer Bottleneck	Deep semantic extraction and gradient flow stability
Channel dimension	1024	Ensure feature expression capacity
Global pooling output dimension	1024 (per branch)	Spatial dimension reduction, generate fusion pre-vector
Feature dimension after concatenation	2048	Gather slow/fast branch information for subsequent fusion

According to the time-frequency characteristics of the audio signal, the Mel frequency domain conversion is performed, and then the long-term and short-term dependencies are modeled with the

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

help of pre-trained convolutional networks and bidirectional temporal networks. The original waveform resampled to 16 kHz is subjected to the short-time Fourier transform:

$$X[t,k] = \sum_{n=0}^{N-1} x[n+tH]w[n]e^{-\frac{j2\pi kn}{N}} \quad (4)$$

In Formula 4, $x[n]$ is the original waveform; $N = 400$; $H = 160$.

The Mel filtering and logarithmic energy spectrum are:

$$S_{\text{mel}}(t,m) = \log\left(\sum_{k=0}^{N-1} |X[t,k]|^2 H_m(k) + \epsilon\right) \quad (5)$$

In Formula 5, $H_m(k)$ is the m-th Mel filter response, and $\epsilon = 10^{-6}$ is used for numerical stability.

The Mel-spectrogram with a shape of $T_a \times 64$ is input into the VGGish network (6 layers of convolution + 2 layers of full connection) to obtain a 128-dimensional embedding $\{e_t\}_{t=1}^{T_a}$ for each frame.

$\{e_t\}$ is further input into the bidirectional LSTM:

$$\vec{h}_t = \overrightarrow{\text{LSTM}}(e_t, \vec{h}_{t-1}) \quad (6)$$

$$\overleftarrow{h}_t = \overleftarrow{\text{LSTM}}(e_t, \overleftarrow{h}_{t+1}) \quad (7)$$

Finally, the hidden states at both ends are concatenated to obtain the full sentence-level context feature:

$$F_a = [\vec{h}_{T_a}; \overleftarrow{h}_1] \in \mathbb{R}^{512} \quad (8)$$

The parameters of VGGish + Bi-LSTM are shown in Table 2.

Table 2.
VGGish + Bi-LSTM parameters.

Parameters	Value	Function
Audio sampling rate	16 kHz	Ensure the accuracy of speech details and spectrum
Window length/stride	25 ms / 10 ms	Get the balance of time-frequency resolution
Number of Mel filters	64	Extract the frequency bands related to human ear perception
VGGish embedding dimension	128	Generate frame-level high-level semantic representation
Number of Bi-LSTM hidden units	256	Model the long-term and short-term temporal dependencies of audio
Context feature dimension	512	Gather forward and backward information to enhance semantics
Activation function	ReLU (Rectified Linear Unit)	Introduce nonlinearity to prevent sparse expression

The text stream relies on the subtitles generated by end-to-end speech recognition and extracts high-level semantics through a multi-layer self-attention network. The Whisper model is used to transcribe the audio into a text sequence with timestamps and then split into sentences $\{s_i\}_{i=1}^{M}$ at 1-2 s intervals.

Each sentence s_i is input into the WordPiece word segmentation map to a token sequence $\{w_{i,j}\}$, which is mapped to the vector space by the shared embedding matrix:

$$\mathrm{w}_{i,j} = E_{\mathrm{WP}}(w_{i,j}) + PE_j \quad (9)$$

The calculations of the l-th layer of Transformer is:

$$Q = W^Q W^{(l-1)} \quad (10)$$

$$K = W^K W^{(l-1)} \quad (11)$$

$$V = W^V W^{(l-1)} \quad (12)$$

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (13)$$

$$\text{MultiHead}\left(W^{(l-1)}\right) = \text{Concat}(\text{Attn}_1, \ldots, \text{Attn}_H)W^O \quad (14)$$

$$W^{(l)} = \text{LayerNorm}\left(W^{(l-1)} + \text{MultiHead}\left(W^{(l-1)}\right)\right) \quad (15)$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

In Formulas 10-15, H is 8; the dimension of each head is $d_k = d_v = 64$; the model width is $d_{model} = 512$.

The output of the L-th layer is average pooled in the temporal dimension to obtain:

$$F_t = \frac{1}{N_i}\sum_{j=1}^{N_i} W_{i,j}^{(L)} \in \mathbb{R}^{768} \quad (16)$$

In Formula 16, N_i is the number of tokens in the i-th sentence.

The parameters of the Transformer text encoder are shown in Table 3.

Table 3.
Transformer text encoder parameters.

Parameters	Value	Function
Number of layers	6	Control model depth and expressiveness
Number of heads	8	Capture multiple subspace semantics in parallel
Model dimension	512	Unify vector dimensions
Feedforward network dimension	2048	Strengthen feature nonlinear transformation
Dropout	0.1	Relieve overfitting of deep networks

In the tri-modal feature fusion stage, the temporal features of the three streams of vision, audio, and text are linearly projected to obtain their respective query, key, and value vectors. For each pair of modalities, the attention weight is calculated:

$$\text{Attn}(Q_v, K_a, V_a) = \text{softmax}\left(\frac{Q_v K_a^T}{\sqrt{d_k}}\right) V_a \quad (17)$$

By calculating the attention of visual query and audio key/value, as well as the combination of vision and text, audio and text, etc., the complementary information and association patterns between different modalities are captured. To adaptively distribute the importance of the three modalities, a lightweight gating network is added before the fusion output:

$$[\alpha_v, \alpha_a, \alpha_t] = \text{softmax}\left(W_g[\bar{F}_v; \bar{F}_a; \bar{F}_t] + b_g\right) \quad (18)$$

$$F_{\text{fuse}}(t) = \alpha_v F_v(t) + \alpha_a F_a^{(v)}(t) + \alpha_t F_t^{(v)}(t) \quad (19)$$

In Formulas 18-19, $\bar{F}_v$, $\bar{F}_a$, and $\bar{F}_t$ are the semantic vectors obtained by global averaging of the three modalities; b_g is the parameter of the gating network; α satisfies $\sum \alpha = 1$ after softmax normalization. Finally, the fused feature matrix is obtained:

$$F_{\text{fuse}} = [F_{\text{fuse}}(1), F_{\text{fuse}}(2), \ldots, F_{\text{fuse}}(T)] \in \mathbb{R}^{T \times D} \quad (20)$$

In Formula 20, D is the dimension of the fused feature, which not only retains the deep interaction between the modalities but also realizes dynamic weight allocation through the gating mechanism.

3.3. Cross-modal Consistency Discriminator

To quantify the cross-modal temporal consistency of the fused feature sequence, the fused temporal feature F_{fuse} is bidirectionally aggregated by GRU (Gated Recurrent Unit):

$$\vec{h}_t = \overrightarrow{\text{GRU}}\left(F_{\text{fuse}}(t), \vec{h}_{t-1}\right) \quad (21)$$

$$\overleftarrow{h}_t = \overleftarrow{\text{GRU}}\left(F_{\text{fuse}}(t), \overleftarrow{\text{h}}_{t+1}\right) \quad (22)$$

The hidden states of the forward at the last moment and the backward at the first moment are concatenated to obtain the temporal aggregation vector, which is passed through a two-layer fully connected classifier:

$$z_1 = \text{ReLU}(W_1 h + b_1) \quad (23)$$

$$z_1' = \text{Dropout}(z_1, p = 0.5) \quad (24)$$

$$\ell = W_2 z_1' + b_2 \quad (25)$$

$$p = \text{Softmax}(\ell) \quad (26)$$

In Formulas 23-26, $W_1 \in \mathbb{R}^{256\times512}$; $W_2 \in \mathbb{R}^{2\times512}$; the output p represents the predicted probability of "highly consistent" (label=1) and "inconsistent" (label=0).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

Positive sample (consistent): visual-audio-text fusion features of the same video and the same timestamp.

Negative sample (inconsistent): random pairing across videos or misaligned pairing within the same video (time offset exceeds 0.5 s).

The consistency loss uses binary cross entropy, and the formula is:

$$\mathcal{L}_{\text{cons}} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log\, p_i + (1-y_i)\log\,(1-p_i)] \tag{27}$$

In parallel with the consistency branch, a fake classifier is built on the same trunk, which also consists of two fully connected layers:

$$\ell_{\text{forgery}} = W_4 \text{Dropout}(\text{ReLU}(W_3 h + b_3)) + b_4 \tag{28}$$

$$p_{\text{forgery}} = \text{Softmax}\big(\ell_{\text{forgery}}\big) \tag{29}$$

3.4. Adversarial Training Strategy

To improve the robustness of the model to small perturbations, FGSM is applied to generate adversarial samples at the data level:

$$x' = x + \epsilon \text{sign}\big(\nabla_x \mathcal{L}_{\text{total}}(x, y)\big) \tag{30}$$

After generating adversarial samples, the consistency and fake classification losses are recalculated. Finally, the total loss of the model is the weighted sum of the original sample and adversarial sample losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forgery}} + \lambda \mathcal{L}_{\text{cons}} + \alpha\big(\mathcal{L}^{\text{adv}}_{\text{forgery}} + \lambda \mathcal{L}^{\text{adv}}_{\text{cons}}\big) \tag{31}$$

In Formula 31, $\lambda = 0.5$, and $\alpha = 1.0$. By adding normal and adversarial samples to each training batch and jointly optimizing the above joint loss, it is ensured that the model has good defense capabilities against data-level perturbations while maintaining high classification accuracy.

4. Experimental Design

4.1. Dataset

This study uses the FaceForensics++ [37, 38] dataset, which is built for deep fake video detection tasks [39, 40] and contains four typical fake methods: DeepFake, FaceSwap, Face2Face, and NeuralTextures, as well as the corresponding original real videos. Each fake method has rich samples at different resolutions. This study uses the original resolution version to maximize the retention of fake details and ensure that the feature extraction module can capture high-quality spatio-temporal and texture information.

The dataset image of FaceForensics++ is shown in Figure 2.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

Figure 2.
Data image display.

In the data preprocessing stage, face detection and alignment are performed on each video to ensure that the face in the input frame is in the center of the picture and scaled to 256×256 size. A fixed-interval frame sampling strategy is adopted: one frame is extracted every 10 frames, and 32 key frames are evenly extracted from each video to construct a frame sequence input of uniform length. This sampling method not only takes into account the coverage of the video time series but also controls the amount of calculation to avoid excessive redundancy. The corresponding audio stream is segmented according to the above frame timestamps and aligned with the visual frames one by one to ensure the synchronous extraction of tri-modal features.

To evaluate the model's generalization and stability, all extracted video samples are divided into training set, validation set, and test set in a ratio of 5:1:1, and strictly balanced on the fake category to ensure that each fake type and the original real video are distributed consistently in each subset. Mild data enhancement (random horizontal flip, color jitter) and audio random noise injection are applied into the training set to improve the model's robustness to environmental changes and noise interference. The validation set and the test set only retain the necessary alignment processing and strictly restore the real usage scene, so as to objectively and impartially evaluate the method's performance.

The distribution of the number of images in the dataset is shown in Table 4.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

Table 4.
Distribution of the number of images.

Dataset		DeepFake	FaceSwap	Face2Face	NeuralTextures
Training set	Real	24685	24686	24687	24688
	Fake	25565	25566	25567	25568
Validation set	Real	4937	4937	4937	4937
	Fake	5113	5114	5115	5116
Test set	Real	4937	4937	4937	4937
	Fake	5113	5114	5115	5116
Total		70350	70354	70358	70362

4.2. Experimental Environment and Evaluation Indicators

This experiment is conducted in a single-machine multi-card environment, using mainstream deep learning frameworks and standard training configurations to ensure the reproducibility and comparability of the results. In terms of hardware, 4 NVIDIA Tesla V100s (32 GB of video memory per card) are used, along with Intel Xeon processors and sufficient memory to achieve efficient parallel computing. The software uses Python, using the AdamW optimizer and Cosine Annealing learning rate scheduling to balance convergence speed and generalization ability. The training hyperparameters are set to batch size 16, and a total of 50 epochs are trained, taking into account training time and model performance.

The experimental environment is shown in Table 5.

Table 5.
Experimental environment.

Parameters	Vlaue	Function
Compute device	4×NVIDIA Tesla V100	Supports large-scale parallel tensor operations and model training
Memory	128 GB	Multi-process loading guarantee
Operating system	Ubuntu 20.04	Stable Linux environment, compatible with mainstream deep learning dependencies
Python version	3.8	Supports modern deep learning libraries and tool chains
Optimizer	AdamW	Adaptive learning rate and weight decay to improve convergence and generalization
Learning rate schedule	CosineAnnealing	Smooth annealing learning rate decay to prevent falling into local optimality

In the task of fake video detection [41, 42] the model evaluation index needs to comprehensively measure its classification performance, robustness, and generalization ability. The evaluation indicators include accuracy, precision, recall, ROC curve, and AUC.

Accuracy is used to measure the model's overall classification ability, which is defined as the ratio of the number of samples predicted correctly to the total number of samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (32)$$

Precision reflects how many of the samples predicted as positive by the model are true positive examples, which measures the false positive rate:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (33)$$

Recall measures the ability of the model to detect all true positive examples, that is, how many of all positive samples are successfully identified:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (34)$$

The F1 calculation formula is:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{(Precision+Recall)} \quad (35)$$

The ROC curve depicts the relationship between the true positive rate and the false positive rate at different judgment thresholds:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

$$\text{TPR} = \frac{TP}{TP+FN} \quad (36)$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad (37)$$

AUC refers to the area under the ROC curve, with a range of [0,1], indicating the model's ability to rank positive and negative samples:

$$\text{AUC} = \int_0^1 \text{TPR}(x)dx \quad (38)$$

4.3. Adversarial Test Scenes

To fully verify the model's ability to defend against cross-modal adversarial attacks, two types of attack scenes are designed: white-box attack (the attacker fully understands the model structure and parameters) and black-box attack (the attacker can only access the model input/output). Both types of attacks inject perturbations synchronously into the visual frame color space and the audio spectrogram domain to simulate the threat of multimodal joint attacks.

This study designs a systematic adversarial test scene to comprehensively evaluate the model's robustness. The white-box attack uses two typical methods, FGSM and PGD (Projected Gradient Descent), to synchronously inject perturbations into the visual frame RGB space and the audio spectrogram domain. The FGSM generates single-step adversarial perturbations based on the model gradient, and the attack parameters are set to ε=0.03 (visual frame) and ε=0.015 (audio spectrogram), strictly following the human perception threshold constraint of $L\infty \leq 8/255$; PGD achieves multi-step optimization perturbation enhancement through iterative attacks (10 iterations, step size $\alpha=\varepsilon/4$), focusing on verifying the model's defense capabilities against high-intensity iterative attacks. Both types of attacks cover single-modal (independent visual/audio perturbations) and multimodal joint attacks (visual-audio collaborative perturbations) scenes, and the model defense effect is quantified by the accuracy drop (ΔAcc).

The black-box attack adopts the transfer attack paradigm and selects ResNet18 (visual backbone) and EfficientNet CrossViT (cross-modal fusion model) as alternative models to generate adversarial samples. The attack parameters are consistent with the white-box scene (ε=0.03 visual/ε=0.015 audio).

5. Results

5.1. Comparison with Mainstream Models

The model in this paper is compared with ResNet18, ShuffleNet, MobileNet, Convolutional ViT (Vision Transformer), and EfficientNet CrossViT. The performance of the model is measured by accuracy and number of parameters. The results are shown in Table 6.

Table 6.
Model accuracy and number of parameters.

Model	DeepFake	FaceSwap	Face2Face	NeuralTextures	Parameter quantity (M)
ResNet18	0.978	0.978	0.967	0.877	11.65
ShuffleNet	0.967	0.951	0.934	0.792	1.23
MobileNet	0.964	0.965	0.954	0.845	3.21
Convolutional ViT	0.942	0.867	0.768	0.712	87.98
EfficientNet CrossViT	0.932	0.912	0.945	0.704	101.22
This paper model	0.989	0.988	0.976	0.928	64.12

As can be seen from Table 6, the accuracy of this paper's model on the four types of fakes is significantly ahead of all the comparison models: it reaches 0.989 and 0.988 on DeepFake and FaceSwap, respectively, which are 1.1% and 1.0% higher than ResNet18; it is 0.9% and 5.1% higher on Face2Face and NeuralTextures, respectively. Although the number of parameters of this paper's model (about 64.12M) is higher than that of lightweight networks (such as ShuffleNet's 1.23M and MobileNet's 3.21M), it is still smaller than Convolutional ViT (87.98M) and EfficientNet CrossViT (101.22M). This

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

shows that the multimodal fusion and cross-modal consistency discrimination mechanism can fully exploit the complementary advantages of visual, audio, and text information while controlling the growth of parameters, thereby significantly improving classification accuracy. Especially in NeuralTextures, the most challenging type of fake, traditional pure visual models generally perform poorly (the highest is only 0.877). However, by adding audio and text consistency discrimination and adversarial enhancement, the model in this paper has stronger robustness to subtle fake traces, thus achieving a high accuracy of 0.928.

The parameter-accuracy ratio (Accuracy/Params) is an important indicator for evaluating the practical value of a model. Taking the model in this paper as an example, its parameter volume is only about 63% of EfficientNet CrossViT, but its accuracy is much higher. The reasons are: first, the multimodal feature extraction module (SlowFast, VGGish+Bi-LSTM, Whisper→Transformer) can capture hidden traces of fake videos from different angles; second, the cross-modal self-attention fusion and the adaptive weight allocation of the gating mechanism effectively focus on the most discriminative modal information; third, the additional cross-modal consistency discriminator and FGSM adversarial training strategy enable the model to not only identify fakes in normal modes but also resist small perturbations and noise interference, ensuring robust detection capabilities in a variety of scenes and fake techniques. This combination enables the model in this paper to achieve the best balance between accuracy and parameter quantity, fully demonstrating the potential of multimodal fusion and adversarial enhancement in deep fake video detection.

In the NeuralTextures dataset, the ROC comparison is shown in Figure 3.

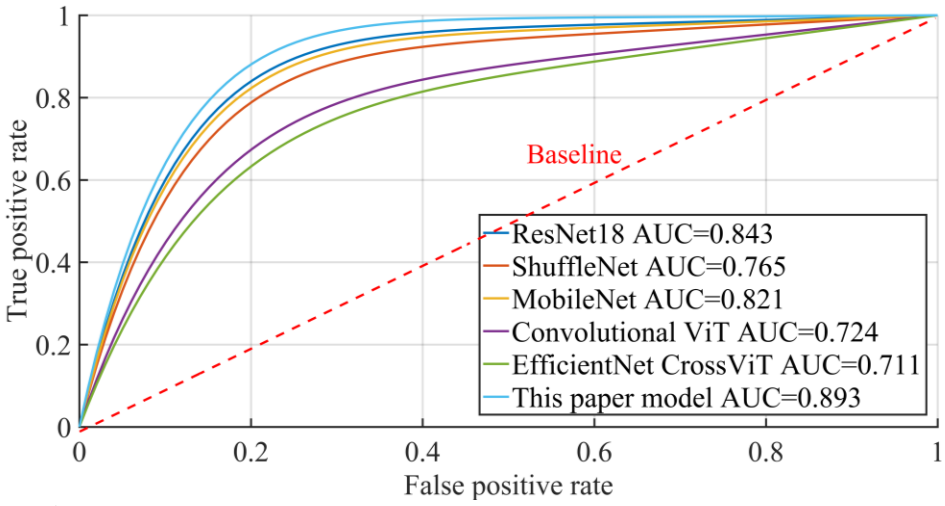

Figure 3.
ROC curve.

The ROC-AUC comparison results on the NeuralTextures subset show that the proposed model leads with a score of 0.893, followed by ResNet18 (0.843), MobileNet (0.821), ShuffleNet (0.765), Convolutional ViT (0.724), and EfficientNet CrossViT (0.711). Although visual backbone networks such as ResNet18 and MobileNet perform well on general fake types, their single frame-level features make it difficult to distinguish the subtle fake traces of NeuralTextures in material details and lighting effects. ShuffleNet is limited by its lightweight design and insufficient capacity, resulting in a significant lack of expression of detailed features. Although the ViT series models have a global attention mechanism, they are prone to texture overfitting due to excessive reliance on visual patterns in the absence of audio and text assistance, resulting in a low recognition rate for complex fake mismatches of NeuralTextures.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

The significant advantage of the proposed model comes from the deep integration of two key factors. Firstly, the cross-modal self-attention fusion mechanism breaks the limitation of single vision and interacts the spatio-temporal visual features captured by SlowFast with the audio context extracted by VGGish+Bi-LSTM and the semantic information encoded by Transformer at the attention level, so that the model can precisely capture the slight asynchrony between the sound rhythm and visual action and the potential contradiction between semantics and expression, so as to more sensitively identify the flaws of NeuralTextures in rendering details and sound-image alignment. Secondly, the FGSM data-level adversarial training strategy effectively constructs difficult samples "near the decision boundary" by injecting small perturbations in the visual frame and audio spectrum domain, so that the model can still maintain stable classification capabilities in the face of slight noise and adversarial interference. The synergy of the two not only improves the discrimination performance of standard fake samples but also significantly enhances the model's generalization and robustness to high-difficulty fake types, thus achieving industry-leading results in the ROC-AUC indicator.

5.2. Noise Robustness

To verify the robustness of the model, Gaussian noise (σ=0.05) and salt and pepper noise (noise density=0.02) are added to the test set, and the accuracy under the influence of noise is statistically analyzed to evaluate the model's stability under noise interference. The noise robustness results are shown in Figure 4.

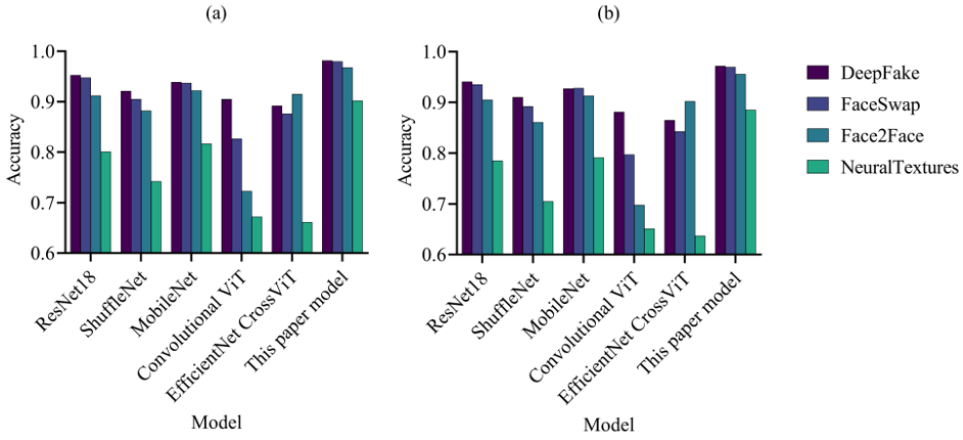

Figure 4.
Noise robustness.

Figure 4 (a). Gaussian noise
Figure 4 (b). Salt and pepper noise

After applying Gaussian noise with σ=0.05, the accuracy of all models decreases to varying degrees, but the degree of decrease is significantly different. The average accuracy of ResNet18 in the four types of fake is 0.904, of which NeuralTextures is greatly affected; ShuffleNet and MobileNet have fewer parameters and are sensitive to noise, and their accuracy drop to 0.742 and 0.817 on NeuralTextures respectively; Convolutional ViT and EfficientNet CrossViT have larger capacity, but because they only rely on visual features, the noise destroys the texture and edge information, resulting in the NeuralTextures retention rate of only 0.672 and 0.661. In contrast, the accuracy of the model in this paper can still be maintained at 0.982/0.980/0.968/0.902 under the same noise conditions, which is higher than all the compared models. The reason is that cross-modal fusion can call on redundant audio and text information to compensate for the loss when the visual flow is disturbed, and the adversarial training strategy enables the model to learn how to resist Gaussian disturbances during training, so

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

that the decision boundary is smoother, and the ability to resist noise interference is significantly enhanced.

In the salt and pepper noise scene with a noise density of 0.02, the discrete bright and dark pixels seriously destroy the local structure, causing the performance of all visual backbone models to further decline. ResNet18 and ShuffleNet can only reach 0.785 and 0.705 on NeuralTextures respectively; MobileNet is slightly better, at 0.791; the ViT series model still performs the worst, at 0.651 and 0.637, respectively. It can be seen that the model that relies solely on visual features and has not undergone adversarial training almost loses its sensitivity to subtle fake traces when facing severe pixel perturbations. The model in this paper still maintains a high accuracy of 0.972/0.970/0.956/0.885 under the same conditions, indicating that the cross-modal consistency discriminator can filter out the interference of isolated noise on the discrimination, and FGSM adversarial training simultaneously injects discrete perturbations in the visual and audio spectrum domains, so that the model has seen similar noises in training and can quickly restore stable predictions. The combination of the two effectively improves the robustness to high-intensity random noise and verifies the superiority of multimodal and adversarial enhancement strategies in actual interference scenes.

5.3. Adversarial Test Results

The adversarial test results are shown in Figure 5.

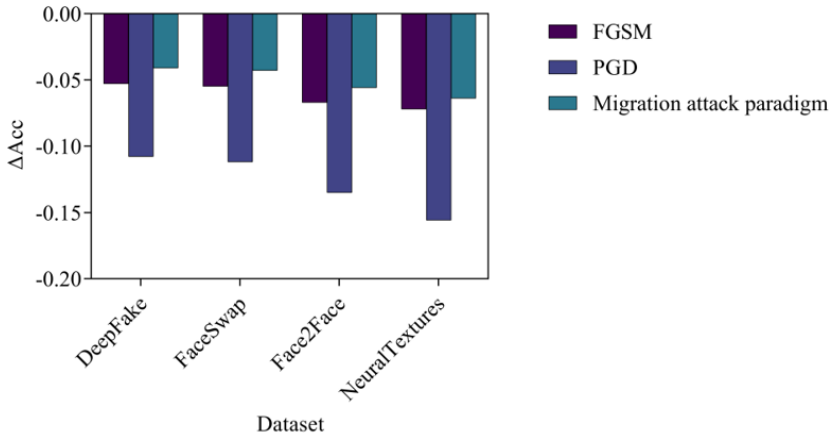

Figure 5.
Adversarial test results.

ΔAcc is calculated as the difference in accuracy after and before the adversarial test. The results in Figure 5 verify the robustness advantage of the model in this paper under cross-modal consistency constraints and adversarial training through adversarial testing. Experimental data shows that in the face of FGSM single-step perturbation in white-box attacks, the ΔAcc of the model on four types of fake videos, DeepFake, FaceSwap, Face2Face, and NeuralTextures, are -0.053, -0.055, -0.067, and -0.072, respectively, indicating that the cross-modal fusion mechanism effectively suppresses the impact of single-modal perturbation on the overall decision. Among them, NeuralTextures has a relatively large absolute value of ΔAcc but is still within an acceptable range because the fake traces are highly dependent on visual-audio synchronization.

PGD iterative attack approaches the optimal perturbation direction through multi-step optimization, resulting in ΔAcc of the four datasets of -0.108, -0.112, -0.135, and -0.156, respectively, but the model still maintains the detection accuracy through the gradient mask defense strategy. Under the transfer attack paradigm, when the adversarial samples generated by ResNet18 and EfficientNet CrossViT attack the model in this paper, ΔAcc only reaches -0.041, -0.043, -0.056, and -0.064,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

respectively, indicating that the decision boundary of the model after adversarial training has stronger smoothness and generalization defense capabilities, and the cross-modal consistency discriminator realizes a multimodal redundancy compensation mechanism through dynamic weight allocation (such as gated weighted fusion). When the visual or audio modality is disturbed, the contribution weight of the undamaged modality is increased, thereby alleviating the impact of single-modal attacks on the overall performance.

5.4. Data Modal

To analyze the impact of data modal, single-modal (visual V, audio A, text T), bi-modal (V+A, V+T, A+T), and tri-modal V+A+T are compared, and the results are shown in Figure 6.

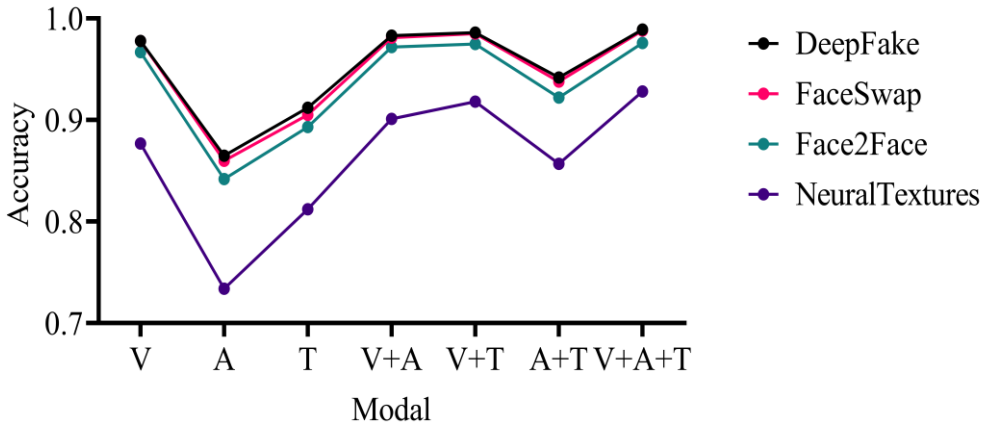

Figure 6.
Data modal analysis results.

From the results in Figure 6, it can be seen that the single-modal vision (V) achieves an accuracy of nearly 0.980 on DeepFake, FaceSwap, and Face2Face, but drops to 0.877 on NeuralTextures, reflecting that a single frame-level visual feature is difficult to capture the details of high-difficulty material rendering fakes. The audio (A) modal only relies on voiceprint and speaking rhythm information, and the detection accuracy for DeepFake and FaceSwap is 0.865 and 0.860, respectively; it is more significant for Face2Face and NeuralTextures, dropping to 0.842 and 0.734, respectively, indicating that although the synchronization errors and acoustic distortions hidden in the audio can assist in detection, the overall discrimination ability is still limited. The text (T) modal is based on ASR→Transformer sentence-level semantic encoding, and the accuracy of DeepFake, FaceSwap, and Face2Face is improved to 0.912, 0.905, and 0.893, reflecting that semantic clues have complementary value in detecting speech content fake and lip movement mismatch, but it is still only 0.812 on NeuralTextures.

After the application of bi-modal fusion, the performance is significantly improved: vision + audio (V + A) jumps to 0.901 on NeuralTextures; vision + text (V + T) further reaches 0.918, indicating that the alignment of text semantics and visual expressions can more accurately capture subtle inconsistencies in deep fakes; although audio + text (A + T) is not as good as the combination with vision, it still reaches 0.942 on DeepFake and 0.857 on NeuralTextures, which exceeds the linear sum of single audio and text, verifying the effectiveness of cross-modal information complementarity. In general, bi-modal fusion can approach the overall performance on most fake types, but a tri-modal is still needed to achieve a near-perfect detection rate.

The tri-modal (V + A + T) model achieves the highest accuracy on all fake types: DeepFake 0.989, FaceSwap 0.988, Face2Face 0.976, NeuralTextures 0.928. This improvement is attributed to the deep

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

coordination of the three-stream information at two levels: cross-modal self-attention fusion can adaptively adjust the contribution of vision, audio, and text according to different video scenes through the Query–Key–Value mechanism and dynamic weight allocation. For high-difficulty fakes such as NeuralTextures, the subtle rendering artifacts captured by the visual stream are often accompanied by a slight synchronization deviation between the audio and text. The fusion module strengthens the consistency detection of audio and text at this moment, so that accurate judgment can still be made when the visual is difficult to distinguish.

Adversarial training makes the model stable under disturbances such as Gaussian noise and FGSM, further strengthening the robust fusion of tri-modal information. Adversarial training not only simulates the most confusing fake and noise interference in the training stage but also forces the model to build a smoother decision boundary, strengthening the compensation effect of weak modal branches in strong noise scenes. For example, when the visual frame is damaged by noise, the audio and text branches can continue to maintain high discrimination ability through pre-learned language-rhythm correspondence and semantic-expression mapping; vice versa. It can be seen that the dual mechanism of tri-modal deep fusion and adversarial training significantly improves the generalization and robustness of the model in various fake scenes, providing solid technical support for high-precision and practical deep fake video detection.

5.5. Ablation Experiment

The ablation experiment is set up in the NeuralTextures dataset to analyze the impact of each part of the model in this paper on the performance, including AT (adversarial training) and CD (consistency discriminator). The ablation experiment results are shown in Figure 7.

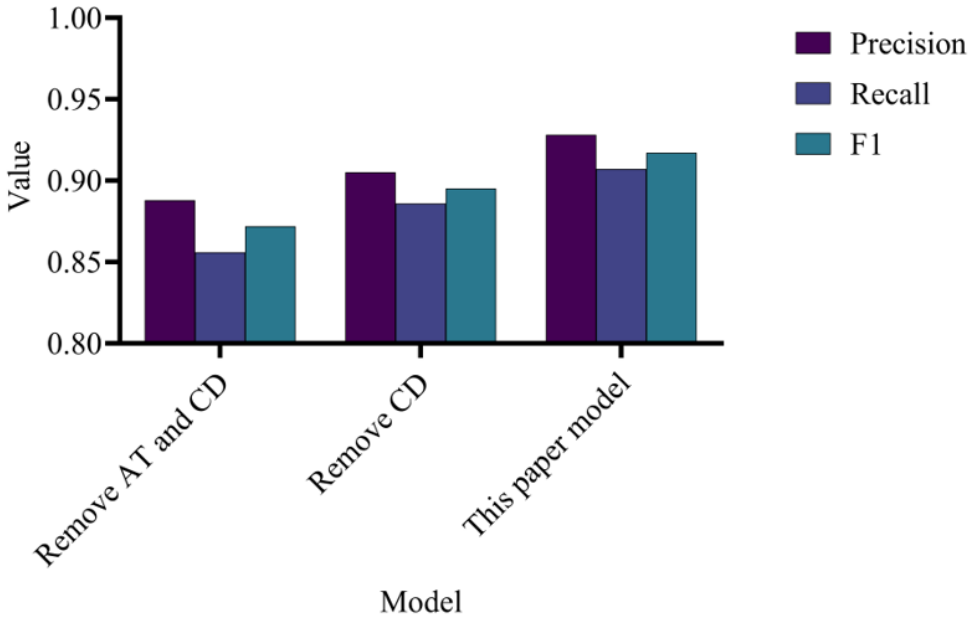

Figure 7.
Ablation experiment results.

From the ablation experiment results, after removing the consistency discriminator, the model precision drops from 0.928 to 0.905; the recall drops from 0.907 to 0.886; the F1 score drops from 0.917 to 0.895. This shows that the consistency discriminator plays a key role in reducing the false positive rate: through the binary classification constraints of "high consistency" and "inconsistency", the model can more accurately identify the fake traces caused by abnormal cross-modal alignment, thereby

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

significantly improving the precision. The slight decrease in recall shows that although the consistency discriminator enhances the model's sensitivity to cross-modal differences, in some borderline cases, a small number of real fake videos may be misclassified as inconsistent, resulting in a slight decrease in recall. Overall, the addition of the consistency discriminator enables the model to effectively compress false positives while maintaining high recall, which is of great significance to improving detection reliability.

Furthermore, when both adversarial training and the consistency discriminator are removed at the same time, the precision drops further to 0.888; the recall drops to 0.856; the F1 score drops to 0.872. This comparison highlights the core value of FGSM data-level adversarial training in strengthening the model's ability to resist disturbances: adversarial training enables the model to learn to deal with tiny noise and adversarial perturbations by constructing difficult samples that approximate the decision boundary, thereby maintaining higher recall and precision when facing detail fakes such as NeuralTextures. Compared with the F1 of 0.917 of the complete model, the F1 of the model without adversarial training drops significantly, indicating that adversarial enhancement is crucial to maintaining high overall detection performance. In addition, the synergistic effect of the consistency discriminator and adversarial training is particularly significant: the former provides structured constraints on cross-modal consistency metrics, and the latter approaches the optimal discrimination boundary under data-level perturbations. The combined effect of the two enables the model to achieve the highest precision and recall on NeuralTextures, reaching 0.928/0.907, verifying the effectiveness of the joint design of multimodal fusion and adversarial enhancement.

6. Conclusions

This study proposes a deep fake video detection framework that integrates visual, audio, and text tri-modal features with adversarial training, achieving industry-leading detection performance (maximum accuracy 0.989) on the FaceForensics++ dataset, and significantly improving the model's sensitivity to cross-modal temporal consistency anomalies through cross-modal self-attention mechanism and dual-branch discriminator. Its core contribution lies in the systematic integration of multimodal deep feature interaction and adversarial defense mechanism, solving the performance bottleneck of traditional single-modal methods under high-quality fake and noise interference, and providing technical support for content review and public opinion security on social platforms. However, the research is still limited by specific fake type datasets (such as the lack of samples generated by the latest diffusion model) and high computational costs. Future can explore lightweight multimodal fusion architectures, apply more modals (such as physiological signals), and optimize detection strategies for new generation technologies to cope with the rapidly evolving threat of deep fakes and promote the practical and universal development of detection technology.

Funding:

This work was supported by Teaching Reform and Innovation Program in Shanxi Province, China (Grant Number: J20241265).

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

References

[1] S. Suratkar and F. Kazi, "Deep fake video detection using transfer learning approach," *Arabian Journal for Science and Engineering*, vol. 48, no. 8, pp. 9727-9737, 2023. https://doi.org/10.1007/s13369-022-07321-3

[2] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, "Regulating deep fakes: Legal and ethical considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24-31, 2020. https://doi.org/10.1093/jiplp/jpz167

[3] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089-1102, 2021. https://doi.org/10.1109/TCSVT.2021.3074259

[4] U. Kosarkar and G. Sakarkar, "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis," *Multimedia Tools and Applications*, vol. 84, no. 7, pp. 3841-3857, 2025. https://doi.org/10.1007/s11042-025-15729-4

[5] Z. Zhu, H. Qiu, C. Yang, and Y. Wang, "A method for determining speech lip-movement consistency based on analysis of specific vowel pronunciation events," *Journal of South China University of Technology (Natural Science Edition*, vol. 48, no. 1, pp. 139-139, 2020. https://doi.org/10.13298/j.cnki.jscut.2020.01.021

[6] Z. Zhu, C. Luo, Q. He, W. Peng, Z. Mao, and S. Zhang, "Multi-view lip-sound consistency discrimination based on lip reconstruction and 3D coupled CNN," *Journal of South China University of Technology (Natural Science Edition)*, vol. 51, no. 5, pp. 70-70, 2023. https://doi.org/10.13298/j.cnki.jscut.2023.05.016

[7] L. Sun, H. Zhang, X. Mao, S. Guo, and Y. Hu, "Strongly compressed deep fake video detection based on super-resolution reconstruction," *Journal of Electronics & Information Technology*, vol. 43, no. 10, pp. 2967-2975, 2021. https://doi.org/10.11999/JEIT200531

[8] Y. Han, G. Hua, and H. Zhang, "Collaborative detection of face-swapping in videos based on eye and mouth regions based on Inception3D network," *Journal of Signal Processing*, vol. 37, no. 4, pp. 567-567, 2021. https://doi.org/10.16798/j.issn.1003-0530.2021.04.010

[9] S. Yang, J. Wang, Y. Sun, and J. TANG, "Fake face detection based on global consistency of multi-level features," *Journal of Image and Graphics*, vol. 27, no. 09, pp. 2708-2720, 2022.

[10] B. Zhang, C. Zhu, Q. Yin, J. Fu, L. Liu, and J. Liu, "Fake face detection based on noise attention," *Chinese Journal of Network & Information Security*, vol. 9, no. 4, pp. 155-155, 2023.

[11] S. T. Ikram, S. Chambial, and D. Sood, "A performance enhancement of deepfake video detection through the use of a hybrid CNN Deep learning model," *International journal of electrical and computer engineering systems*, vol. 14, no. 2, pp. 169-178, 2023. https://doi.org/10.32985/ijeces.14.2.6

[12] S. Tambe, A. Pawar, and S. Yadav, "Deep fake videos identification using ANN and LSTM," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 24, no. 8, pp. 2353-2364, 2021. https://doi.org/10.1080/09720529.2021.2014140

[13] S. Borade *et al.*, "ResNet50 deepfake detector: Unmasking reality," *Indian Journal of Science and Technology*, vol. 17, no. 13, pp. 1263-1271, 2024. https://doi.org/10.20852/ijisae.4782

[14] V. Rajakumareswaran, S. Raguvaran, V. Chandrasekar, S. Rajkumar, and V. Arun, "DeepFake detection using transfer learning-based Xception model," *Advanced Information Systems*, vol. 8, no. 2, pp. 89-98, 2024. https://doi.org/10.20998/2522-9052.2024.2.10

[15] R. Zhou, C. Jiang, Q. Xu, Y. Li, C. Zhang, and Y. Song, "Text-to-video synthesis method based on multi-conditional generative adversarial networks," *Journal of Computer-Aided Design & Computer Graphics/Jisuanji Fuzhu Sheji Yu Tuxingxue Xuebao*, vol. 34, no. 10, pp. 1567-1567, 2022. https://doi.org/10.3724/SP.J.1089.2022.19731

[16] J. Lin and B. Yang, "From perception to creation: Frontier research on image and video generative methods," *Acta Optica Sinica* vol. 43, no. 15, pp. 1510002-1510002, 2023.

[17] F. Guan, H. Zhang, S. Lu, H. Lai, X. Du, and Y. Zheng, "Research status of diffusion model in computer vision," *CAAI Transactions on Intelligent Systems*, vol. 20, no. 2, pp. 265-282, 2025.

[18] S. Tipper, H. F. Atlam, and H. S. Lallie, "An investigation into the utilisation of cnn with lstm for video deepfake detection," *Applied Sciences*, vol. 14, no. 21, p. 9754, 2024.

[19] U. Masud, M. Sadiq, S. Masood, M. Ahmad, and A. A. Abd El-Latif, "LW-DeepFakeNet: A lightweight time distributed CNN-LSTM network for real-time DeepFake video detection," *Signal, Image and Video Processing*, vol. 17, no. 8, pp. 4029-4037, 2023.

[20] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, "Voice-face homogeneity tells deepfake," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1-22, 2023.

[21] D. Salvi *et al.*, "A robust approach to multimodal deepfake detection," *Journal of Imaging*, vol. 9, no. 6, p. 122, 2023.

[22] G. Zhang, M. Gao, Q. Li, W. Zhai, and G. Jeon, "Multi-modal generative deepfake detection via visual-language pretraining with gate fusion for cognitive computation," *Cognitive Computation*, vol. 16, no. 6, pp. 2953-2966, 2024.

[23] Y. YUAN, L.-q. HUANG, F. YE, T.-q. HUANG, H.-f. LUO, and C. XU, "A real-time facial manipulation video detection model based on ensemble learning dual-stream neural network," *Computer Engineering & Science*, vol. 45, no. 03, p. 470, 2023.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119

[24] X. Li and K. Yu, "A deepfakes detection technology based on two-stream network," *Journal of Cyber Security*, vol. 5, no. 2, pp. 84-91, 2020.
[25] S. Tyagi and D. Yadav, "A detailed analysis of image and video forgery detection techniques," *The Visual Computer*, vol. 39, no. 3, pp. 813-833, 2023.
[26] S. Mohiuddin, S. Malakar, M. Kumar, and R. Sarkar, "A comprehensive survey on state-of-the-art video forgery detection techniques," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 33499-33539, 2023.
[27] W. El-Shafai, M. A. Fouda, E.-S. M. El-Rabaie, and N. A. El-Salam, "A comprehensive taxonomy on multimedia video forgery detection techniques: Challenges and novel trends," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4241-4307, 2024.
[28] Y. Wang, C. Peng, D. Liu, N. Wang, and X. Gao, "Spatial-temporal frequency forgery clue for video forgery detection in VIS and NIR scenario," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7943-7956, 2023.
[29] Q. Bao, Y. Wang, H. Hua, K. Dong, and F. Lee, "An anti-forensics video forgery detection method based on noise transfer matrix analysis," *Sensors*, vol. 24, no. 16, p. 5341, 2024.
[30] G. Yang, K. Xu, X. Fang, and J. Zhang, "Video face forgery detection via facial motion-assisted capturing dense optical flow truncation," *The Visual Computer*, vol. 39, no. 11, pp. 5589-5608, 2023.
[31] N. A. Shelke and S. S. Kasana, "Multiple forgery detection in digital video with VGG-16-based deep neural network and KPCA," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5415-5435, 2024.
[32] H. D. Panchal and H. B. Shah, "Multiple forgery detection in digital video based on inconsistency in video quality assessment attributes," *Multimedia Systems*, vol. 29, no. 4, pp. 2439-2454, 2023.
[33] Q. Li, R. Wang, and D. Xu, "A video splicing forgery detection and localization algorithm based on sensor pattern noise," *Electronics*, vol. 12, no. 6, p. 1362, 2023.
[34] G. Singh and K. Singh, "Copy-move video forgery detection techniques: A systematic survey with comparisons, challenges and future directions," *Wireless Personal Communications*, vol. 134, no. 3, pp. 1863-1913, 2024.
[35] X. Li, S. Ji, C. Wu, Z. Liu, S. Deng, and P. Cheng, "A review of deepfakes and detection technologies," *Journal of Software*, vol. 32, no. 2, pp. 496-518, 2021.
[36] W. Zhou, Z. Weiming, and Y. Nenghai, "An overviw of Deepfake forgery and defense techniques," *Journal of Signal Processing*, vol. 37, no. 12, pp. 2338-2355, 2021.
[37] Y. Zhang, G. Li, Y. Cao, and X. Zhao, "A face tampering video detection method based on inter-frame difference," *Journal of Cyber Security*, vol. 5, no. 2, pp. 49-72, 2020.
[38] J. Li, B. Li, and W. Lin, "AdfNet: An adaptive deepfake detection network based on diversified features," *Journal of South China University of Technology (Natural Science Edition)*, vol. 51, no. 9, pp. 82-82, 2023.
[39] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *Let Biometrics*, vol. 10, no. 6, pp. 607-624, 2021.
[40] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: Challenges and opportunities," *Artificial Intelligence Review*, vol. 57, no. 6, p. 159, 2024.
[41] H. Wang, Z. Liu, and S. Wang, "Exploiting complementary dynamic incoherence for deepfake video detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4027-4040, 2023.
[42] A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "DeepFake video detection: A time-distributed approach," *SN Computer Science*, vol. 1, no. 4, p. 212, 2020.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 1342-1359, 2025
DOI: 10.55214/25768484.v9i6.8119