

Article

SDEQ-Net: A Deepfake Video Anomaly Detection Method Integrating Stochastic Differential Equations and Hermitian-Symmetric Quantum Representations

Ruixing Zhang, Bin Li and Degang Xu *

College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China; 2024910062@stu.haut.edu.cn (R.Z.); lbgis@haut.edu.cn (B.L.)

* Correspondence: xudegang@haut.edu.cn

Abstract

With the rapid advancement of deepfake generation technologies, forged videos have become increasingly realistic in visual quality and temporal consistency, posing serious threats to multimedia security. Existing detection methods often struggle to effectively model temporal dynamics and capture subtle inter-frame anomalies. To address these challenges, we propose a Stochastic Differential Equation and Quantum Uncertainty Network (SDEQ-Net), a novel deepfake video anomaly detection framework that integrates continuous time stochastic modeling with quantum uncertainty mechanisms. First, a Continuous Time Neural Stochastic Differential Filtering Module (CNSDFM) is introduced to characterize the continuous evolution of latent inter-frame states using neural stochastic differential equations, enabling robust temporal filtering and uncertainty estimation. Second, a Quantum Uncertainty Aware Fusion Module (QUAFM) incorporates Hermitian-symmetric density matrix representations and von Neumann entropy to enhance feature fusion under uncertainty, leveraging the mathematical symmetry properties of quantum state representations for principled uncertainty quantification. Third, a Fractional Order Temporal Anomaly Detection Module (FOTADM) is proposed to generate fine grained temporal anomaly scores based on fractional order residuals, which are used as dynamic weights to guide attention toward anomalous frames. Extensive experiments on three benchmark datasets, including FaceForensics++, Celeb-DF, and DFDC, demonstrate the effectiveness of the proposed method. SDEQ-Net achieves AUC scores of 99.81% on FF++ (c23) and 97.91% on FF++ (c40). In cross dataset evaluations, it obtains 89.55% AUC on Celeb-DF and 86.21% AUC on DFDC, consistently outperforming existing state-of-the-art methods in both detection accuracy and generalization capability.

Keywords: deepfake detection; stochastic differential equations; quantum uncertainty; continuous-time modeling; Hermitian symmetry

Received: 11 December 2025

Revised: 12 January 2026

Accepted: 14 January 2026

Published: 30 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, with the rapid development of deep learning technologies such as Generative Adversarial Networks (GANs) and diffusion models, deepfake techniques have made significant progress [1,2]. These techniques are capable of generating highly realistic forged videos, which have greatly facilitated advancements in fields such as film production, virtual reality, and digital entertainment. However, the negative impacts of

deepfake technology have also become increasingly prominent. Forged videos are frequently exploited for malicious purposes including disinformation dissemination, defamation, and political manipulation, thereby posing serious threats to public safety and the trust infrastructure of modern society [3]. A notable example occurred during the recent Russia–Ukraine conflict, in which a video depicting President Zelensky allegedly urging soldiers to lay down their arms was circulated online and later confirmed to be a deepfake—marking the first known application of deepfake technology in a military context [4]. Consequently, the development of efficient, robust, and generalizable deepfake video detection methods has emerged as a critical research focus in the fields of computer vision and AI security.

Current deepfake video detection methods face substantial challenges, primarily due to the nature of the forged content itself. Generated by powerful deep learning models, deepfake content often exhibits high fidelity in visual details and structural features, rendering it nearly indistinguishable from real content. This significantly undermines the effectiveness of traditional feature extraction and video analysis techniques. To address these issues, a wide range of deepfake detection approaches have been proposed in recent years, primarily falling into two major categories. The first category focuses on temporal consistency modeling, which aims to capture subtle temporal discontinuities or anomalies by analyzing the dynamic evolution between consecutive video frames [5–11]. The second category emphasizes intra-frame forensic analysis, which seeks to detect local artifacts by examining spatial cues such as texture irregularities, frequency domain distributions, and illumination inconsistencies within individual frames [12–16]. These two categories represent the most prominent directions in current deepfake video detection research, each addressing the problem from a complementary perspective: temporal modeling and spatial feature analysis.

However, as deepfake generation techniques continue to evolve, existing detection methods are increasingly challenged in complex and dynamic real-world scenarios. Inter-frame approaches may fail to capture meaningful temporal anomalies in low-motion or static scenes, while intra-frame methods, although effective in modeling local textures and structural inconsistencies, often struggle to distinguish high-fidelity forged content from genuine images and tend to overlook latent temporal dependencies across frames. Consequently, developing a unified framework that effectively integrates spatial perception with temporal modeling remains a central challenge in deepfake detection.

Meanwhile, uncertainty quantification has emerged as a fundamental research direction for improving the reliability and robustness of deep learning systems. Existing studies have systematically categorized uncertainty estimation methods and established the distinction between data-related and model-related uncertainty [17,18], while practical Bayesian learning techniques such as Monte Carlo Dropout have demonstrated that uncertainty can be effectively extracted from standard neural networks [19]. In parallel, recent advances indicate that continuous-time stochastic modeling provides a principled framework for capturing non-stationary temporal dynamics [20], quantum-inspired entropy measures such as von Neumann entropy offer advantages in representation manipulation [21], and fractional-order operators are effective in detecting anomalies with long-range dependencies due to their inherent non-local memory property [22]. Collectively, these findings motivate the development of a unified uncertainty-aware spatiotemporal modeling framework for robust deepfake detection.

To address the above challenges, we propose a novel deepfake video anomaly detection framework, termed SDEQ-Net, which integrates stochastic differential equations with quantum-based uncertainty modeling. The proposed method is designed to jointly capture spatial texture cues and inter-frame dynamic evolution patterns. SDEQ-Net com-

prises three key modules: a Continuous-time Neural Stochastic Differential Filtering Module (CNSDFM) for modeling temporally continuous frame transitions; a Quantum Uncertainty-Aware Fusion Module (QUAFM) for uncertainty-aware feature integration; and a Fractional-Order Temporal Anomaly Detection Module (FOTADM) for frame-level anomaly localization. To verify the effectiveness of the proposed framework, extensive experiments are conducted on three widely used deepfake video datasets, including FaceForensics++, Celeb-DF, and DFDC. Experimental results demonstrate that SDEQ-Net significantly outperforms existing state-of-the-art methods across various metrics and exhibits superior generalization capability in cross-dataset evaluations.

The main contributions of this study are summarized as follows:

- We develop a novel CNSDFM that models the continuous evolution of video frame states using neural stochastic differential equations. By combining LSTM-based temporal encoding with gain filtering, the module enables dynamic modeling and temporal uncertainty estimation.
- We propose a QUAFM that leverages density matrix representations and von Neumann entropy to quantify and fuse uncertainty-aware features, thereby improving the stability and robustness of uncertain region identification.
- We design a FOTADM that introduces a residual-driven, fractional-order scoring mechanism for fine-grained frame-level anomaly detection, guiding the model to adaptively focus on anomalous dynamic patterns.
- We perform comprehensive evaluations on three benchmark datasets (FF++, Celeb-DF, and DFDC), demonstrating the superior performance and generalizability of the proposed method compared to mainstream baselines.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed methodology and component modules. Section 4 presents experimental settings, evaluation metrics, and results. Section 5 concludes the study and outlines future research directions.

2. Related Work

This section first reviews the mainstream techniques for facial forgery generation, including methods based on Autoencoders, Generative Adversarial Networks (GANs), Diffusion Models, and Transformer-based architectures. Subsequently, we discuss existing detection strategies for facial forgery, which can be broadly categorized into intra-frame-based and inter-frame temporal-based detection approaches.

2.1. Deepfake Generation Techniques

Deepfake generation has evolved through several paradigms. Early approaches employed autoencoders to learn latent facial representations for face swapping [23,24]. Generative Adversarial Networks (GANs) subsequently became dominant, with variants such as StyleGAN producing highly realistic results [25,26]. More recently, diffusion models have emerged as powerful alternatives, achieving state-of-the-art image quality through iterative denoising [27,28]. Transformer-based architectures have also been explored for their ability to model long-range dependencies [29,30]. These advances have significantly increased the realism of synthetic content, posing substantial challenges for detection methods.

2.2. Deepfake Detection Methods

With the rapid evolution of deepfake generation techniques, the challenge of accurately detecting manipulated content has become increasingly urgent. Deepfake detection

methods primarily aim to determine the authenticity of visual content by identifying abnormal patterns or inconsistencies within images or videos. Based on the processing strategy, existing detection approaches can be broadly divided into two categories:

(1) Inter-frame-based deepfake detection. Real video sequences typically exhibit smooth and coherent temporal transitions, while manipulated videos often contain abrupt or inconsistent temporal dynamics. Therefore, detecting subtle discrepancies between consecutive frames has proven effective for identifying forgeries. For instance, the work in [8] utilized optical flow estimation to capture latent differences across frames, which were subsequently fed into a Convolutional Neural Network (CNN) for classification. In [31], a Gated Recurrent Unit (GRU) was integrated after the CNN to incorporate temporal dependencies for improved distinction of frame-level inconsistencies. The study in [10] employed 3D Convolutional Neural Networks (3D CNNs) to jointly model spatial artifacts and temporal motion inconsistencies, enabling comprehensive spatiotemporal detection. To further enhance performance, several works have introduced domain-specific priors, such as eye blinking patterns [32], lip movements [11], and emotional expressions [33], as discriminative features. In addition, recent approaches have focused on temporal representation learning. For example, [34] proposed a framework that explicitly learns spatiotemporal inconsistencies for deepfake detection. The work in [35] adopted a self-supervised cross-modal learning strategy by leveraging the natural synchrony between visual and auditory modalities in real videos to learn temporally dense representations. Moreover, [5] introduced a self-supervised framework that learns consistency representations from authentic facial videos, aiming to improve the generalization and robustness of detection models against various types of manipulations.

(2) Intra-frame-based deepfake detection. Intra-frame detection methods focus on analyzing the content of individual video frames, aiming to identify forgeries by extracting local image features such as texture artifacts, illumination inconsistencies, and frequency domain cues. The study in ref. [36] ignored fine-grained facial features and instead concentrated on detecting global artifacts for deepfake identification. In ref. [37], a method was proposed to estimate 3D head pose from facial images, where an angular deviation beyond a certain threshold between two directional vectors was considered indicative of forgery. The work in ref. [38] extracted and fused features from both YCbCr and RGB color spaces to improve detection accuracy. In ref. [39], original images were partitioned into equal-sized blocks prior to spatial feature extraction, encouraging the model to explore more discriminative tampering traces. In ref. [40], a deepfake detection framework with adaptively weighted multi-scale attention features was introduced to enhance feature representation. To improve generalization, [41] utilized unsupervised domain adaptation to bridge the distribution gap between real and fake samples. To further boost detection performance, several studies have incorporated frequency domain information. The work in ref. [42] analyzed both global and local temporal inconsistencies from spatial and frequency perspectives, proposing a locally frequency-guided dynamic inconsistency network. In ref. [43], a frequency-aware learning strategy was introduced to enhance generalization performance by focusing on frequency domain patterns. In ref. [44], a spatial-phase shallow learning method was developed by integrating spatial image content with phase spectrum features, effectively capturing upsampling artifacts in manipulated faces. The study in ref. [45] employed the Laplacian of Gaussian (LoG) filter for frequency-based enhancement, aiming to suppress low-level image content while amplifying high-frequency forgery cues.

In summary, existing detection methods predominantly focus on either temporal or spatial features in isolation. Inter-frame approaches may fail in low-motion scenarios, while intra-frame methods struggle against high-fidelity forgeries. These limitations motivate the development of unified frameworks that effectively integrate both perspectives.

3. Proposed Method

This section introduces SDEQ-Net, a unified framework that integrates neural stochastic differential modeling with quantum-inspired uncertainty fusion and fractional-order temporal analysis for deepfake video detection. The key motivation is to jointly capture temporal uncertainty at complementary scales, including instantaneous dynamics, aggregated temporal correlations, and long-range dependency patterns, which has been shown to be more effective than single-paradigm approaches [46]. SDEQ-Net adopts a pre-trained ResNet3D-50 backbone for spatiotemporal feature extraction. While ResNet3D-50 effectively captures local spatiotemporal patterns through 3D convolutions, its temporal receptive field is inherently limited to short clip-level windows (typically 16–32 frames). To model long-range temporal dependencies and continuous-time dynamics across entire video sequences, SDEQ-Net incorporates LSTM-based recurrent processing within the CNSDFM. This hierarchical architecture separates local and global temporal modeling: the 3D CNN backbone extracts clip-level spatiotemporal representations, while the LSTM captures sequence-level temporal evolution and uncertainty propagation [47].

The proposed framework comprises three specialized modules organized in a coordinated pipeline, as illustrated in Figure 1: (1) CNSDFM receives the backbone features $\mathbf{Z} \in \mathbb{R}^{B \times C \times T}$ and outputs refined temporal representations F along with frame-level uncertainty estimates U ; (2) QUAFM takes both F and U as inputs, constructing density matrices to perform uncertainty-aware feature fusion; and (3) FOTADM computes frame-wise anomaly scores A from the residuals between original and refined features. During training, the anomaly scores serve as dynamic weights in the loss function (Equation (17)). During inference, the fused features from QUAFM are concatenated with global features from the backbone and passed through fully connected layers for final classification.

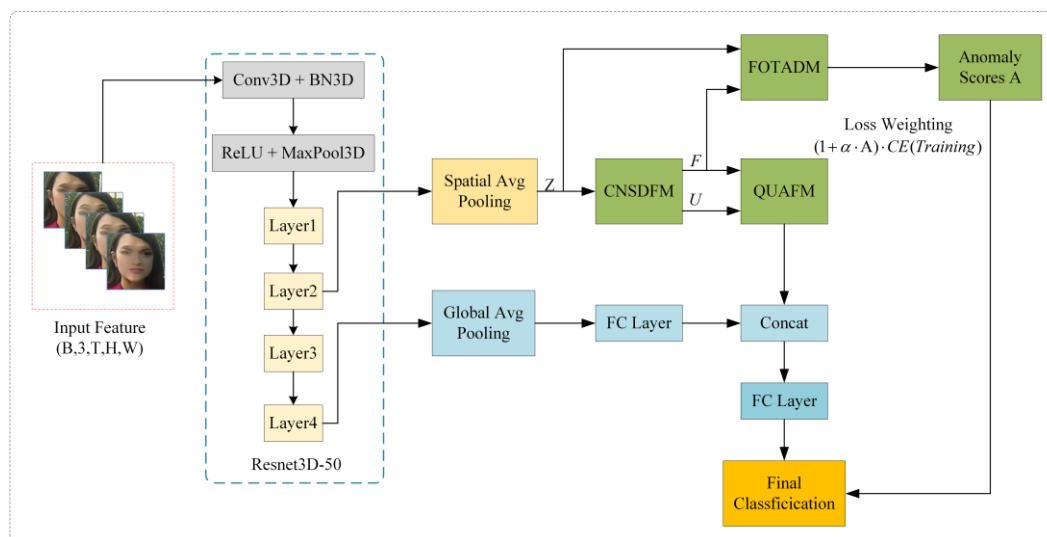


Figure 1. The overall architecture of the proposed SDEQ-Net framework. The coordinated pipeline integrates continuous-time stochastic modeling, quantum uncertainty fusion, and fractional-order anomaly detection to jointly capture temporal dynamics and spatial artifacts for robust deepfake detection.

3.1. Continuous-Time Neural Stochastic Differential Filter Module

Deepfake videos often exhibit subtle temporal inconsistencies that manifest as continuous, non-stationary dynamic variations during inter-frame propagation. To model the evolution of latent inter-frame states and their associated uncertainties, we propose the Continuous-time Neural Stochastic Differential Filter Module (CNSDFM). Unlike standard LSTM or RNN-based temporal models that operate on discrete time steps and lack

principled uncertainty modeling, Neural SDEs enable continuous-time latent dynamics with mathematically grounded stochastic components [20]. CNSDFM adopts the Itô integral formulation compatible with the Euler–Maruyama discretization scheme [48], and integrates LSTM with an adaptive gain mechanism. This hybrid architecture leverages the complementary strengths of both paradigms rather than being a simple enhancement of either [49]. The overall workflow is illustrated in Figure 2.

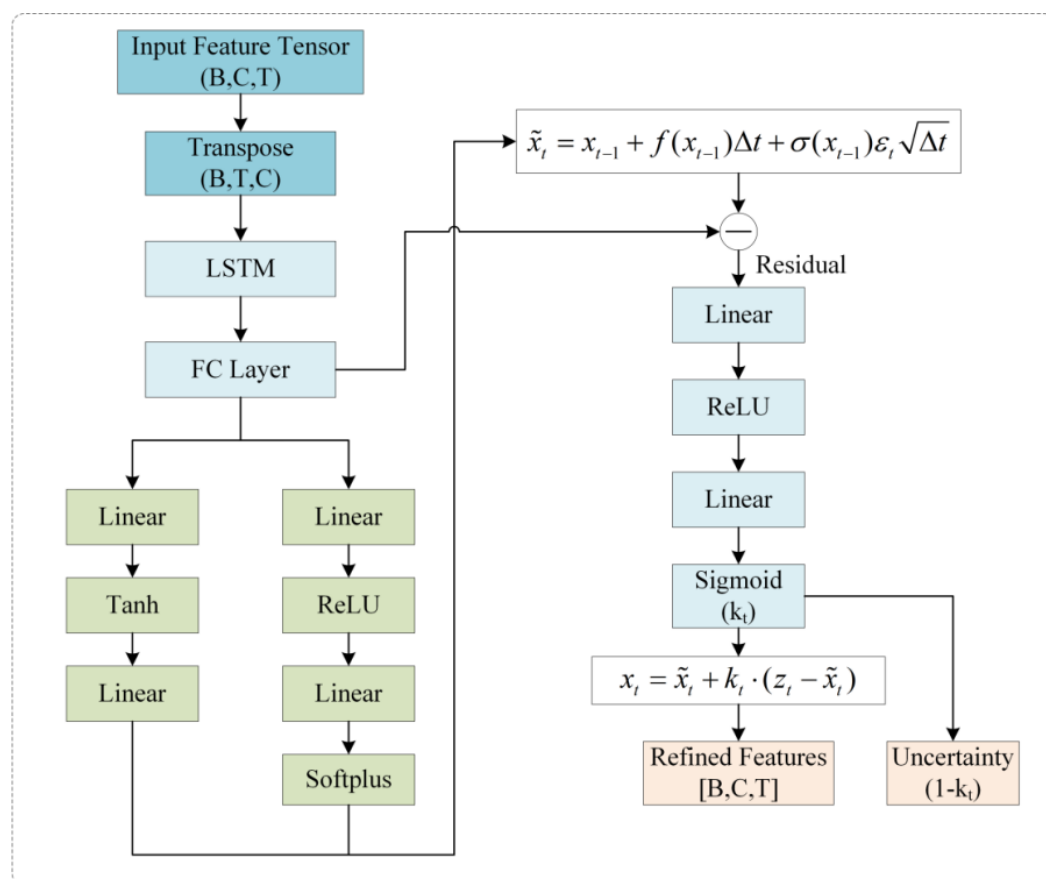


Figure 2. The workflow of CNSDFM. This module enables continuous-time latent state evolution with uncertainty estimation, effectively capturing non-stationary temporal inconsistencies that discrete-time models often miss.

Let the high-level semantic feature representation of the input video frame sequence be denoted by the tensor $Z \in \mathbb{R}^{B \times C \times T}$, where B is the batch size, C is the feature dimension, and T is the number of temporal steps. This tensor is generated by a pretrained visual encoder and serves as the observation input for the CNSDFM.

To capture long-range temporal dependencies, the tensor Z is transposed to shape $[B, T, C]$ and fed into an LSTM network, producing a sequence of hidden states $h_t \in \mathbb{R}^H, t = 1, 2, \dots, T$. These hidden states are subsequently mapped back to the feature space via a fully connected transformation, yielding the prior prediction $\hat{x}_t = f_{\text{pred}}(h_t) \in \mathbb{R}^C$ of the current latent state.

To model the continuous temporal evolution of the latent state, we introduce a Neural Stochastic Differential Equation (Neural SDE) defined as follows:

$$dx_t = f(x_t)dt + \sigma(x_t)dW_t \quad (1)$$

where $f(\cdot)$ represents the drift function, $\sigma(\cdot)$ denotes the diffusion term, and W_t is a standard Brownian motion under the Itô interpretation.

In practice, we approximate this continuous evolution using the Euler–Maruyama scheme in discrete time. Let the current time step be t and the step size be Δt , and the discretized state transition can be approximated by:

$$\tilde{x}_t = x_{t-1} + f(x_{t-1})\Delta t + \sigma(x_{t-1})\varepsilon_t\sqrt{\Delta t} \quad (2)$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$ is a standard Gaussian noise term. This continuous-time modeling process effectively captures non-stationary phenomena such as stochastic jumps and non-linear drifts that may occur in deepfake regions. As illustrated in Figure 2, the drift function employs tanh activation to produce bounded outputs, while the diffusion function uses ReLU to ensure non-negative coefficients, following standard practices in neural SDE implementations [48,49].

To address the uncertainty inherent in video frame observations, we further introduce an adaptive Kalman-inspired gain mechanism to dynamically balance the contributions from observation and prediction. Although the proposed update rule resembles the measurement update of a Kalman filter, it does not constitute a formal Kalman filter, as no explicit covariance propagation or Bayesian optimality assumptions are imposed. Instead, the design is motivated by the correction principle of Kalman filtering, where a residual between prediction and observation is used to adaptively refine the latent state. The observation residual is defined as:

$$r_t = z_t - \hat{x}_t \quad (3)$$

where z_t is the actual observation at time step t . The residual r_t is passed through a two-layer fully connected network to estimate the adaptive fusion weight:

$$K_t = \text{Sigmoid}(W_2 \text{ReLU}(W_1 r_t)) \in [0,1] \quad (4)$$

where W_1 and W_2 are learnable parameters, and $\text{Sigmoid}(\cdot)$ denotes the element-wise Sigmoid activation function. The final state update equation fuses the prior prediction from the SDE and the observation-guided correction term as follows:

$$x_t = \tilde{x}_t + K_t \odot (z_t - \tilde{x}_t) \quad (5)$$

where $\odot$ denotes element-wise multiplication. This formulation enables filtering updates from prior (predictive) to posterior (fused) latent states based on continuous-time evolution and observed evidence. To quantify the uncertainty in the latent state estimation at each time step, we further define the following metric:

$$\text{Uncertainty}_t = 1 - K_t \quad (6)$$

This uncertainty measure can be utilized in downstream decision modules or incorporated into loss function design to reflect the model's confidence in the current observation.

3.2. Quantum Uncertainty-Aware Fusion Module

Fusing smoothed features with their associated uncertainty information is critical for enhancing discriminative power in deepfake detection. Conventional methods typically rely on simple concatenation or attention-based fusion strategies, which struggle to capture the intrinsic correlation between these two types of information. To address this issue, we propose the Quantum Uncertainty-Aware Fusion Module (QUAFM), which introduces density matrices and von Neumann entropy from quantum information theory. Unlike classical covariance matrices that only capture second-order statistics and linear correlations, density matrices can encode both classical uncertainty and the intrinsic correlations among feature states in a unified framework [50]. The von Neumann entropy operates on the eigenvalue spectrum of the density matrix and captures the correlations between feature states that classical entropy measures cannot capture [21]. As depicted in

Figure 3, QUAFM receives the refined states x_t (Equation (5)) as input features F and the uncertainty estimates (Equation (6)) as U from CNSDFM (Figure 2). The overall workflow is illustrated in Figure 3a, and the density matrix construction process is depicted in Figure 3b.

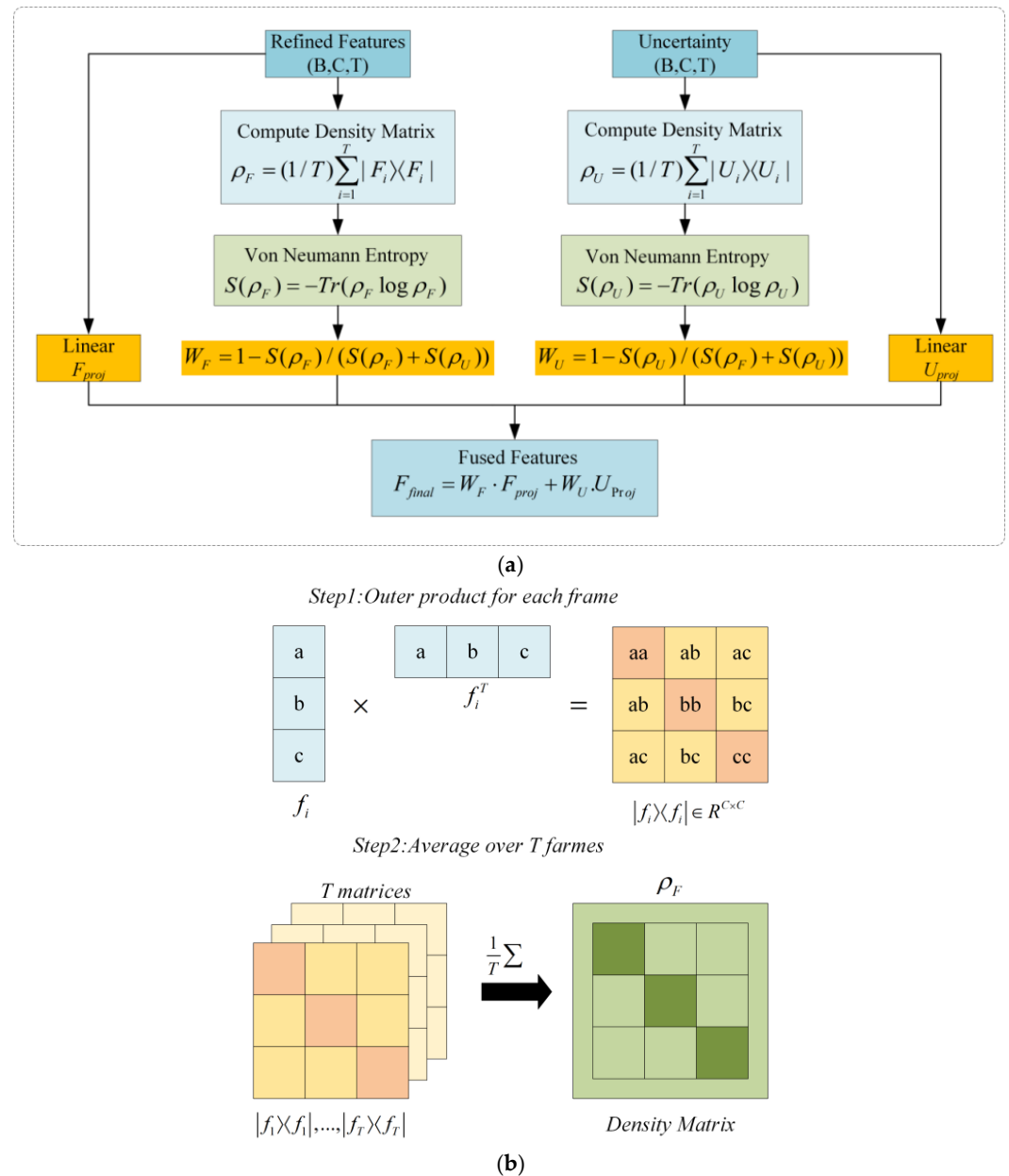


Figure 3. (a) The workflow of QUAFM. By leveraging density matrices and von Neumann entropy, this module achieves uncertainty-aware feature fusion that improves robustness to local forgery artifacts. (b) Illustration of density matrix construction in QUAFM. The density matrix encodes both feature distributions and inter-state correlations, providing richer representations than conventional covariance-based methods.

Let the input consist of the preprocessed smoothed features denoted by $F \in \mathbb{R}^{B \times C \times T}$, and the corresponding uncertainty information be denoted by $U \in \mathbb{R}^{B \times C \times T}$. First, both types of features are independently normalized along the temporal dimension L_2 , and their density matrix representations are constructed as follows:

$$\begin{cases} \rho_F = \frac{1}{T} \sum_{i=1}^T |f_i\rangle\langle f_i| \\ \rho_U = \frac{1}{T} \sum_{i=1}^T |u_i\rangle\langle u_i| \end{cases} \quad (7)$$

The density matrix can be interpreted as a statistical superposition of all temporal feature states, characterizing their spatial distribution structure. Notably, these density matrices are Hermitian by construction, since each outer product used in the summation satisfies this symmetry property. This Hermitian symmetry ensures that the eigenvalue spectrum consists entirely of real non-negative values, which is essential for the subsequent von Neumann entropy computation. To quantify the uncertainty associated with each representation, we introduce the von Neumann entropy, defined as:

$$S(\rho) = -\text{Tr}(\rho \log \rho) \quad (8)$$

where ρ is the density matrix and $\text{Tr}(\cdot)$ denotes the trace operator. The Hermitian symmetry of the density matrix guarantees that all eigenvalues are real and non-negative with unit sum, which makes the von Neumann entropy well-defined. Unlike classical Shannon entropy, which only measures the uncertainty of individual random variables, the von Neumann entropy operates on the eigenvalue spectrum of the density matrix and thus captures the correlations and entanglement between feature states [21]. This property enables more accurate quantification of the uncertainty within temporal features, as it accounts for the statistical dependencies that classical entropy measures cannot capture. Next, we compute fusion weights based on the entropy values. Features with lower uncertainty are assigned higher weights to enhance their contribution in the fusion process. The fusion weights W_F and W_U for the two inputs are calculated as:

$$\begin{cases} W_F = 1 - \frac{S(\rho_F)}{S(\rho_F) + S(\rho_U) + \epsilon} \\ W_U = 1 - \frac{S(\rho_U)}{S(\rho_F) + S(\rho_U) + \epsilon} \end{cases} \quad (9)$$

Finally, we perform temporal average pooling on both feature streams and project them into a unified feature space via a linear transformation. The final fused representation is obtained by the weighted combination of the two branches, as given by:

$$\begin{cases} F_{\text{proj}} = W \cdot \text{AvgPool}_T(F) \\ U_{\text{proj}} = W \cdot \text{AvgPool}_T(U) \\ F_{\text{final}} = W_F \cdot F_{\text{proj}} + W_U \cdot U_{\text{proj}} \end{cases} \quad (10)$$

3.3. Fractional-Order Temporal Anomaly Detection Module

In deepfake video detection, manipulations typically affect only a small number of frames or localized regions, resulting in sparse and subtle forgery signals. Traditional supervised learning frameworks treat all frames with equal importance, making it difficult to model the temporal dynamics of key anomalous frames. To address this challenge, we propose the Fractional-Order Temporal Anomaly Detection Module (FOTADM), which generates frame-wise anomaly scores and uses them as loss weighting factors. FOTADM introduces fractional-order differential operators to capture long-range dependencies and multi-scale temporal variations through learnable fractional orders. The overall workflow is illustrated in Figure 4.

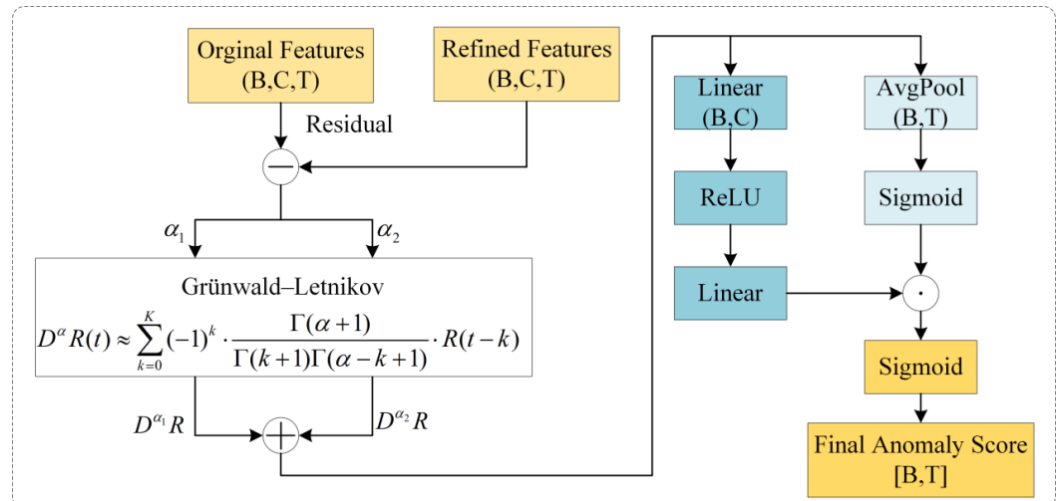


Figure 4. The workflow of FOTADM. The fractional-order operators capture long-range temporal dependencies through their non-local memory property, enabling fine-grained anomaly localization at the frame level.

Let the original temporal feature sequence and its smoothed reference be denoted by $X_{\text{orig}}, X_{\text{refined}} \in \mathbb{R}^{B \times C \times T}$, where B , C , and T denote the batch size, number of channels, and number of temporal steps, respectively. The residual sequence is first computed frame-by-frame as:

$$R = |X_{\text{orig}} - X_{\text{refined}}| \in \mathbb{R}^{B \times C \times T} \quad (11)$$

To model more discriminative temporal anomaly responses, FOTADM introduces the Grünwald–Letnikov fractional-order differential operator, defined as:

$$D^{\alpha}R(t) \approx \sum_{k=0}^L (-1)^k \cdot \frac{\Gamma(\alpha+1)}{\Gamma(k+1)\Gamma(\alpha-k+1)} \cdot R(t-k) \quad (12)$$

where $\alpha \in (0,2)$ is a learnable fractional order, $\Gamma(\cdot)$ denotes the Gamma function, and L is the truncation length. Compared to integer-order derivatives, fractional-order operators exhibit non-local memory characteristics, enabling the modeling of long-range temporal dependencies. Physically, the Grünwald–Letnikov coefficients decay according to a power-law with exponent determined by α , where smaller α yields slower decay and thus longer memory span [22].

To enhance the representation of multi-scale temporal anomalies, FOTADM applies two parallel fractional orders α_1 and α_2 , generating two distinct residual sequences, which are then fused:

$$R_{\text{frac}} = D^{\alpha_1}(R) + D^{\alpha_2}(R) \quad (13)$$

Subsequently, global average pooling is applied along the channel dimension to extract a single scalar response for each time step:

$$R_{\text{avg}} = \frac{1}{C} \sum_{c=1}^C R_{\text{frac}} \quad (14)$$

This global response sequence is then passed through a two-layer perceptron (MLP) to produce normalized anomaly scores:

$$A = \text{Sigmoid}\left(W_2 \cdot \text{ReLU}(W_1 \cdot R_{\text{avg}}^T)\right) \in [0,1]^{B \times T} \quad (15)$$

Here, W_1 and W_2 are learnable weight matrices, and $\text{Sigmoid}(\cdot)$ denotes the Sigmoid activation function. The final output A represents the anomaly degree at each time

step, where higher values indicate more significant dynamic variations in the residual space, potentially corresponding to forged regions.

The anomaly scores produced by FOTADM are not used directly as detection outputs. Instead, they are integrated into the loss function of the backbone network to adaptively weight training on anomalous frames. This design follows the weakly supervised learning paradigm commonly adopted in video anomaly detection, where only video-level labels are available during training while frame-level predictions are desired [51]. Critically, the frame-level anomaly scores function as soft attention weights rather than hard pseudo-labels, which avoids the self-confirmation bias typically associated with pseudo-label-based methods [52]. Let $Z \in \mathbb{R}^{B \times T}$ be the prediction of the backbone model and Y the ground-truth label. The standard cross-entropy loss is given by:

$$\mathcal{L}_{CE} = \frac{1}{B} \sum_{i=1}^B CE(Z_i, Y_i) \quad (16)$$

To incorporate anomaly-aware weighting, the loss is reformulated as:

$$\mathcal{L}_{weighted} = \frac{1}{B} \sum_{i=1}^B (1 + \alpha \cdot A_i) \cdot CE(Z_i, Y_i) \quad (17)$$

Here, $\alpha > 0$ is a modulation factor that controls the influence of anomaly scores on the training loss. Frames assigned higher anomaly scores contribute more significantly to the overall loss, encouraging the model to focus on improving recognition accuracy at these time steps during optimization. Notably, the supervision signal is solely derived from the video-level ground-truth label Y , rather than the generated anomaly scores, which prevents the model from reinforcing its own predictions and ensures stable learning. During inference, video-level predictions are obtained by averaging the frame-level outputs [51].

4. Experiments

This section presents the experimental setup and result analysis of the proposed method. We begin by detailing the experimental configuration, including the datasets used, preprocessing procedures, and implementation details. Subsequently, we conduct a comprehensive evaluation of the proposed algorithm through a series of experiments, including accuracy comparison, cross-dataset generalization, ablation studies, visualization analysis, and robustness assessment, thereby validating its effectiveness and robustness from multiple perspectives.

4.1. Datasets

To comprehensively evaluate the detection performance of the proposed method, experiments were conducted on three widely used deepfake benchmark datasets: FaceForensics++ [53], Celeb-DF [54], and DFDC [55]. These datasets contain large-scale collections of both real and manipulated videos or images, and are representative and challenging benchmarks commonly adopted in state-of-the-art deepfake detection research.

(1) FaceForensics++ (FF++): The FF++ dataset consists of 1000 real videos and 4000 fake videos generated using four manipulation techniques: DeepFake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT), with 1000 videos per technique. The dataset is available in three different compression levels: uncompressed (c0), light compression (c23), and strong compression (c40). In this study, we adopt the c23 and c40 versions to assess the robustness of the model under varying compression conditions. For each forgery method, video samples are divided into training, validation, and testing sets in a ratio of 720:140:140. To mitigate the issue of class imbalance, 128 frames are uniformly

sampled from each real video, and 32 frames are sampled from each fake video. All images are cropped to the facial region using the official face masks provided by the dataset, and resized to 224×224 pixels as input to the model.

(2) Celeb-DF (CDF): The Celeb-DF dataset consists of 590 real videos and 5693 forged videos, involving subjects of diverse genders, ethnicities, and age groups. The fake videos are generated using an enhanced DeepFake algorithm, which significantly reduces common visual artifacts—such as unnatural boundaries and color inconsistencies—frequently found in earlier methods. This improvement leads to higher-quality forgeries and increases the difficulty of detection. In this study, we follow the official dataset partitioning protocol, extracting facial images from the designated videos. A total of 128 frames are sampled from each video, and face regions are detected and cropped using the Dlib library. All face images are then resized to a uniform resolution of 224×224 pixels for input into the model.

(3) Deepfake Detection Challenge (DFDC): Released by Meta AI in 2019, the DFDC dataset is one of the largest and most challenging benchmarks for deepfake detection. It comprises 19,197 real videos recorded by 430 actors, along with approximately 100,000 fake videos generated using a variety of forgery techniques, including DFAE, MM/NN Face Swap, NTH, and FSGAN, among others. The dataset exhibits high diversity and complexity in both content and manipulation methods. Given the massive scale of the dataset, we selected the last 10 compressed archive files out of the 50 officially provided and constructed a subset consisting of 1000 real videos and 1000 fake videos. For each video, 64 frames were uniformly sampled, facial regions were extracted using Dlib, and all frames were resized to 224×224 pixels for training and evaluation purposes.

4.2. Implementation Details

The proposed network was implemented using the PyTorch 2.7.1 deep learning framework. A ResNet3D-50 backbone pre-trained on the Kinetics-400 dataset was adopted. All experiments were conducted on an NVIDIA RTX A6000 GPU. The model was trained for 20 epochs with a batch size of 32 using the Adam optimizer, where both the learning rate and weight decay were set to 1×10^{-4} . A custom anomaly score-based loss function was employed to guide the optimization. Accuracy (ACC) and Area Under the ROC Curve (AUC) were used as evaluation metrics. ACC reflects the model's classification performance, while AUC provides a threshold-independent assessment of detection capability. To ensure fair comparison, video-level ACC and AUC were reported following the protocol in [11]. We strictly followed the official train/validation/test split without any overlap. In addition, no cross-video data augmentation was applied to avoid data leakage.

4.3. Accuracy Comparison

The proposed method was evaluated on two standard datasets, FF++ (c23) and FF++ (c40), and compared with several state-of-the-art deepfake detection approaches. The results, summarized in Table 1, demonstrate that the proposed method achieved an ACC of 99.12% and an AUC of 99.81% on FF++ (c23), and an ACC of 94.89% with an AUC of 97.91% on FF++ (c40). While the ACC ranks second among all compared methods, the AUC consistently outperforms existing approaches. These results indicate the robustness of the proposed model in maintaining high detection accuracy under varying compression levels.

Table 1. Comparative experiments on the FF++ dataset, the best results are highlighted in bold.

Methods	FF++(c23)		FF++(c40)	
	ACC (%)	AUC (%)	ACC (%)	AUC (%)
Xception [56]	95.73	96.30	86.86	89.30
MesoNet [57]	83.10	87.68	70.47	78.42
EN-B4 [58]	96.63	99.18	86.67	88.20
Two Branch [45]	96.43	98.70	86.34	86.59
Face X-ray [59]	83.56	87.35	57.42	61.60
MADD [60]	97.60	99.29	88.69	90.40
Add-Net [61]	96.78	97.74	87.50	91.01
RFM [62]	95.69	98.79	87.06	89.83
MF-Net [12]	98.21	99.70	90.26	92.09
CLAF [3]	97.12	99.49	87.58	95.03
TAN-GFD [13]	97.17	99.21	87.42	89.26
F3-Net [63]	97.52	98.10	90.43	93.30
ID3 [15]	93.76	97.82	83.64	86.38
SFICnv [64]	93.61	98.46	89.41	95.90
MRL [65]	93.82	98.27	91.81	96.18
DFAD [66]	–	98.96	–	96.69
PDCNet [67]	97.60	99.00	96.30	96.40
D2Fusion [68]	97.77	99.42	91.43	93.33
CUTA [69]	98.21	99.63	92.35	96.51
SDEQ-Net	99.12	99.81	94.89	97.91

4.4. Generalization Experiments

Two sets of experiments were conducted to evaluate the generalization capability of the proposed method. In the first set of experiments, the model was trained on the FF++ (c23) dataset and tested on two unseen datasets, Celeb-DF and DFDC, following a cross-dataset evaluation protocol. Several state-of-the-art deepfake detection methods were included for comparison. The Area Under the ROC Curve (AUC) was used as the evaluation metric, and the results are summarized in Table 2. On the Celeb-DF dataset, the proposed method achieved an AUC of 89.55%, outperforming all competing methods. On the DFDC dataset, it obtained an AUC of 86.21%, also achieving the best performance among the compared approaches.

Table 2. Generalization performance on the Celeb-DF and DFDC datasets, the best results are highlighted in bold.

Train Set	Methods	Test Set AUC (%)	
		Celeb-DF	DFDC
FF++(c23)	Xception [56]	65.27	69.80
	EN-B4 [58]	68.52	70.12
	Two Branch [45]	73.41	–
	F3-Net [63]	72.88	71.21
	MADD [60]	67.44	–
	TAN-GFD [13]	72.33	73.46
	CLAF [3]	78.94	80.02
	Local-relation [70]	78.26	76.53
	Face X-ray [59]	74.76	71.15
	NACO [5]	89.50	76.70
	FTCNT [6]	86.90	74.00
	DFAD [66]	71.74	72.66
	RFFR [71]	81.97	72.08
	PDCNet [67]	79.82	70.03
	SDEQ-Net	89.55	86.21

In the second set of experiments, a leave-one-subset-out strategy was applied. Each of the four subsets in FF++ (c23)—DF, F2F, FS, and NT—was used as the training set, while the remaining three subsets served as the test sets. This procedure was repeated four times, and the AUC was reported as the evaluation metric. The results are shown in Table 3. When DF was used for training, the average AUC across the other three subsets reached 90.21%, representing the best performance among all methods. When trained on F2F, the average AUC was 92.61%, again outperforming all baseline models. Using FS as the training set yielded an average AUC of 92.00%, while training on NT achieved an average AUC of 91.18%, the best result in this setting as well. To provide a more intuitive understanding of the contribution of each module, we visualize the ablation results in Figure 5.

These results consistently demonstrate that the proposed deepfake face detection method exhibits superior generalization ability compared to existing approaches when applied to previously unseen or cross-dataset fake video scenarios.

Table 3. Generalization performance on FF++ subset datasets. The best results are highlighted in bold.

Train Set	Methods	Test Set AUC (%)				Avg	Cross Avg
		DF	F2F	FS	NT		
DF	Xception [56]	99.38	75.05	49.13	80.39	75.99	68.19
	EN-B4 [58]	99.65	73.60	40.73	73.94	71.87	62.61
	MADD [60]	99.51	66.41	67.33	66.01	74.82	66.58
	MF-Net [12]	99.97	70.82	74.54	82.37	81.93	75.91
	RFFR [71]	99.19	76.61	68.96	74.83	79.90	—
	D2Fsuion [68]	99.98	77.88	62.65	75.23	78.93	—
	SDEQ-Net	99.43	86.35	89.47	85.59	90.21	88.83
F2F	Xception [56]	87.56	99.53	65.23	65.90	79.56	72.90
	EN-B4 [58]	87.15	99.26	51.60	66.95	76.24	68.57
	MADD [60]	73.04	97.96	65.10	71.88	77.00	70.01
	MF-Net [12]	86.74	99.96	67.51	90.42	86.16	81.56
	RFFR [71]	93.75	99.61	78.62	79.56	87.81	—
	D2Fsuion [68]	89.50	99.86	62.47	75.23	81.76	—
	SDEQ-Net	91.05	99.90	89.89	89.59	92.61	90.77
FS	Xception [56]	70.12	61.70	99.36	68.71	74.97	66.84
	EN-B4 [58]	61.44	68.96	99.57	49.83	69.95	60.08
	MADD [60]	82.33	61.65	98.82	54.79	74.40	66.26
	MF-Net [12]	76.62	70.12	99.91	60.34	76.75	69.06
	RFFR [71]	87.46	75.96	99.42	55.87	79.68	—
	D2Fsuion [68]	77.50	69.76	99.92	58.42	76.40	—
	SDEQ-Net	89.06	89.46	99.99	89.51	92.00	90.82
NT	Xception [56]	93.09	84.82	47.98	99.50	81.35	75.30
	EN-B4 [58]	83.98	69.08	46.32	97.59	74.24	66.46
	MADD [60]	74.56	80.61	60.90	93.34	77.35	72.02
	MF-Net [12]	89.68	74.97	64.59	99.36	82.15	76.41
	RFFR [71]	84.31	81.04	54.67	96.17	79.05	—
	D2Fsuion [68]	94.44	71.08	80.75	99.43	86.42	—
	SDEQ-Net	85.51	89.72	89.73	99.77	91.18	87.72

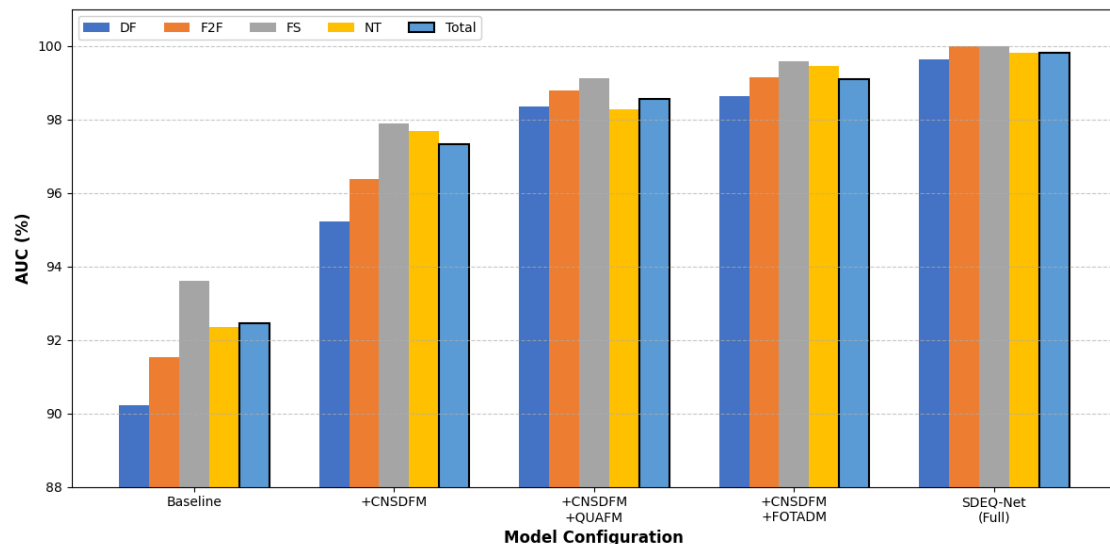


Figure 5. Ablation study results on FF++ (c23). The progressive integration of CNSDFM, QUAFM, and FOTADM leads to consistent performance improvements across all forgery subsets, with the full SDEQ-Net achieving the highest AUC of 99.81%.

4.5. Ablation Study

To evaluate the effectiveness of the three proposed modules in the deepfake face detection framework, an ablation study was conducted. The model was trained on the FF++ (c23) dataset, and tested separately on its four subsets: DF, F2F, FS, and NT. The Area Under the ROC Curve (AUC) was used as the evaluation metric. The results are summarized in Table 4. Compared to the baseline model, incorporating the Continuous-time Neural Stochastic Differential Filtering Module (CNSDFM) led to a 4.87% improvement in AUC on FF++ (c23). Adding the Quantum Uncertainty-Aware Fusion Module (QUAFM) on top of CNSDFM further improved the AUC by 1.23%. Finally, introducing the Fractional-order Temporal Anomaly Detection Module (FOTADM) resulted in an additional gain of 1.25%. Overall, the progressive integration of these three modules significantly enhanced the model's performance across all subsets. Specifically, the AUC increased by 9.42% on DF, 8.46% on F2F, 6.37% on FS, and 7.47% on NT, with an overall improvement of 7.35% on the entire FF++ (c23) dataset. These results clearly demonstrate the effectiveness and complementary contributions of the proposed CNSDFM, QUAFM, and FOTADM in boosting detection performance on both the full dataset and its individual subsets.

Table 4. Ablation study on the FF++ dataset. The best results are highlighted in bold.

Train Set	ID	CNSDFM	QUAFM	FOTADM	Test Set AUC (%)				Total
					DF	F2F	FS	NT	
FF++(C23)	1				90.23	91.53	93.62	92.35	92.46
	2	✓			95.23	96.38	97.89	97.68	97.33
	3	✓	✓		98.36	98.78	99.12	98.29	98.56
	4	✓		✓	98.63	99.16	99.59	99.45	99.10
	5	✓	✓	✓	99.65	99.99	99.99	99.82	99.81

To further investigate the effectiveness of the fractional-order differential operator in FOTADM, we conducted additional experiments comparing different temporal weighting strategies. As shown in Table 5, four approaches were examined: (1) Uniform weighting, which treats all frames equally ($\alpha = 0$ in Equation (17)); (2) L1-Residual

weighting, which uses absolute residual magnitudes directly; and (3) Integer-order weighting, which employs standard first-order temporal difference ($\alpha = 1$); and (4) FOTADM with learnable fractional orders.

Table 5. Comparison of temporal weighting strategies in FOTADM on FF++ (c23). The best results are highlighted in bold.

Weighting Strategy	AUC (%)	Learned α_1/α_2
Uniform ($\alpha = 0$)	98.56	—
L1-Residual	98.92	—
Integer-order ($\alpha = 1$)	99.23	1.0/1.0
FOTADM	99.81	0.73/1.42

The results demonstrate that anomaly-based weighting consistently outperforms uniform weighting, and the proposed FOTADM with learnable fractional orders surpasses the integer-order variant by 0.58%. Notably, the learned orders converged to $\alpha_1 = 0.73$ and $\alpha_2 = 1.42$. According to the power-law decay property of Grünwald–Letnikov coefficients, $\alpha_1 = 0.73$ corresponds to slower coefficient decay and thus longer memory span, making the operator more sensitive to accumulated temporal inconsistencies. Conversely, $\alpha_2 = 1.42$ exhibits faster decay, emphasizing recent frame differences and capturing local discontinuities. The combination of these two complementary fractional orders enables FOTADM to simultaneously detect both long-range temporal drift and short-term abrupt artifacts in deepfake videos.

4.6. Grad-CAM Visualization

To further validate the effectiveness of the proposed detection method, Grad-CAM was employed to visualize the attention distribution across different convolutional layers. Figure 6 presents the visualization results from Layer1 to Layer4, clearly illustrating the evolution of the model’s focus as the network depth increases. At the shallow layer (Layer1), the model primarily responds to global facial structures and low-level texture features, with dispersed activation regions mostly located around the eyes and forehead. As the network deepens, the attention in Layer2 and Layer3 gradually shifts towards more representative local facial regions, such as the eye contour, nasal bridge, and mouth corners—regions that are inherently prone to forgery artifacts due to the technical limitations of deepfake generation pipelines. Specifically, the eye region frequently exhibits artifacts such as irregular pupil shapes, inconsistent corneal reflections between the two eyes, and unnatural blinking patterns, as generative models struggle to reproduce the physical and physiological constraints of human eyes [72]. The mouth and nasal regions are particularly susceptible to blending boundary artifacts and texture inconsistencies, since most face-swapping and face-reenactment methods require warping and blending the synthesized facial content onto the target frame, inevitably introducing visible seams and color mismatches around these semantically complex areas [73,74]. At the deepest level (Layer4), attention becomes highly concentrated around the nasal alae and mouth corners, indicating that the model is capable of accurately localizing the most discriminative forged regions in high-level semantic space.



Figure 6. Grad-CAM visualization results from different convolutional layers. The progressive attention shift from global structures to localized forgery-prone regions (eyes, nose, mouth) demonstrates the model's interpretability and its ability to learn semantically meaningful features.

Figure 7 shows Grad-CAM results for four types of forgeries (DF, F2F, FS, and NT) from the FF++ (c23) dataset. It can be observed that the proposed method accurately identifies forged regions across different manipulation types, with strong activation in key facial features. This consistent focus on the eye, nose, and mouth regions across diverse manipulation techniques aligns with findings from recent forensic studies, which demonstrate that these facial areas are the most manipulation-prone regions where geometric asymmetries, lighting irregularities, and up-sampling artifacts from the generator's decoder are most pronounced [75]. This suggests that the method not only achieves high classification accuracy but also exhibits good interpretability and region-aware capability. These findings further confirm the effectiveness and feasibility of the proposed method in the task of deepfake detection.

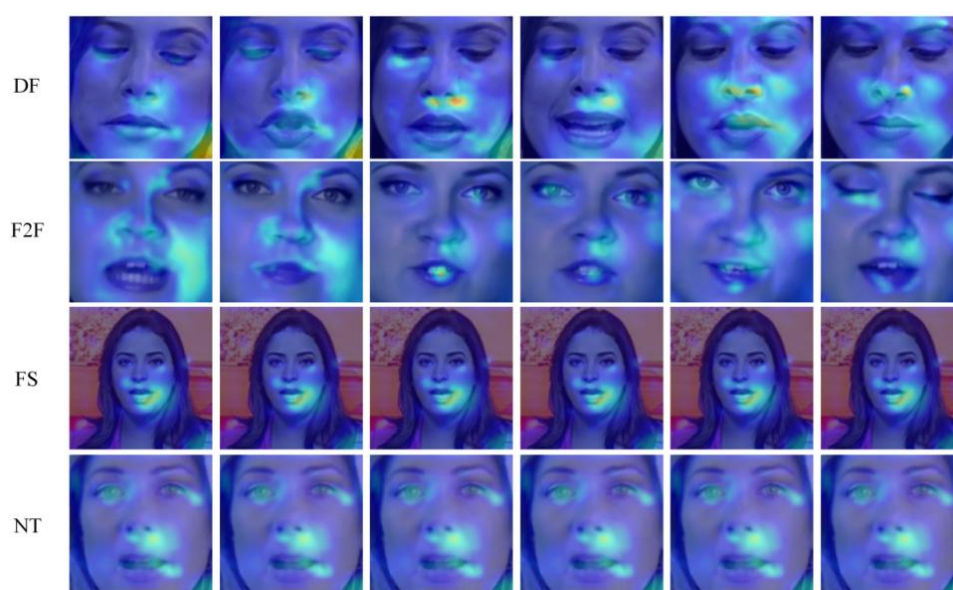


Figure 7. Grad-CAM visualization results under different types of forgery. The consistent focus on manipulation-prone facial regions across diverse forgery methods (DF, F2F, FS, NT) validates the generalization capability of the learned representations.

4.7. Robustness Evaluation

To assess the robustness of the proposed method under various visual degradation conditions, the model was trained on the FF++ (c23) dataset and systematically evaluated under seven types of degradation transformations: saturation variation, contrast variation, block occlusion, Gaussian noise, Gaussian blur, pixelation, and JPEG compression. Each degradation type was applied at five severity levels to simulate a spectrum of distortions ranging from mild to severe. The proposed method was compared against three representative deepfake detection models: F3-Net (frequency domain-based), Xception, and MADD (both spatial domain-based). The results are illustrated in Figure 8. Under color perturbations—namely saturation and contrast variations—the proposed method consistently achieved high AUC scores (97.86–99.02%), significantly outperforming the baselines. In block occlusion scenarios (with the number of occlusion blocks increasing from 10 to 50), our method exhibited the smallest performance degradation, maintaining strong stability. For Gaussian noise (with standard deviation σ ranging from 2 to 10), the proposed method achieved an AUC of 72.06% under the highest noise level, which was notably superior to F3-Net (55.05%) and Xception (65.37%).

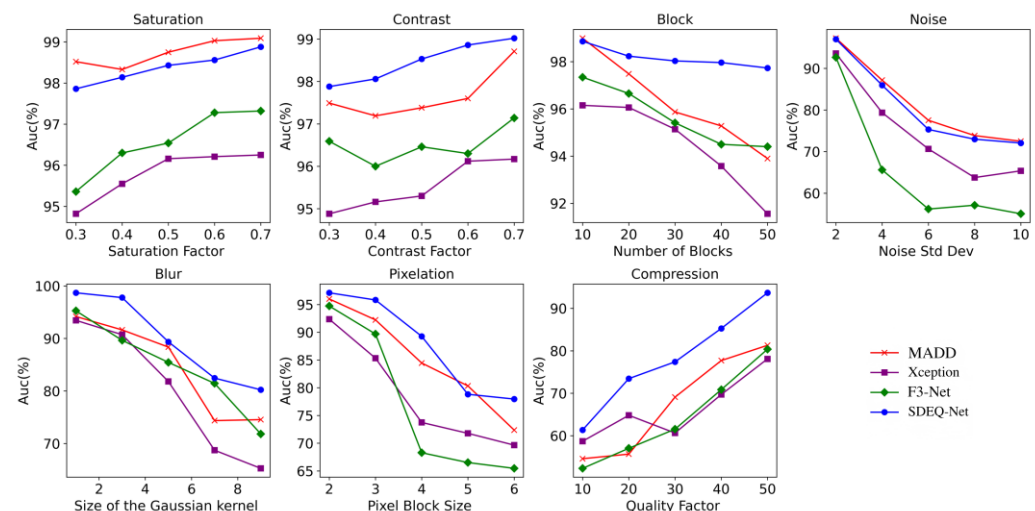


Figure 8. Robustness evaluation under seven types of perturbations. SDEQ-Net consistently outperforms baseline methods across all degradation conditions, demonstrating its practical reliability for real-world deployment scenarios.

Under Gaussian blur, with kernel sizes increasing from 1 to 9, the AUC of the proposed method declined from 98.69% to 80.23%, but still exceeded MADD (74.55%) and F3-Net (71.82%). Regarding pixelation (with block sizes from 2 to 6), the proposed method demonstrated greater adaptability, maintaining an AUC of 77.96% even at the most severe level. In the scenario of heavy JPEG compression (quality factor reduced from 50 to 10), the AUC remained at 61.35%, indicating stronger robustness compared to the competing approaches.

Overall, the results confirm that the proposed deepfake detection method exhibits superior stability under various complex visual degradation conditions, highlighting its strong potential for real-world video deepfake detection applications.

4.8. Model Complexity and Inference Efficiency Analysis

To comprehensively evaluate the deployment potential of different types of deepfake detection models in real-world scenarios, we systematically compare representative 2D CNNs, 3D CNNs, Transformer-based architectures, and the proposed SDEQ-Net across two critical dimensions: model complexity (measured by parameter count and FLOPs) and inference efficiency (measured by frames per second, FPS, and average inference time per video clip, ms/clip). All experiments were conducted on the same hardware platform (NVIDIA RTX A6000 GPU), using standardized input clips consisting of 16 frames with a spatial resolution of 224×224 pixels to ensure fair and reproducible comparisons.

Table 6 presents the comparison results in terms of parameter count, FLOPs, average inference time, and FPS under FP32 precision. R3D-50 demonstrated the highest efficiency, achieving an inference speed of 567.17 FPS with an average inference time of only 28.21 ms per clip. While the proposed SDEQ-Net exhibits slightly higher complexity (60.16 GFLOPs and 47.83 M parameters), it maintains competitive inference performance, reaching 280.94 FPS and 56.95 ms/clip—significantly outperforming more complex models such as Swin3D. Regarding real-time applicability, standard video streams typically operate at 24–30 FPS, while high-frame-rate content may reach 60 FPS. The achieved inference speed of 280.94 FPS is approximately 9× higher than the 30 FPS broadcast standard, indicating that SDEQ-Net is well-suited for real-time deepfake detection on server-grade GPUs. The end-to-end latency of 56.95 ms per 16-frame clip remains within acceptable bounds for online monitoring applications such as video conferencing platforms and social media content moderation. For edge deployment or latency-critical scenarios, optimization techniques including INT8 quantization and TensorRT acceleration can be employed to further reduce computational overhead.

Table 6. Comparison of model complexity and inference efficiency. The best results are highlighted in bold.

Method	Params (M)	FLOPs (G)	Inference Speed (FPS)	Inference Time/Clip (ms)
ResNet50 [76]	25.56	123.36	113.37	116.37
EN-B4 [58]	19.30	24.32	62.70	254.88
Xception [56]	22.90	134.40	142.60	112.16
Swin3D [77]	88.05	101.90	356.58	44.87
R3D-50 [78]	46.20	40.12	567.17	28.21
SDEQ-Net	47.83	60.16	280.94	56.95

To analyze the computational overhead introduced by each proposed module, we report the cumulative cost as each component is progressively added, as shown in Table 7. CNSDFM introduces the largest overhead (18.2 GFLOPs, 12.3 ms) due to LSTM encoding and Euler–Maruyama discretization. QUAFM adds 1.4 GFLOPs and 8.6 ms for density matrix construction and entropy computation, while FOTADM contributes only 0.4 GFLOPs and 7.8 ms for fractional-order differentiation. Overall, the three modules collectively add 20.0 GFLOPs and 28.7 ms to the backbone, enabling SDEQ-Net to strike a favorable balance among accuracy, complexity, and latency.

Table 7. Computational overhead analysis of each proposed module.

Method	Params (M)	FLOPs (G)	Inference Speed (FPS)	Inference Time/Clip (ms)
R3D-50 (baseline)	46.20	40.12	567.17	28.21
+CNSDFM	47.62	58.36	394.85	40.51
CNSDFM + QUAFM	47.77	59.76	325.53	49.15
+CNSDFM+QUAFM+FOTADM (SDEQ-Net)	47.83	60.16	280.94	56.95

4.9. Sensitivity Analysis

4.9.1. Backbone Architecture

The proposed CNSDFM, QUAFM, and FOTADM modules operate on intermediate feature representations, enabling integration with various backbone architectures. To evaluate this flexibility, we conducted experiments with four representative backbones: ResNet3D-50, EfficientNet-B4, Xception, and ViT-B. All models were trained on FF++ (c23) and evaluated on both intra-dataset and cross-dataset benchmarks. As shown in Table 8, the proposed modules yield consistent performance improvements across all backbone architectures. With EfficientNet-B4, the model achieves 98.92% AUC on FF++ (c23) and 87.43% on Celeb-DF. The Xception-based variant obtains 98.45% and 86.21% on the respective datasets. ViT-B, representing transformer-based architectures, achieves 99.15% on FF++ (c23) and 88.12% on Celeb-DF. ResNet3D-50 achieves the highest overall performance with 99.81% and 89.55% on FF++ (c23) and Celeb-DF, respectively, which motivates its selection as the default backbone in our framework.

Table 8. Performance comparison across different backbone architectures. All backbones are integrated with the proposed CNSDFM, QUAFM, and FOTADM. The best results are highlighted in bold.

Backbone	FF++ (c23) AUC (%)	Celeb-DF AUC (%)	DFDC AUC (%)
R3D-50 [78]	99.81	89.55	86.21
EN-B4 [58]	98.92	87.43	84.15
Xception [56]	98.45	86.21	82.89
Swin3D [77]	99.15	88.12	85.03

4.9.2. Training Data Ratio

We investigated model performance under varying amounts of training data. The models were trained using 100%, 75%, 50%, 25% of the FF++ (c23) training set and evaluated on both FF++ (c23) and Celeb-DF. Xception and EfficientNet-B4 were included as baselines. As shown in Table 9, the proposed method achieves 98.76% AUC on FF++ (c23) with 50% training data, compared to 96.45% for Xception and 97.23% for EfficientNet-B4. At 25% training data, the proposed method maintains 96.82% AUC, outperforming both baselines by 3.70% and 2.26%, respectively. In the cross-dataset evaluation, the proposed method exhibits a slower rate of performance degradation as training data decreases. With only 25% training data, the proposed method achieves 81.24% AUC on Celeb-DF, while Xception and EfficientNet-B4 obtain 62.45% and 57.89%, respectively.

Table 9. Performance comparison under different training data ratios. All models were trained on subsets of FF++ (c23) and evaluated on both FF++ (c23) and Celeb-DF. The best results are highlighted in bold.

Training Ratio	Method	FF++ (c23) AUC (%)	Celeb-DF AUC (%)
100%	SDEQ-Net	99.81	89.55
	Xception [56]	99.21	73.41
	EN-B4 [58]	99.63	68.52
75%	SDEQ-Net	99.45	88.21
	Xception [56]	98.34	71.85
	EN-B4 [58]	98.87	66.78
50%	SDEQ-Net	98.76	85.63
	Xception [56]	96.45	68.23
	EN-B4 [58]	97.23	63.12
25%	SDEQ-Net	96.82	81.24
	Xception [56]	93.12	62.45
	EN-B4 [58]	94.56	57.89

4.9.3. Regularization Strategy

We analyzed the effects of weight decay and dropout on model generalization. As shown in Table 10, without regularization, the model achieves 99.56% AUC on FF++ (c23) but only 76.82% on Celeb-DF, resulting in a significant performance gap of 22.74%. Applying weight decay (1×10^{-4}) reduces the FF++ (c23) performance slightly to 99.48% while improving cross-dataset generalization to 83.67%. Adding dropout (0.3) alone yields 99.35% on FF++ (c23) and 85.41% on Celeb-DF. The combination of weight decay (1×10^{-4}) and dropout (0.3) achieves the optimal balance, with 99.81% on FF++ (c23) and 89.55% on Celeb-DF. Increasing the weight decay to 1×10^{-3} leads to performance degradation on both datasets, suggesting that excessive regularization impairs the model's representational capacity.

Table 10. Ablation study on regularization strategies. Δ denotes the performance gap between FF++ (c23) and Celeb-DF. The best results are highlighted in bold.

Weight Decay	Dropout	FF++ (c23) AUC (%)	Celeb-DF AUC (%)	AUC Δ (%)
0	0	99.56	76.82	22.74
1×10^{-4}	0	99.48	83.67	15.81
0	0.3	99.35	85.41	13.94
1×10^{-4}	0.3	99.81	89.55	10.26
1×10^{-3}	0.3	98.67	86.23	12.44

5. Conclusions

In this paper, we propose SDEQ-Net, a novel deepfake video anomaly detection framework that integrates stochastic differential equation modeling, quantum uncertainty-aware fusion, and fractional-order temporal analysis. The proposed method addresses the challenges of continuous temporal dynamics modeling and fine-grained anomaly localization through three complementary modules: CNSDFM for continuous-time state evolution, QUAFM for uncertainty-aware feature fusion based on Hermitian-symmetric density matrices, and FOTADM for frame-level anomaly scoring. Extensive experiments on FaceForensics++, Celeb-DF, and DFDC demonstrate that SDEQ-Net con-

sistently outperforms state-of-the-art methods in detection accuracy, cross-dataset generalization, and robustness. Future work will focus on improving computational efficiency and extending the framework to multimodal deepfake detection scenarios.

Author Contributions: R.Z.: Conceptualization, Methodology, Investigation, Writing—Original Draft. B.L.: Investigation, Formal Analysis. D.X.: Supervision, Writing—Review & Editing, Project Administration, Funding Acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Major Science and Technology Special Project of Henan Province (Grant No. 241100210100).

Data Availability Statement: Code (<https://github.com/1090176349/DeepFakeVideo>, accessed on 13 January 2026). The datasets that support the findings of this study are publicly available from the following sources: FaceForensics++ (<https://github.com/ondyari/FaceForensics>, accessed on 13 January 2026), Celeb-DF (<https://www.kaggle.com/datasets/reubensuju/celeb-df-v2>, accessed on 13 January 2026), and DFDC (<https://www.kaggle.com/c/deepfake-detection-challenge>, accessed on 13 January 2026).

Acknowledgments: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Zhang, H.; Chen, B.; Wang, J.; Zhao, G. A local perturbation generation method for GAN-generated face anti-forensics. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 661–676.
2. Wang, J.; Lei, J.; Li, S.; Zhang, J. STA-3D: Combining Spatiotemporal Attention and 3D Convolutional Networks for Robust Deepfake Detection. *Symmetry* **2025**, *17*, 1037.
3. Yang, G.; Zuo, B.; Fang, X.; Zhang, J. Exploiting optimized forgery representation space for general fake face detection. *Pattern Anal. Appl.* **2025**, *28*, 9.
4. Wang, F.; Chen, Q.; Jing, B.; Tang, Y.; Song, Z.; Wang, B. Deepfake Detection Based on the Adaptive Fusion of Spatial-Frequency Features. *Int. J. Intell. Syst.* **2024**, *2024*, 7578036.
5. Zhang, D.; Xiao, Z.; Li, S.; Lin, F.; Li, J.; Ge, S. Learning natural consistency representation for face forgery video detection. In *Computer Vision—ECCV 2024, 18th European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 407–424.
6. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international Conference on computer Vision, Montreal, QC, Canada, 10–17 October 2021*; pp. 15044–15054.
7. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Li, H. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; pp. 4129–4138.
8. Amerini, I.; Galteri, L.; Caldelli, R.; Del Bimbo, A. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF International Conference on Computer vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019*; pp. 1205–1207.
9. Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; Verdoliva, L. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; pp. 15108–15117.
10. De Lima, O.; Franklin, S.; Basu, S.; Karwoski, B.; George, A. Deepfake detection using spatiotemporal convolutional networks. *arXiv* **2020**, arXiv:2006.14749.
11. Haliassos, A.; Vougioukas, K.; Petridis, S.; Pantic, M. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021*; pp. 5039–5049.

12. Duan, H.; Jiang, Q.; Jin, X.; Wozniak, M.; Zhao, Y.; Wu, L.; Yao, S.; Zhou, W. Mf-net: Multi-feature fusion network based on two-stream extraction and multi-scale enhancement for face forgery detection. *Complex Intell. Syst.* **2025**, *11*, 11.
13. Bai, N.; Wang, X.; Han, R.; Hou, J.; Wang, Q.; Pang, S. Towards generalizable face forgery detection via mitigating spurious correlation. *Neural Netw.* **2025**, *182*, 106909.
14. Siddiqui, F.; Yang, J.; Xiao, S.; Fahad, M. Enhanced deepfake detection with DenseNet and Cross-ViT. *Expert Syst. Appl.* **2025**, *267*, 126150.
15. Yin, Z.; Wang, J.; Xiao, Y.; Zhao, H.; Li, T.; Zhou, W.; Liu, A.; Liu, X. Improving Deepfake Detection Generalization by Invariant Risk Minimization. *IEEE Trans. Multimed.* **2024**, *26*, 6785–6798.
16. Gao, J.; Micheletto, M.; Orrù, G.; Concas, S.; Feng, X.; Marcialis, G.L.; Roli, F. Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108450.
17. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589.
18. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **2021**, *110*, 457–506.
19. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York City, NY, USA, 9–24 June 2016; pp. 1050–1059.
20. Oh, Y.; Lim, D.-Y.; Kim, S. Stable neural stochastic differential equations in analyzing irregular time series data. *arXiv* **2024**, arXiv:2402.14989.
21. Kim, J.; Kang, S.; Hwang, D.; Shin, J.; Rhee, W. Vne: An effective method for improving deep representation by manipulating eigenvalue distribution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3799–3810.
22. Coelho, C.; Costa, M.; Ferrás, L. Fractional calculus meets neural networks for computer Vision: A survey. *AI* **2024**, *5*, 1391–1426.
23. Khalid, H.; Woo, S.S. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 656–657.
24. Child, R. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv* **2020**, arXiv:2011.10650.
25. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
26. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.
27. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
28. Aghasanli, A.; Kangin, D.; Angelov, P. Interpretable-through-prototypes deepfake detection for diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 4–6 October 2023; pp. 467–474.
29. Mubarak, R.; Alsboui, T.; Alshaikh, O.; Inuwa-Dutse, I.; Khan, S.; Parkinson, S. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE Access* **2023**, *11*, 144497–144529.
30. Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A.G.; Yang, M.-H.; Hao, Y.; Essa, I. Magvit: Masked generative video transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10459–10469.
31. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)* **2019**, *3*, 80–87.
32. Li, Y.; Chang, M.-C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 11–13 December 2018; pp. 1–7.
33. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM International Conference on Multimedia, Virtual, 12–16 October 2020; pp. 2823–2832.
34. Gu, Z.; Chen, Y.; Yao, T.; Ding, S.; Li, J.; Huang, F.; Ma, L. Spatiotemporal inconsistency learning for deepfake video detection. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 3473–3481.
35. Haliassos, A.; Mira, R.; Petridis, S.; Pantic, M. Leveraging real talking faces via self-supervision for robust forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14950–14962.

36. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 83–92.
37. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8261–8265.
38. Xia, Z.; Qiao, T.; Xu, M.; Zheng, N.; Xie, S. Towards DeepFake video forensics based on facial textural disparities in multi-color channels. *Inf. Sci.* **2022**, *607*, 654–669.
39. Zhang, J.; Ni, J.; Nie, F.; Huang, J. Domain-invariant and patch-discriminative feature learning for general deepfake detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *21*, 47.
40. Yadav, A.; Vishwakarma, D.K. AW-MSA: Adaptively weighted multi-scale attentional features for DeepFake detection. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107443.
41. Zhou, X.; Han, H.; Shan, S.; Chen, X. Fine-grained Open-set Deepfake Detection via Unsupervised Domain Adaptation. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 7536–7547.
42. Yue, P.; Chen, B.; Fu, Z. Local region frequency guided dynamic inconsistency network for deepfake video detection. *Big Data Min. Anal.* **2024**, *7*, 889–904.
43. Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; Wei, Y. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; pp. 5052–5060.
44. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 772–781.
45. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; proceedings, part VII 16, 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 667–684.
46. Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J.Y.; Zhou, J. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv* **2024**, arXiv:2405.14616.
47. Saoudi, E.M.; Jaafari, J.; Andaloussi, S.J. Advancing human action recognition: A hybrid approach using attention-based LSTM and 3D CNN. *Sci. Afr.* **2023**, *21*, e01796.
48. Kidger, P.; Foster, J.; Li, X.C.; Lyons, T. Efficient and accurate gradients for neural sdes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18747–18761.
49. Kidger, P.; Foster, J.; Li, X.; Lyons, T.J. Neural sdes as infinite-dimensional gans. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 5453–5463.
50. González, F.A.; Gallego, A.; Toledo-Cortés, S.; Vargas-Calderón, V. Learning with density matrices and random features. *Quantum Mach. Intell.* **2022**, *4*, 23.
51. Zhang, C.; Li, G.; Qi, Y.; Wang, S.; Qing, L.; Huang, Q.; Yang, M.-H. Exploiting completeness and uncertainty of pseudo labels for weakly supervised video anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16271–16280.
52. Li, S.; Liu, F.; Jiao, L. Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 1395–1403.
53. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv* **2018**, arXiv:1803.09179.
54. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
55. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397.
56. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
57. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 11–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–7.

58. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Singapore, 20–23 February 2019; pp. 6105–6114.
59. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
60. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
61. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.-G. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390.
62. Wang, C.; Deng, W. Representative forgery mining for fake face detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14923–14932.
63. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Computer Vision – ECCV 2020, 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 86–103.
64. Guo, Z.; Jia, Z.; Wang, L.; Wang, D.; Yang, G.; Kasabov, N. Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *19*, 401–413.
65. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.-Y.; He, R. Masked relation learning for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708.
66. Shao, R.; Wu, T.; Nie, L.; Liu, Z. Deepfake-adapter: Dual-level adapter for deepfake detection. *Int. J. Comput. Vis.* **2025**, *133*, 3613–3628.
67. Mao, M.; Yan, C.; Wang, J.; Yang, J. Leveraging Pixel Difference Feature for Deepfake Detection. *IEEE Trans. Emerg. Top. Comput. Intell.* **2025**, *9*, 3178–3188.
68. Qiu, X.; Miao, X.; Wan, F.; Duan, H.; Shah, T.; Ojha, V.; Long, Y.; Ranjan, R. D2Fusion: Dual-domain fusion with feature superposition for Deepfake detection. *Inf. Fusion* **2025**, *120*, 103087.
69. Shi, Z.; Chen, H.; Jia, Y.; Zhang, D.; Lu, W.; Yang, X. Customized Transformer Adapter with Frequency Masking for Deepfake Detection. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 5904–5918.
70. Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local relation learning for face forgery detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 19–21 May 2021; pp. 1081–1088.
71. Shi, L.; Zhang, J.; Ji, Z.; Bai, J.; Shan, S. Real face foundation representation learning for generalized deepfake detection. *Pattern Recognit.* **2025**, *161*, 111299.
72. Tchaptchet, E.; Tagne, E.F.; Acosta, J.; Rawat, D.B.; Kamhoua, C. The Eyes: A Source of Information for Detecting Deepfakes. *Information* **2025**, *16*, 371.
73. Chen, H.; Li, Y.; Lin, D.; Li, B.; Wu, J. Watching the big artifacts: Exposing deepfake videos via bi-granularity artifacts. *Pattern Recognit.* **2023**, *135*, 109179.
74. Belguesmia, S.; Allili, M.S.; Hamadene, A. Unmasking Facial DeepFakes: A Robust Multiview Detection Framework for Natural Images. *arXiv* **2025**, arXiv:2510.15576.
75. Ma, L.; Yan, Z.; Xu, J.; Chen, Y.; Guo, Q.; Bi, Z.; Liao, Y.; Lin, H. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *arXiv* **2025**, arXiv:2504.04827.
76. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
77. Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3202–3211.
78. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; LeCun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6450–6459.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.