Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ECDF-NET: AN EXPLAINABLE CONTINUAL DEEPFAKE FORENSIC NETWORK WITH ADAPTIVE SPATIAL-FREQUENCY FUSION

G.V. ASHRITHA*[1], Dr. ANNALAKSHMI GOVINDARAJ[2]

[1*]Research Scholar, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad -500075, Telangana, India.

[2]Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad -500075, Telangana, India

E-mail: [1*]julieashritha@gmail.com, [2]annalakshmi@klh.edu.in

ABSTRACT

The rapid evolution of DeepFake and diffusion-based face generation methods has made the reliable forensic analysis of facial media increasingly difficult, because detectors trained on known manipulations degrade under cross-domain distribution shifts, common media transformations, and newly emerging forgery techniques. Existing studies address spatial-frequency learning, continual adaptation, forgery localization, and explainability largely in isolation, and rarely within a single model. This paper addresses this gap by proposing the Explainable Continual DeepFake Forensic Network (ECDF-Net). ECDF-Net adaptively fuses spatial and frequency-domain features to capture both visual and spectral manipulation artifacts, maintains a continual prototype memory that is updated only by reliable samples so that new forgery characteristics are learned while previous knowledge is retained, and jointly produces a detection decision, a forgery localization map, and multi-source forensic evidence. Robustness, consistency, and explanation consistency learning are further introduced to stabilize predictions and explanations under common media transformations. The framework is evaluated on FaceForensics++, while Celeb-DF v2 is used for cross-dataset evaluation and continual adaptation. On FaceForensics++, ECDF-Net achieves 98.42% accuracy, a 98.39% F1-score, and an AUC of 0.993. On Celeb-DF v2, it achieves 93.76% accuracy in direct cross-dataset evaluation, which rises to 97.21% after continual adaptation, while accuracy on FaceForensics++ decreases by only 0.46 percentage points. The ablation study confirms the contribution of adaptive fusion, prototype memory, robustness consistency, and explanation consistency learning. The principal contribution of this work is new evidence that detection, continual adaptation, localization, and stable explanation can be achieved from a single spatial-frequency representation, with reliability-gated prototype updates reducing the cross-dataset accuracy gap from 4.66 to 0.75 percentage points without retraining from scratch.

Keywords: *DeepFake detection, spatial-frequency learning, adaptive feature fusion, continual learning, prototype memory, forgery localization, explainable artificial intelligence, forensic analysis.*

1. INTRODUCTION

The rapid advancement of generative artificial intelligence has enabled the creation of highly realistic manipulated images and videos, commonly referred to as DeepFakes. Although these technologies have beneficial applications in entertainment, education, and digital media, their misuse presents serious concerns related to misinformation, identity manipulation, privacy violations, and the authenticity of digital content. As DeepFake generation methods become increasingly sophisticated, detecting manipulated media has become a major challenge in digital forensics. Conventional detection approaches often depend on manipulation-specific artifacts, which can limit their effectiveness when encountering unseen generation techniques, datasets, or real-world media variations [1] [2].

Recent research has increasingly focused on improving the generalization capability of DeepFake detection systems. Frequency-aware learning has been explored to capture manipulation traces that are less dependent on specific visual content and more resistant to variations across forgery methods [1]. Similarly, joint reconstruction and classification strategies have been introduced to learn complementary representations for distinguishing authentic and manipulated facial

content [2]. Multi-scale and dual-stream architectures have also been investigated to jointly exploit spatial and frequency-domain information, enabling the detection of subtle and diverse manipulation artifacts [3]. These developments demonstrate the importance of learning complementary forensic representations for improving DeepFake detection performance.

DeepFake detection has also evolved beyond conventional spatial feature analysis toward more comprehensive representation learning. Spatiotemporal models have been developed to capture inconsistencies in both visual appearance and temporal motion, improving the detection of manipulated video content [4]. Other approaches model the intrinsic characteristics of authentic facial data and identify manipulated samples based on deviations from learned real-face representations [5]. Self-contrastive learning has further been explored to reduce overfitting to manipulation-specific patterns and improve generalization across different forgery techniques [6]. These studies indicate that robust DeepFake detection requires feature representations that are less dependent on the characteristics of individual datasets or generation methods.

Cross-domain generalization remains another important research challenge. Several approaches have incorporated facial landmarks, adaptive feature selection, pre-trained vision-language representations, and spatial attention mechanisms to improve the ability of DeepFake detectors to identify previously unseen manipulations [7]-[10]. Continual learning has also emerged as an important direction because DeepFake generation techniques continuously evolve. Instead of repeatedly training a model from scratch, continual detection frameworks aim to adapt to newly emerging forgery types while minimizing catastrophic forgetting of previously acquired knowledge [11]. Such adaptation is essential for maintaining the long-term effectiveness of forensic systems in dynamic real-world environments.

In addition, the integration of spatial and frequency-domain information has shown considerable potential for capturing complementary forgery evidence. Spatial representations can identify visible inconsistencies related to facial structure, texture, blending, and semantic appearance, whereas frequency-domain features can reveal subtle spectral artifacts introduced during the generation and manipulation processes [3], [12]. Attention-based feature integration further improves the ability of a model to emphasize informative forensic patterns while suppressing irrelevant information [12]. However, many existing approaches focus primarily on detection accuracy or generalization and address spatial-frequency learning, continual adaptation, forgery localization, robustness, and explainability as separate research problems.

Despite the significant progress achieved in recent DeepFake detection research, several limitations remain. First, static detection models may struggle to adapt effectively to newly emerging manipulation techniques. Second, continual learning approaches may address knowledge retention but do not always provide comprehensive spatial-frequency forensic analysis. Third, many detection systems provide only a binary classification result without explicitly localizing manipulated regions. Furthermore, although explainable DeepFake detection has received increasing attention, the stability and consistency of generated explanations under different input transformations remain insufficiently explored. These limitations highlight the need for a unified forensic framework that can jointly support accurate detection, cross-dataset generalization, continual adaptation, knowledge retention, forgery localization, robustness, and consistent explanations.

The core concern of this research, namely the joint treatment of evolving forgeries and verifiable forensic evidence rather than detection accuracy alone, was selected for three reasons. First, new generation techniques appear faster than detectors can be redesigned, and retraining a detector from scratch whenever a new generator emerges is computationally expensive and requires continued access to all earlier training data, which may be restricted for privacy or storage reasons [11], [30]. A detector that adapts incrementally while retaining earlier knowledge therefore addresses a practical rather than a purely academic need. Second, in forensic and legal settings, a binary label is rarely sufficient: investigators must also know where a manipulation is located and why a decision was made, and this evidence should not change when the same media file is re-compressed or resized during sharing [24], [32]. Third, the reviewed literature indicates that spatial and frequency cues fail under different conditions. Spatial cues capture visible blending and texture inconsistencies but can be weakened by blurring and compression, whereas frequency cues capture spectral traces of synthesis and up-sampling that are

invisible in the pixel domain but can be altered by resizing [3], [12], [27]. Relying on either cue alone leaves a detector exposed to the conditions under which that cue fails, which justifies an adaptive combination of the two.

The choice of datasets follows the same reasoning. FaceForensics++ [33] was selected as the primary dataset because it contains several manipulation methods and provides manipulation masks, which allows detection and localization to be learned and assessed within one benchmark. Celeb-DF v2 [34] was selected as the unseen domain because its forgeries are of higher visual quality and contain fewer visible artifacts; moving from FaceForensics++ to Celeb-DF v2 therefore represents the kind of realistic distribution shift that a deployed detector faces when a new generation method appears.

To address these challenges, this study proposes the Explainable Continual DeepFake Forensic Network (ECDF-Net). The proposed framework integrates spatial and frequency-domain feature extraction through an adaptive feature fusion mechanism to capture complementary manipulation artifacts. A reliability-aware prototype memory is incorporated to facilitate continual adaptation to newly encountered forgery patterns while preserving representative knowledge acquired from previous domains. In addition to DeepFake classification, the framework generates forgery localization maps and multiple forms of forensic evidence to support the final decision. Robustness consistency learning and explanation consistency learning are further incorporated to improve the stability of predictions and explanations under common image transformations.

The proposed ECDF-Net is evaluated using FaceForensics++ for primary detection, component analysis, forgery localization, and robustness experiments, while Celeb-DF v2 is employed to examine cross-dataset generalization and continual adaptation.

The main contributions of this work are summarized as follows:

1. **A unified ECDF-Net framework** that integrates spatial and frequency-domain feature learning through adaptive fusion for comprehensive DeepFake forensic representation.
2. **A continual prototype memory mechanism** with reliability-aware updates to support adaptation to emerging forgery techniques while preserving previously acquired knowledge.
3. **Joint DeepFake detection and forgery localization**, enabling both classification and identification of suspicious manipulated regions.
4. **Robustness and explanation consistency learning** to improve the stability of predictions and forensic evidence under common media transformations.

Beyond these architectural contributions, the study generates new knowledge on how continual adaptation interacts with spatial-frequency forensic representations. It shows empirically that a reliability-gated prototype memory built on a fused spatial-frequency representation can close most of the cross-dataset gap, reducing it from 4.66 to 0.75 percentage points, while limiting forgetting on the source domain to less than 0.5 percentage points, and that consistency constraints on predictions and explanations improve, rather than reduce, detection accuracy.

The remainder of this paper is organized as follows. Section 2 reviews related work and presents the research gap, problem statement, hypotheses, and objectives. Section 3 describes the proposed ECDF-Net. Section 4 presents the experimental results and discussion. Section 5 explains how this work differs from prior studies and analyzes its advantages and disadvantages. Section 6 discusses the limitations of the study, and Section 7 concludes the paper by revisiting the hypotheses in light of the results and outlining future work.

2. RELATED WORK

The rapid advancement of DeepFake generation technologies has increased the need for reliable forensic methods capable of identifying manipulated visual content under diverse and evolving conditions. Recent studies have explored DeepFake detection from several perspectives, including hybrid deep learning, spatial-frequency feature analysis, multi-domain representation learning, domain generalization, robustness enhancement, continual adaptation, forgery localization, and explainable artificial intelligence. However, the continuous emergence of new manipulation techniques and variations in data distributions make it difficult for conventional static detection models to maintain reliable performance.

Sardhara *et al.* (2026) [13] proposed DeepForgeryNet, a hybrid CNN-LSTM and transfer learning framework for robust image forgery and DeepFake detection. The study combines convolutional feature extraction with sequential learning to capture complementary forgery characteristics. Although the framework improves the robustness of DeepFake detection, it primarily focuses on hybrid feature learning and does not explicitly address continual adaptation to emerging manipulation techniques.

Aljabri (2026) [14] introduced a DeepFake detection approach that integrates enhanced feature fusion, an adaptive hidden semi-Markov model, and deep neural networks for IoT-based edge computing environments. The study emphasizes adaptive feature integration and robust detection under resource-constrained settings. However, the approach does not specifically incorporate spatial-frequency forensic representation, forgery localization, or continual knowledge retention.

Alshammari and Alshammari (2026) [15] proposed DeepLayer-ID, a lightweight multi-domain forensic framework for real-time DeepFake detection on resource-constrained UAV sensor platforms. Their work focuses on balancing forensic detection capability with computational efficiency. While the framework supports practical deployment, continual learning and consistency-based explainability are not explicitly addressed.

Potluri *et al.* (2026) [16] developed a hybrid deep feature integration framework using transfer-learned neural networks for robust DeepFake detection. The model integrates complementary deep features to strengthen the distinction between authentic and manipulated content. However, the method mainly concentrates on robust classification and does not jointly address evolving forgery patterns, localization, and explanation consistency.

Xu *et al.* (2026) [17] presented a continual face forgery detection approach based on relation-aware spatial-frequency interaction aggregation and contrastive learning. The study addresses the challenge of adapting DeepFake detectors to newly emerging manipulation techniques while preserving previously acquired knowledge. Although it contributes to continual adaptation and spatial-frequency representation learning, the joint integration of continual learning with forgery localization, robustness consistency, and explainability remains limited.

Yang *et al.* (2026) [18] proposed SFD-ViT, a spatial-frequency dual-stream Transformer with patch embedding for face forgery detection. The framework simultaneously learns spatial and frequency-domain features to capture complementary forgery artifacts. By combining dual-stream representation learning with Transformer-based modeling, the method improves the ability to identify subtle manipulation traces. However, the approach mainly focuses on detection and does not explicitly address continual adaptation, forgery localization, or consistency-based explainability.

Kumar *et al.* (2026) [19] introduced a hybrid spatial-frequency attention-based approach using EfficientNet for robust and interpretable DeepFake detection. The method combines spatial and frequency-domain information with an attention mechanism to improve discriminative feature learning and provide interpretable detection results. Although the framework addresses robustness and interpretability, it does not incorporate continual learning or explicit knowledge retention for adapting to evolving forgery techniques.

Kou *et al.* (2025) [20] presented SFIAD, which integrates spatial-frequency features with dynamic margin optimization for DeepFake detection. The approach enhances feature discrimination by exploiting complementary information from both domains and optimizing the decision boundary. However, the method is primarily designed for classification and does not jointly address localization, continual adaptation, and explanation consistency.

Zheng *et al.* (2025) [21] proposed a spatio-frequency cross-fusion model for both DeepFake detection and segmentation. The model exploits interactions between spatial and frequency-domain features to identify manipulated content while also localizing forged regions. This work is particularly relevant to multi-task forensic analysis. However, continual adaptation and the preservation of knowledge from previously learned forgery types are not explicitly considered.

Li *et al.* (2025) [22] developed a DeepFake detection approach based on domain generalization and mask-guided supervision. The framework aims to improve generalization across different manipulation domains while using mask information to guide the learning of forgery-related features. Although the method addresses cross-domain robustness and manipulation-aware

learning, it does not explicitly incorporate continual learning or explanation consistency.

Lai *et al.* (2025) [23] proposed a high-frequency diversified augmentation strategy for general DeepFake detection. Their approach improves the model's ability to learn generalized forgery representations by exposing it to diverse high-frequency variations during training. While this method enhances robustness against distribution shifts and unseen manipulations, it does not address forgery localization or continual knowledge retention.

Al-Imran *et al.* (2025) [24] introduced DeFaX, a cross-attention fusion framework for robust and explainable DeepFake detection. The method uses cross-attention to integrate complementary feature representations while generating interpretable evidence for model predictions. This work demonstrates the importance of combining robust detection with explainability. However, continual adaptation and knowledge retention across emerging DeepFake generation techniques remain outside its primary scope.

Zhao *et al.* (2025) [25] proposed a multi-definition DeepFake detection framework based on semantics reduction and cross-domain training. The method aims to reduce dependence on manipulation-specific semantic information and improve generalization across different DeepFake domains. Although the framework strengthens cross-domain detection, it does not explicitly combine continual adaptation, localization, and explainability within a unified forensic framework.

Han *et al.* (2025) [26] proposed HSFF-Net, a hierarchical spectral-feature fusion network for DeepFake detection and localization. The framework learns hierarchical spectral representations and integrates them for both classification and manipulation-region localization. While the method supports joint detection and localization, it does not explicitly address continual adaptation to emerging forgery techniques or consistency in forensic explanations.

Wang *et al.* (2024) [27] proposed an adaptive spatial-frequency feature fusion framework for DeepFake detection. The method combines spatial and frequency-domain information to capture complementary forgery artifacts and employs adaptive fusion to strengthen discriminative feature representation. This study highlights the importance of jointly exploiting visual and frequency cues. However, the framework primarily focuses on static detection and does not explicitly address continual adaptation, forgery localization, or explanation consistency.

Wang *et al.* (2024) [28] introduced a spatial-frequency feature fusion approach enhanced through knowledge distillation. The method transfers complementary forensic knowledge to improve DeepFake detection and strengthen robustness against variations in manipulated content. Although the approach improves feature learning and detection robustness, it does not jointly incorporate continual learning, explicit localization of manipulated regions, and explainable forensic evidence.

Zhai *et al.* (2024) [29] proposed a spatial-frequency interaction framework for generalizable DeepFake detection. The approach models the relationship between spatial and frequency-domain features to reduce dependence on manipulation-specific artifacts and improve generalization across different forgery domains. While the method effectively addresses cross-domain generalization, continual adaptation and the integration of localization and explanation consistency are not explicitly considered.

Tassone *et al.* (2024) [30] investigated continuous fake media detection to adapt DeepFake detectors to newly emerging generative techniques. The study addresses the evolving nature of manipulated media by focusing on continual adaptation while reducing the loss of previously acquired knowledge. Although this work provides an important direction for handling sequentially emerging forgery techniques, it does not integrate adaptive spatial-frequency fusion, forgery localization, and explanation consistency within a unified framework.

Amin *et al.* (2024) [31] developed a multi-domain Transformer-based approach using cross-domain local characteristic analysis for DeepFake detection. The method extracts complementary characteristics from different domains to improve the detection of manipulation traces and enhance generalization. However, the framework primarily concentrates on multi-domain representation learning and does not explicitly incorporate continual knowledge retention or consistency-based explainability.

Yang *et al.* (2024) [32] presented SNIPPET, a framework for the subjective evaluation of visual explanations in DeepFake detection. The study emphasizes the importance of assessing the quality, relevance, and reliability of visual explanations generated by DeepFake

detection systems. While it contributes to the understanding of explainability in forensic analysis, it focuses on evaluating explanations rather than jointly integrating explanation consistency with adaptive detection, localization, and continual learning.

2.1. Research Gap

The review of recent DeepFake detection studies reveals several important research gaps. First, although many approaches exploit spatial and frequency-domain information, most focus primarily on improving classification or feature discrimination. The integration of adaptive spatial-frequency learning with other forensic tasks, such as localization and continual adaptation, remains limited.

Second, existing continual learning approaches address adaptation to newly emerging forgery techniques and knowledge preservation, but they are generally not integrated with comprehensive spatial-frequency feature fusion, robustness consistency, and explainable forensic analysis. Consequently, there is still a need for a framework that can adapt to evolving manipulation patterns while maintaining reliable knowledge of previously encountered forgeries.

Third, forgery localization and explainability are often treated as separate components. Existing methods either identify manipulated regions or generate visual explanations, but the consistency and reliability of these outputs under different transformations or changing data conditions are not sufficiently addressed. This limits the forensic reliability of model predictions in practical scenarios.

Finally, although domain generalization and augmentation strategies improve robustness to unseen manipulations and distribution shifts, most methods remain focused on static training environments. A unified mechanism that combines adaptive spatial-frequency feature fusion, continual knowledge retention, robustness-oriented consistency learning, forgery localization, and explanation consistency is still insufficiently explored.

Based on these limitations, the proposed ECDF-Net is designed to address this research gap through a unified forensic framework. It integrates spatial and frequency-domain feature extraction with adaptive fusion to construct a comprehensive forensic representation, while incorporating continual forgery prototype memory to support adaptation to emerging manipulation techniques. In addition, the framework jointly supports DeepFake detection, forgery localization, robustness consistency, and explainability with forensic evidence, thereby providing a more comprehensive approach to DeepFake analysis under evolving and diverse manipulation scenarios.

2.2. Problem Statement

The literature reviewed above shows that progress has been made along four largely separate lines. Spatial-frequency methods [18]-[20], [27]-[29] improve discrimination but are trained once and remain static. Continual detectors [11], [17], [30] adapt to new forgeries but provide neither localization nor explanation. Detection-localization models [21], [26] indicate manipulated regions but are not designed to adapt to new domains, and explainable detectors [19], [24] provide visual evidence without ensuring that this evidence remains stable when the input is compressed, blurred, or resized, a concern also raised by explanation-evaluation studies [32]. Consequently, a practitioner who needs a detector that remains accurate as new forgeries emerge and that also supports its decisions with stable, localized evidence must currently combine several independently trained models whose outputs are not guaranteed to agree.

Problem statement: *Given a DeepFake detector trained on a source domain of known manipulations (FaceForensics++), how can it be adapted to a previously unseen DeepFake domain (Celeb-DF v2) without retraining from scratch and without substantial loss of source-domain accuracy, while simultaneously localizing manipulated regions and producing explanations that remain consistent under common media transformations?*

This study is required because each of these requirements has so far been addressed in isolation, whereas a deployed forensic system must satisfy them together. Without continual adaptation, accuracy falls when a new generator appears; without localization and explanation, a correct prediction cannot be verified by an investigator; and without consistency under transformations, the same video can receive different evidence after it is re-shared. Determining whether a single representation can serve all of these requirements, and at what computational cost, is the question that motivates the present work.

2.3. Research Hypotheses and Objectives

Based on this problem statement, the following hypotheses are formulated and tested in this study:

H1: Adaptive fusion of spatial and frequency-domain features yields higher DeepFake detection performance than either feature domain alone or their non-adaptive combination.

H2: A continual prototype memory updated only with reliable samples allows the detector to adapt to an unseen DeepFake domain and substantially reduce the cross-dataset performance gap, while limiting the loss of accuracy on the source domain (catastrophic forgetting).

H3: Learning forgery localization jointly with detection on the shared fused representation yields more accurate localization of manipulated regions than CNN-based, attention-based, and spatial-frequency localization approaches.

H4: Robustness consistency and explanation consistency learning make predictions and explanations more stable under common media transformations without reducing detection accuracy.

Correspondingly, the objectives of this study are (O1) to design an adaptive spatial-frequency fusion mechanism and verify its contribution to detection (H1); (O2) to develop a reliability-aware continual prototype memory and assess cross-dataset adaptation and knowledge retention (H2); (O3) to integrate forgery localization into the same framework and evaluate it using IoU and localization accuracy (H3); and (O4) to introduce and evaluate consistency learning for robust predictions and stable explanations (H4). The achievement of these objectives is assessed in Section 5.5, and each hypothesis is revisited in Section 7 in light of the experimental evidence.

3. PROPOSED METHOD

This study proposes an Explainable Continual DeepFake Forensic Network (ECDF-Net) for robust and adaptive analysis of manipulated facial media. Conventional DeepFake detection systems are generally designed as static binary classifiers that distinguish authentic media from manipulated media based on patterns learned during training. However, their performance may decrease when the test data originate from a different manipulation domain or when new DeepFake generation characteristics are encountered.

To address these limitations, ECDF-Net integrates spatial-frequency feature learning, adaptive feature fusion, prototype-guided continual learning, multi-task forensic analysis, and explanation consistency learning within a unified framework. The proposed architecture is designed to perform the following tasks:

1. Binary classification of authentic and manipulated facial media.
2. Manipulation-aware forensic representation learning.
3. Localization of suspicious or manipulated facial regions.
4. Cross-dataset generalization analysis.
5. Continual adaptation to emerging DeepFake patterns.
6. Generation of stable and interpretable forensic evidence.

The methodology uses FaceForensics++ as the primary dataset for model development and evaluation, and Celeb-DF v2 as an independent dataset for cross-domain testing and continual adaptation.

The overall mapping of ECDF-Net can be expressed as

$$\mathcal{Y} = \Phi(\mathcal{X}; \Theta) \quad (1)$$

where $\mathcal{X}$ represents the input facial image or extracted video frame, Θ denotes the learnable parameters, and $\mathcal{Y}$ represents the final forensic output.

The output contains the DeepFake prediction, manipulation-related representation, localization information, prototype similarity, reliability score, and explanation.

3.1 Overall Architecture of ECDF-Net

Figure 1 presents the overall architecture of the proposed ECDF-Net, an integrated framework for robust DeepFake detection, forgery localization, continual adaptation, and explainable forensic analysis.

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

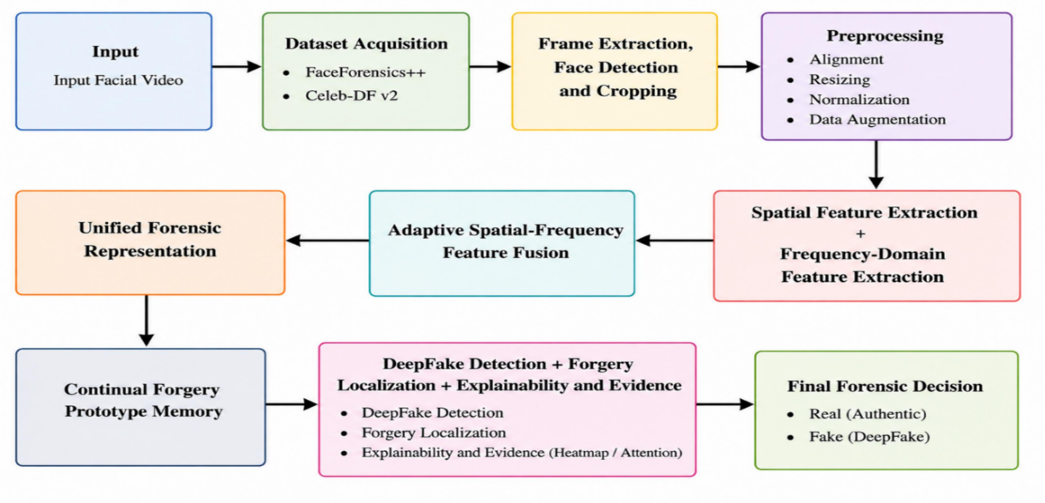


Figure 1: Proposed ECDF-Net Architecture

The proposed ECDF-Net follows a unified DeepFake forensic pipeline. Input facial videos are obtained from FaceForensics++ and Celeb-DF v2, followed by frame extraction, face detection, and cropping. The facial images then undergo alignment, resizing, normalization, and data augmentation.

Next, the framework performs parallel spatial feature extraction and frequency-domain feature extraction to capture visible and spectral manipulation artifacts. These complementary features are integrated through adaptive spatial-frequency feature fusion to generate a unified forensic representation.

The learned representation is maintained through continual forgery prototype memory, enabling the framework to retain previously learned patterns and support adaptation to new DeepFake characteristics. Finally, the representation is used for DeepFake detection, forgery localization, and explainability with forensic evidence generation, such as heatmaps or attention maps.

The architecture produces the final forensic decision by classifying the input as Real (Authentic) or Fake (DeepFake).

3.2 Dataset Description and Experimental Data Organization

The proposed ECDF-Net is evaluated using two complementary DeepFake benchmark datasets: FaceForensics++ [33] and Celeb-DF v2 [34]. FaceForensics++ is used as the primary dataset for model training, validation, and internal testing because it contains authentic and manipulated facial videos generated using multiple manipulation techniques. It supports the learning of spatial artifacts, facial inconsistencies, and frequency-domain forensic patterns.

Celeb-DF v2 is used as an independent dataset to evaluate the cross-dataset generalization and continual adaptation capability of ECDF-Net. Initially, the model trained on FaceForensics++ is directly evaluated on Celeb-DF v2 to assess its ability to generalize to an unseen DeepFake domain. Subsequently, Celeb-DF v2 samples are introduced sequentially for reliability-aware continual adaptation.

The complete experimental dataset is represented as

$$\mathcal{D} = \mathcal{D}FF \cup \mathcal{D}CD \qquad (2)$$

where $\mathcal{D}$FF denotes FaceForensics++ and $\mathcal{D}$CD denotes Celeb-DF v2. Each extracted facial sample is represented as

$$\mathcal{D} = \{(Xi, yi, ci)\}_{i=1}^{N} \qquad (3)$$

where $\mathcal{X}_i$ represents the input sample, y_i represents the authenticity label, and c_i represents the manipulation category when available.

3.3 Video Frame Extraction and Facial Region Preparation

Since both FaceForensics++ and Celeb-DF v2 consist of video-based facial media, representative frames are extracted from each video for subsequent analysis. Let an input video be represented as

$$\mathcal{V}_i = \{\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}T\} \qquad (4)$$

where T denotes the number of frames. A sampling function selects representative frames as

$$\mathcal{S}_i = \Psi(\mathcal{V}_i) \quad (5)$$

where Ψ(·) denotes the frame extraction operation. The selected frames are then processed using face detection to identify the primary facial region. Each detected face is cropped and aligned to reduce background interference and minimize variations caused by facial position and orientation.

The prepared facial sample is expressed as

$$\mathcal{X}^f_i = \Omega(\mathcal{X}_i) \quad (6)$$

where Ω (·) represents the face detection, cropping, and alignment operations. The resulting facial images are resized to a uniform resolution and forwarded to the subsequent preprocessing and feature extraction stages of ECDF-Net

3.4 Image Preprocessing and Normalization

The extracted facial regions are preprocessed to ensure consistent input quality and reduce variations caused by differences in illumination, scale, orientation, and pixel intensity. Each detected face is resized to a fixed spatial resolution and normalized before being supplied to the proposed ECDF-Net.

Let the aligned facial image be represented as $\mathcal{X}^a_i$. The normalized image is obtained as

$$\mathcal{X}^n_i = (\mathcal{X}^a_i - \mu) / \sigma \quad (7)$$

where μ and σ denote the mean and standard deviation of the pixel intensity distribution, respectively.

This preprocessing stage standardizes the input facial samples, reduces unwanted visual variations, and provides consistent representations for the subsequent spatial and frequency-domain feature extraction modules.

3.5 Data Augmentation and Robustness-Oriented Learning

To improve the robustness of ECDF-Net against variations commonly introduced during video compression, transmission, and post-processing, data augmentation is applied to the training samples. The augmentation process generates diverse versions of the input facial images while preserving their original authenticity labels.

Let the set of augmentation transformations be defined as

$$T = \{T_1, T_2, \ldots . T_K\} \quad (8)$$

where K denotes the number of applied transformations. For an input image X^n_i an augmented sample is generated as

$$\mathcal{X}^t_i = \tau(\mathcal{X}^n_i) \quad (9)$$

The augmentation operations include controlled compression, Gaussian noise, blurring, resolution variation, and brightness and contrast adjustments. These transformations simulate common real-world distortions and encourage the model to learn stable forensic representations rather than relying on specific visual conditions.

By exposing ECDF-Net to diverse input variations during training, the proposed robustness-oriented learning strategy improves its ability to maintain consistent DeepFake detection performance under degraded and previously unseen media conditions.

3.6 Spatial Feature Extraction

The spatial feature extraction module is designed to learn discriminative visual patterns associated with authentic and manipulated facial images. The normalized facial image is processed through a deep feature encoder to capture local texture variations, facial boundary inconsistencies, blending artifacts, and structural irregularities.

Let the spatial feature extractor be denoted by $\mathcal{E}_s$ The spatial representation is obtained as

$$\mathcal{F}_s = \mathcal{E}_s(\mathcal{X}^n_i) \quad (10)$$

where $\mathcal{X}^n_i$ represents the normalized input facial image and $\mathcal{F}_s$denotes the extracted spatial feature representation.

To capture forensic evidence at different levels of abstraction, features from multiple stages of the encoder are aggregated as

$$\mathcal{F}_s = \mathcal{A}(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}L) \quad (11)$$

where $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}L$ represent multi-level feature maps and $\mathcal{A}(\cdot)$ denotes the feature aggregation operation.

The resulting spatial representation preserves both low-level forensic artifacts and higher-level facial structural information. These features are subsequently combined with frequency-domain features through the adaptive feature fusion module to generate a comprehensive representation for DeepFake detection.

3.7 Frequency-Domain Feature Extraction

Although manipulated facial images may appear visually realistic, DeepFake generation can introduce subtle inconsistencies in the frequency domain that are not easily captured through spatial analysis alone. Therefore, ECDF-Net incorporates a frequency-domain feature extraction module to identify spectral patterns associated with synthetic image generation and facial manipulation.

For a normalized facial image $\mathcal{X}^{n}_{i}$, the frequency representation is obtained using a Fourier transformation:

$$\mathcal{Z}_f = \mathcal{F}(\mathcal{X}^{n}_{i}) \quad (12)$$

where F(·) denotes the Fourier transformation and $\mathcal{Z}_f$ represents the resulting frequency-domain representation. The corresponding magnitude spectrum is calculated as

$$\mathcal{A}_f = |\mathcal{Z}_f| \quad (13)$$

The magnitude information is subsequently processed by a frequency feature encoder $\mathcal{E}f$ to obtain discriminative spectral features:

$$\boldsymbol{\mathcal{F}}_f = \boldsymbol{\mathcal{E}}_f(\mathcal{A}_f) \quad (14)$$

where $\boldsymbol{\mathcal{F}}_f$ represents the extracted frequency-domain feature representation.

This module enables ECDF-Net to capture complementary forensic evidence related to spectral irregularities and synthesis artifacts. The extracted frequency features are subsequently combined with the spatial feature representation through the adaptive spatial-frequency feature fusion module.

3.8 Adaptive Spatial-Frequency Feature Fusion

The spatial and frequency-domain branches provide complementary forensic information. Spatial features capture visible artifacts and structural inconsistencies, whereas frequency features identify spectral irregularities associated with synthetic image generation. To effectively integrate these complementary representations, ECDF-Net employs an adaptive spatial-frequency feature fusion mechanism.

Let $\boldsymbol{\mathcal{F}}s$ and $\boldsymbol{\mathcal{F}}f$ denote the spatial and frequency-domain feature representations, respectively. An adaptive weighting coefficient is computed as

$$\alpha = \sigma(\mathcal{W}s\boldsymbol{\mathcal{F}}s + \mathcal{W}f\boldsymbol{\mathcal{F}}f + b) \quad (15)$$

where$\mathcal{W}s$ and $\mathcal{W}f$ are learnable weight parameters, b represents the bias term, and $\sigma(\cdot)$ denotes the sigmoid activation function.

The unified forensic representation is then obtained as

$$\boldsymbol{\mathcal{F}}u = \alpha \odot \boldsymbol{\mathcal{F}}s + (1 - \alpha) \odot \boldsymbol{\mathcal{F}}(16)$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{\mathcal{F}}u$ represents the fused feature representation.

This adaptive fusion strategy enables ECDF-Net to dynamically adjust the contribution of spatial and frequency-domain information according to the characteristics of each input sample. The resulting unified representation is subsequently used for continual prototype learning, DeepFake classification, forgery localization, and explainability analysis.

3.9 Continual Forgery Prototype Learning

The continual forgery prototype learning module is designed to maintain representative feature patterns of authentic and manipulated facial media while supporting adaptation to previously unseen DeepFake characteristics. The fused representation obtained from the adaptive spatial-frequency feature fusion module is used to construct a prototype memory.

Let the prototype memory be represented as

$$\mathcal{P} = \{p_1, p_2, \ldots, pK\} \quad (17)$$

where K denotes the total number of stored prototypes, and pk represents the K-th prototype. Each prototype summarizes the characteristic feature distribution of a specific class or manipulation pattern.

For an input sample with fused representation $\boldsymbol{\mathcal{F}}u^{\mathrm{I}}$, its similarity to each prototype is calculated using cosine similarity:

$$s_{\mathrm{i}}k = (\boldsymbol{\mathcal{F}}u^{\mathrm{I}} \cdot pk) / (||\boldsymbol{\mathcal{F}}u^{\mathrm{I}}|| \, ||pk||)(18)$$

The most relevant prototype is identified as

$$k* = arg\ max(s_{\mathrm{i}}k) \quad (19)$$

Initially, the prototype memory is constructed using the feature representations learned from FaceForensics++. During continual adaptation, new samples from the unseen DeepFake domain are compared with the existing prototypes. Reliable samples can then contribute to updating or refining the prototype memory, allowing ECDF-Net to incorporate new manipulation characteristics while retaining previously learned forensic knowledge.

The resulting prototype-based representation supports the subsequent reliability-aware adaptation and helps reduce the dependence of the model on fixed DeepFake patterns learned from a single training domain.

3.10 Reliability-Aware Continual Adaptation

To enable reliable adaptation to previously unseen DeepFake patterns, ECDF-Net incorporates a reliability-aware continual adaptation mechanism. Rather than updating the learned prototype memory using every incoming sample, the model evaluates the reliability of each sample before allowing it to influence the adaptation process. This reduces the risk of incorporating incorrectly classified or uncertain samples into the learned forensic knowledge.

For an incoming sample, the reliability score is determined by combining prediction confidence, prototype similarity, and uncertainty information:

$$r_i = \lambda_1 q_i + \lambda_2 s_i + \lambda_3 u_i \quad (20)$$

where q_i denotes the prediction confidence, s_i represents the similarity between the input representation and the corresponding prototype, and u_i denotes the reliability derived from uncertainty estimation. The weighting coefficients satisfy

$$\lambda_1 + \lambda_2 + \lambda_3 = 1 \quad (21)$$

A sample is considered sufficiently reliable for continual adaptation when

$$r_i \geq \delta \quad (22)$$

where δ represents the predefined reliability threshold. Only samples satisfying this condition are used to update the corresponding prototype.

The prototype update is performed as

$$p_k^{new} = \eta p_k^{old} + (1-\eta) F_u^i \quad (23)$$

where p_k^{old} and p_k^{new} represent the prototype before and after adaptation, respectively; η is the update coefficient, and F_u^i denotes the fused feature representation of the reliable incoming sample.

This reliability-aware adaptation strategy enables ECDF-Net to progressively incorporate new DeepFake characteristics while reducing the influence of uncertain samples and preserving previously learned forensic knowledge.

3.11 DeepFake Detection Module

The DeepFake detection module performs the final classification of the input facial sample as either authentic or manipulated. It receives the unified forensic representation generated by the adaptive spatial-frequency feature fusion module and enhanced through continual prototype learning.

Let $\boldsymbol{\mathcal{F}}_u$ denote the final fused feature representation. The probability of an input sample being classified as manipulated is computed as

$$\hat{y}_i = \sigma(\mathcal{W}_d \boldsymbol{\mathcal{F}}_u + b_d) \quad (24)$$

where $\mathcal{W}_d$ and b_d represent the learnable weight and bias parameters of the detection layer, respectively, and $\sigma(\cdot)$ denotes the sigmoid activation function. The predicted probability $\hat{y}_i$ ranges from 0 to 1, indicating the likelihood that the input sample is manipulated.

The detection module is optimized using the binary cross-entropy loss:

$$\boldsymbol{\mathcal{L}}_d = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1 - y_i)\, log(1 - \hat{y}_i)] \quad (25)$$

where N denotes the number of samples, y_i represents the ground-truth authenticity label, and $\hat{y}_i$ denotes the predicted probability.

The final decision is obtained by applying a predefined classification threshold. Samples with predicted probabilities above the threshold are classified as manipulated, while those below the threshold are classified as authentic. This module enables ECDF-Net to use the learned spatial, frequency-domain, and prototype-based forensic representations for reliable DeepFake detection.

3.12 Forgery Localization Module

The forgery localization module is designed to identify the facial regions that contribute most to the DeepFake prediction. While the detection module determines whether an input sample is authentic or manipulated, the localization module provides spatial evidence by highlighting regions that are likely to contain manipulation-related artifacts.

The unified forensic representation $\boldsymbol{\mathcal{F}}u$ is processed by a localization function $G(\cdot)$ to generate a spatial forgery map:

$$\boldsymbol{\mathcal{M}}_i = \mathcal{G}(\boldsymbol{\mathcal{F}}u) \quad (26)$$

where $\boldsymbol{\mathcal{M}}_i$ denotes the predicted forgery localization map for the i-th input sample. The

generated map assigns higher responses to regions that contain stronger manipulation evidence.

When ground-truth localization masks are available, the localization branch is optimized using a pixel-level binary cross-entropy loss:

$$\mathcal{L}_m = -\frac{1}{P}\sum_{P-1}^{P}[m_p \log(\hat{m}_p) - (1-m_p)\log(1-\hat{m}_p)] \quad (27)$$

where P represents the total number of pixels, m_p denotes the ground-truth label of the p-th pixel, and $\hat{m}_p$ represents the corresponding predicted value.

The localization performance can be evaluated using the Intersection over Union metric:

$$IoU = \frac{|\mathcal{M} \cap \mathcal{M}|}{|\mathcal{M} \cup \mathcal{M}} \quad (28)$$

where $\mathcal{M}$ and $\mathcal{M}$ represent the ground-truth and predicted manipulation regions, respectively.

By providing a spatial representation of suspected manipulated areas, the forgery localization module complements the binary detection result and improves the interpretability of ECDF-Net by indicating where the potential manipulation evidence is located within the facial image.

3.13 Cross-Dataset Generalization Strategy

DeepFake detection models often exhibit reduced performance when evaluated on data generated using manipulation methods or data distributions that differ from those used during training. Therefore, ECDF-Net incorporates a cross-dataset generalization strategy to evaluate its ability to learn transferable forensic representations.

The model is initially trained and optimized using the FaceForensics++ dataset. Let the optimized model parameters be represented as

$$\Theta_{FF}^{*} = arg\ min\ \Theta\ \mathcal{L}(\mathcal{D}_{FF}) \quad (29)$$

where $\mathcal{D}_{FF}$ denotes the primary training dataset, Θ represents the learnable model parameters, and $L(\cdot)$ denotes the overall training objective.

After training, the learned model is directly evaluated on Celeb-DF v2 without incorporating samples from this dataset during the initial training stage. The cross-dataset prediction is expressed as

$$\hat{Y}_{CD} = \Phi(\mathcal{D}_{CD};\ \Theta_{FF}^{*}) \quad (30)$$

where $\mathcal{D}_{CD}$ represents the Celeb-DF v2 dataset and $\hat{Y}_{CD}$ denotes the corresponding predicted outputs.

To quantify the generalization gap, the difference between the performance obtained on the primary dataset and the independent dataset is calculated as

$$\Delta G = G_{FF} - G_{CD} \quad (31)$$

where G_{FF} and G_{CD} represent the selected performance measure obtained on FaceForensics++ and Celeb-DF v2, respectively. A smaller value of ΔG indicates stronger cross-dataset generalization.

This evaluation strategy enables a direct assessment of whether ECDF-Net learns general forensic characteristics that remain effective across different DeepFake data distributions, rather than relying primarily on dataset-specific artifacts.

3.14 Continual Adaptation Using Celeb-DF v2

Following the initial cross-dataset evaluation, Celeb-DF v2 is used to examine the ability of ECDF-Net to adapt to previously unseen DeepFake characteristics. Unlike conventional static learning, the proposed continual adaptation strategy allows the model to progressively incorporate new forensic patterns while preserving the knowledge learned from the FaceForensics++ dataset.

The incoming Celeb-DF v2 data are organized into a sequence of adaptation stages:

$$\mathcal{D}CD = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\} \quad (32)$$

where T denotes the total number of sequential adaptation stages. At each stage, the incoming samples are processed to obtain their fused forensic representations and corresponding reliability scores.

Only samples satisfying the predefined reliability criterion are used to update the prototype memory and adapt the model. The continual learning objective is expressed as

$$\mathcal{L}_c = \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f \quad (33)$$

where $\mathcal{L}_d$ represents the DeepFake detection loss, $\mathcal{L}_p$ denotes the prototype consistency loss, and $\mathcal{L}_f$ represents the knowledge preservation loss. The parameters λ_p and λ_f control the contribution of the corresponding objectives.

This continual adaptation strategy enables ECDF-Net to progressively learn new manipulation

characteristics from Celeb-DF v2 while reducing catastrophic forgetting of previously acquired forensic knowledge. The adapted model is subsequently evaluated on both the original and newly encountered data domains to assess its adaptation capability and knowledge retention.

3.15 Explainability and Consistency Learning

To improve the transparency and reliability of DeepFake detection, ECDF-Net incorporates an explainability and consistency learning mechanism. In addition to generating a binary prediction, the framework provides spatial evidence indicating the regions that contribute to the final decision. This allows the predicted outcome to be supported by interpretable forensic information.

Let the explanation generated for an input sample $\mathcal{X}_{\mathrm{i}}$ be represented as

$$\boldsymbol{\mathcal{E}}_{\mathrm{i}} = \Gamma(\mathcal{X}_{\mathrm{i}}) \quad (34)$$

where $\Gamma(\cdot)$ denotes the explanation generation function, and $\boldsymbol{\mathcal{E}}_{\mathrm{i}}$ represents the corresponding forensic explanation map.

To ensure that the explanation remains stable under moderate input variations, an augmented version of the same sample is generated as

$$\mathcal{X}^{\mathrm{t}}{}_{\mathrm{i}} = \tau(\mathcal{X}_{\mathrm{i}}) \quad (35)$$

where $\tau(\cdot)$ represents a transformation such as compression, noise, or image-quality variation. The explanation generated for the transformed sample is expressed as

$$\boldsymbol{\mathcal{E}}^{\mathrm{t}}{}_{\mathrm{i}} = \Gamma(\mathcal{X}^{\mathrm{t}}{}_{\mathrm{i}}) \quad (36)$$

The consistency between the explanations of the original and transformed samples is enforced using the explanation consistency loss:

$$\boldsymbol{\mathcal{L}}e = ||\boldsymbol{\mathcal{E}}_{\mathrm{i}} - \boldsymbol{\mathcal{E}}^{\mathrm{t}}{}_{\mathrm{i}}||^2{}_2 \quad (37)$$

Minimizing this loss encourages ECDF-Net to generate consistent forensic evidence when the input undergoes moderate transformations. This mechanism improves the reliability and interpretability of the proposed framework by ensuring that its explanations are less sensitive to minor variations in image quality or appearance.

3.16 Robustness Consistency Learning

DeepFake media may undergo compression, resizing, noise addition, or other quality degradations during storage and transmission. Such variations can alter the visual appearance of facial images and affect the reliability of forensic predictions. Therefore, ECDF-Net incorporates robustness consistency learning to encourage stable feature representations under moderate input transformations. For an input sample X_{i} and its transformed version $X_{\mathrm{i}}{}^{\mathrm{t}}$, the corresponding fused feature representations are denoted as

$$\mathrm{F_u}(\mathrm{X}_i) \text{ and } \mathrm{F_u}(X_i^t) \quad (38)$$

respectively. The robustness consistency loss is defined as

$$\boldsymbol{\mathcal{L}}_r = ||\,\mathrm{F_u}(\mathcal{X}_{\mathrm{i}}) - \mathrm{F_u}(\tau(X_i^t))||_2^2 \quad (39)$$

where $\boldsymbol{\mathcal{L}}_r$ measures the difference between the feature representations of the original and transformed samples.

By minimizing this loss, ECDF-Net is encouraged to preserve consistent forensic representations despite moderate variations in image quality. This robustness-oriented learning strategy reduces the sensitivity of the model to common real-world distortions and supports more reliable DeepFake detection across diverse media conditions.

3.17 Multi-Task Optimization

The proposed ECDF-Net integrates multiple learning objectives to jointly optimize DeepFake detection, forgery localization, continual adaptation, robustness, and explanation consistency. Rather than optimizing these components independently, a unified multi-task objective is employed to learn complementary forensic representations.

The overall loss function is defined as

$$L_{total} = \lambda_{\boldsymbol{d}} L_d + \lambda_{\boldsymbol{m}} L_m + \lambda_{\boldsymbol{c}} L_c + \lambda_{\boldsymbol{r}} L_r + \lambda_{\boldsymbol{e}} L_e \quad (40)$$

where L_d denotes the DeepFake detection loss, L_m represents the forgery localization loss, L_c denotes the continual adaptation loss, L_r represents the robustness consistency loss, and L_e denotes the explanation consistency loss. The weighting coefficients $\lambda_{\boldsymbol{d}}$, $\lambda_{\boldsymbol{m}}$, $\lambda_{\boldsymbol{c}}$, $\lambda_{\boldsymbol{r}}$, and $\lambda_{\boldsymbol{e}}$ control the relative contribution of each learning objective.

The model parameters are optimized by minimizing the combined objective:

$$\Theta^* = arg\ min\ \Theta\ L_{Total} \quad (41)$$

where Θ^* represents the optimized parameters of ECDF-Net.

This unified optimization strategy enables the model to jointly learn discriminative, robust, and interpretable forensic representations while supporting adaptation to previously unseen DeepFake patterns.

3.18 Algorithm of the Proposed ECDF-Net

Algorithm 1: ECDF-Net for Adaptive DeepFake Detection, Localization, and Explainability

Input: Facial video/image dataset D containing authentic and manipulated samples

Output: Final forensic decision, manipulated region localization, and explainable evidence

Step 1: Acquire authentic and manipulated videos from FaceForensics++ and Celeb-DF v2.
Step 2: Extract representative frames from each video.
Step 3: Detect, crop, align, resize, and normalize facial regions.
Step 4: Apply robustness-oriented augmentation to training samples.
Step 5: Train the spatial feature extraction branch using FaceForensics++.
Step 6: Generate frequency-domain representations and extract spectral features.
Step 7: Perform adaptive fusion of spatial and frequency features.
Step 8: Construct the initial prototype memory using FaceForensics++.
Step 9: Generate DeepFake prediction and forgery localization map.
Step 10: Optimize the multi-task learning objective.
Step 11: Evaluate on the independent Celeb-DF v2 dataset.
Step 12: Calculate prediction confidence and prototype similarity.
Step 13: Estimate reliability and update prototype memory only when the threshold is satisfied.
Step 14: Preserve previous knowledge while adapting to Celeb-DF v2.
Step 15: Generate spatial, frequency, prototype, and localization evidence.
Step 16: Calculate robustness and explanation consistency.
Step 17: Output the final forensic decision.

3.19 Computational Complexity

The computational complexity of ECDF-Net is primarily determined by the spatial feature extraction, frequency-domain transformation, adaptive feature fusion, and prototype matching operations. Let N denote the number of input samples, D represent the dimensionality of the extracted feature representation, and K denote the number of prototypes maintained in the continual memory.

The spatial feature extraction complexity can be approximated as

$$C_{spatial} = O(ND) \quad (42)$$

The frequency-domain transformation, implemented using an efficient Fourier transform, has a complexity of

$$C_{frequency} = O(N \log N) \quad (43)$$

The adaptive fusion operation combines the spatial and frequency representations and introduces approximately linear complexity with respect to the feature dimension:

$$C_{\text{fusion}} = O(ND) \quad (44)$$

During continual learning, each input representation is compared with K stored prototypes. Therefore, the prototype matching complexity is

$$C_{prototype} = O(KD) \quad (45)$$

Accordingly, the overall computational complexity of the proposed framework can be expressed as

$$C_{total} = O(ND + N \log N + KD \quad (46)$$

Since the prototype memory contains only a limited number of representative embeddings, the additional computational cost of continual adaptation remains relatively low. The proposed architecture therefore provides a balance between robust forensic feature learning, continual adaptation, and computational efficiency, making it suitable for practical DeepFake detection under cross-domain conditions.

3.20 Evaluation Protocol

The proposed ECDF-Net was evaluated through multiple experiments to comprehensively assess its DeepFake detection capability. FaceForensics++ was used as the primary dataset for internal evaluation, ablation analysis, forgery localization, and robustness testing, while Celeb-DF v2 was employed for cross-dataset evaluation and continual adaptation analysis.

The evaluation protocol includes internal detection evaluation, ablation studies, forgery

localization analysis, robustness evaluation under media transformations, cross-dataset generalization, and continual adaptation. Internal evaluation measures standard detection performance, whereas the ablation study examines the contribution of individual ECDF-Net components. Localization analysis evaluates the ability to identify suspicious manipulated regions, while robustness experiments assess resistance to transformations and media degradation. Cross-dataset evaluation examines generalization from FaceForensics++ to Celeb-DF v2, and continual adaptation evaluates adaptation to a new forgery domain while retaining previously acquired knowledge.

For binary DeepFake detection, performance was assessed using accuracy, precision, recall, F1-score, specificity, and ROC-AUC, providing a comprehensive evaluation of classification effectiveness and discriminative capability. The primary metrics are

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (47)$$

$$Precision = TP / (TP + FP) \quad (48)$$

$$Recall = TP / (TP + FN) \quad (49)$$

$$F_1 = 2(Precision \times Recall) / (Precision + Recall) \quad (50)$$

4. RESULTS AND DISCUSSION

This section evaluates the proposed Explainable Continual DeepFake Forensic Network (ECDF-Net) in terms of DeepFake detection performance, cross-dataset generalization, continual adaptation, forgery localization, robustness, explainability consistency, and computational efficiency. The experiments were conducted using FaceForensics++ as the primary dataset and Celeb-DF v2 as the independent dataset for evaluating domain generalization and adaptation capability.

The evaluation employed standard metrics, including Accuracy, Precision, Recall, F1-score, and AUC, while localization performance was measured using Intersection over Union (IoU). The robustness and explanation modules were further assessed under common image transformations.

4.1 DeepFake Detection Performance

To evaluate the effectiveness of the proposed ECDF-Net, its DeepFake detection performance on the FaceForensics++ dataset was compared with representative CNN-based, transformer-based, and spatial-frequency methods, including XceptionNet, EfficientNet-B4, ResNet50, Vision Transformer (ViT-B/16), and the Spatial-Frequency Model. Table 1 presents the comparative results in terms of Accuracy, Precision, Recall, F1-score, and AUC, providing a comprehensive assessment of the detection accuracy and reliability of the proposed framework.

Table 1: DeepFake Detection Performance on FaceForensics++

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
XceptionNet [13]	96.8	96.5	96.9	96.7	0.97
EfficientNet-B4 [14]	97.65	97.68	97.68	97.65	0.977
ResNet50 [15]	90.9	91.5	89	90.2	0.942
Vision Transformer (ViT-B/16) [16]	91.6	90.8	91.2	91	0.87
Spatial-Frequency Model [27]	97.06	96.88	96.95	96.91	0.985
ECDF-Net (Proposed)	98.42	98.31	98.47	98.39	0.993

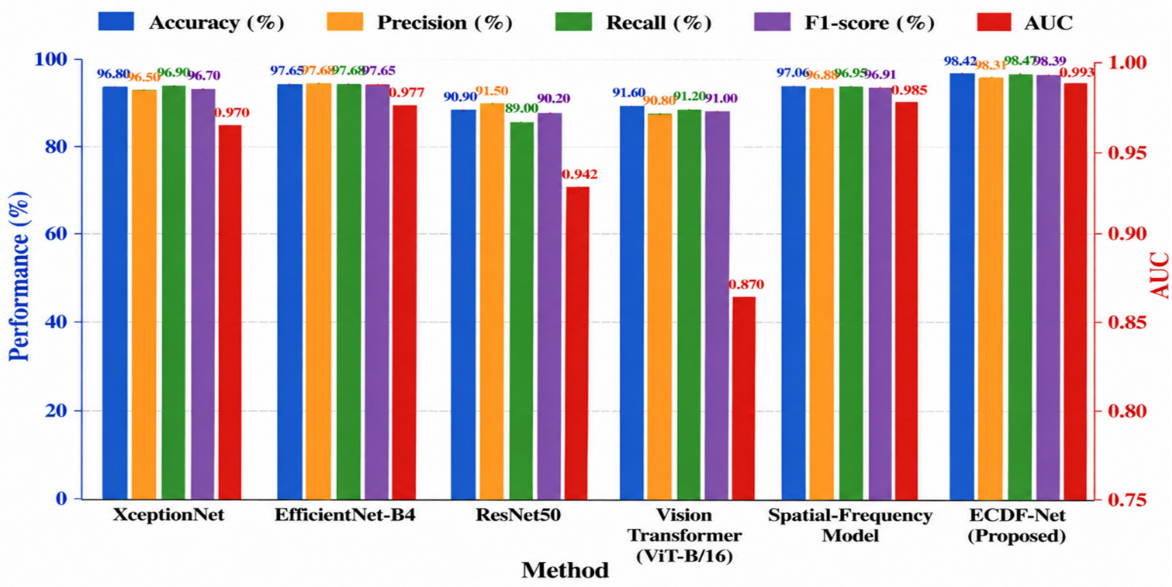


Figure 2: Comparative DeepFake Detection Performance on FaceForensics++

Figure 2 compares the DeepFake detection performance of the proposed ECDF-Net with XceptionNet, EfficientNet-B4, ResNet50, Vision Transformer (ViT-B/16), and the Spatial-Frequency Model using Accuracy, Precision, Recall, F1-score, and AUC. ECDF-Net achieves the best overall performance, obtaining 98.42% accuracy, 98.31% precision, 98.47% recall, 98.39% F1-score, and 0.993 AUC. The results demonstrate the effectiveness of ECDF-Net in providing accurate and reliable DeepFake detection through complementary spatial-frequency feature learning and its integrated forensic learning strategy.

4.2 Cross-Dataset Generalization

To assess the ability of the proposed ECDF-Net to generalize across different DeepFake data distributions, cross-dataset experiments were conducted using FaceForensics++ and Celeb-DF v2**.** The model was evaluated under both intra-dataset and cross-dataset settings. Table 2 presents the results in terms of Accuracy, Precision, Recall, F1-score, and AUC, providing an assessment of the model's capability to transfer learned forensic representations to previously unseen DeepFake data.

Table 2: Cross-Dataset Generalization Performance

Training Dataset	Testing Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC
FaceForensics++	FaceForensics++	98.42	98.31	98.47	98.39	0.993
FaceForensics++	Celeb-DF v2	93.76	93.41	93.92	93.66	0.962
Celeb-DF v2	Celeb-DF v2	97.18	96.94	97.06	97	0.984

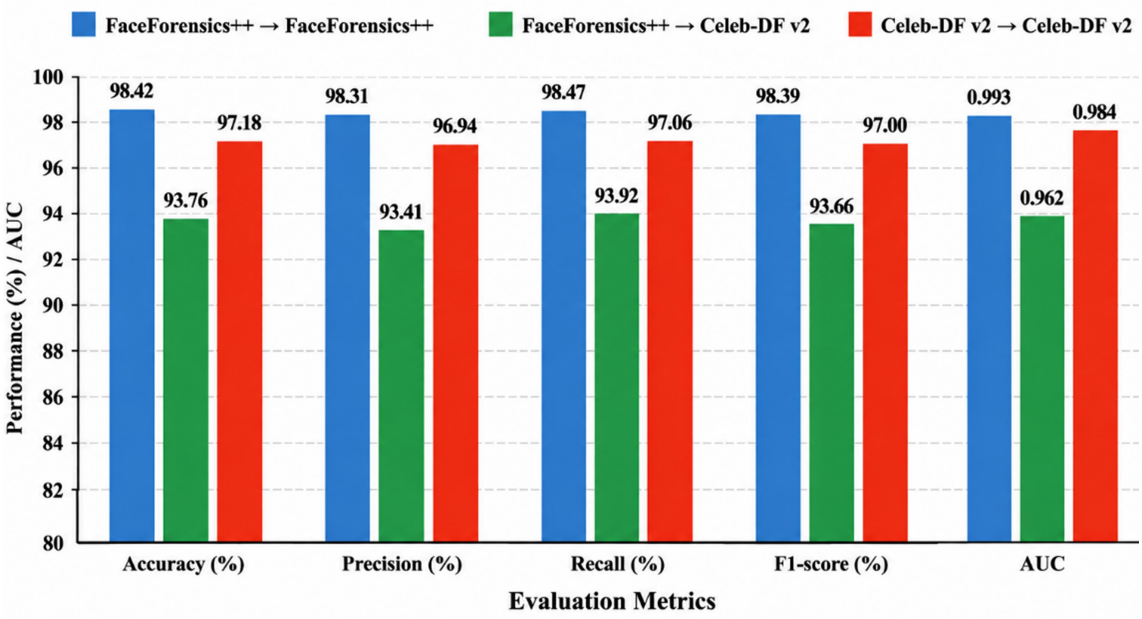


Figure 3: Cross-Dataset Generalization Performance of ECDF-Net

Figure 3 compares ECDF-Net under intra-dataset and cross-dataset evaluation settings using Accuracy, Precision, Recall, F1-score, and AUC. ECDF-Net achieved its highest performance when trained and tested on FaceForensics++, while direct testing on the unseen Celeb-DF v2 dataset resulted in a moderate performance reduction. Nevertheless, the model maintained 93.76% accuracy and 0.962 AUC, demonstrating effective cross-dataset generalization. When trained and tested on Celeb-DF v2, ECDF-Net achieved 97.18% accuracy and 0.984 AUC, further confirming its strong DeepFake detection capability across different data distributions.

4.3 Ablation Analysis

To analyze the contribution of each major component in the proposed framework, an ablation study was conducted by progressively adding the spatial-frequency feature extraction, adaptive feature fusion, prototype memory, robustness consistency learning, and explainability consistency learning modules. Table 3 presents the corresponding Accuracy, F1-score, and AUC values, demonstrating the individual and cumulative impact of each component on the overall performance of ECDF-Net.

Table 3: Ablation Study of ECDF-Net Components

Configuration	Accuracy (%)	F1-score (%)	AUC
Spatial Features Only	95.36	95.1	0.971
Frequency Features Only	94.82	94.56	0.966
Spatial + Frequency Features	96.74	96.58	0.983
+ Adaptive Feature Fusion	97.31	97.18	0.987
+ Prototype Memory	97.68	97.55	0.989
+ Robustness Consistency Learning	98.07	97.96	0.991
+ Explainability Consistency Learning	**98.42**	**98.39**	**0.993**

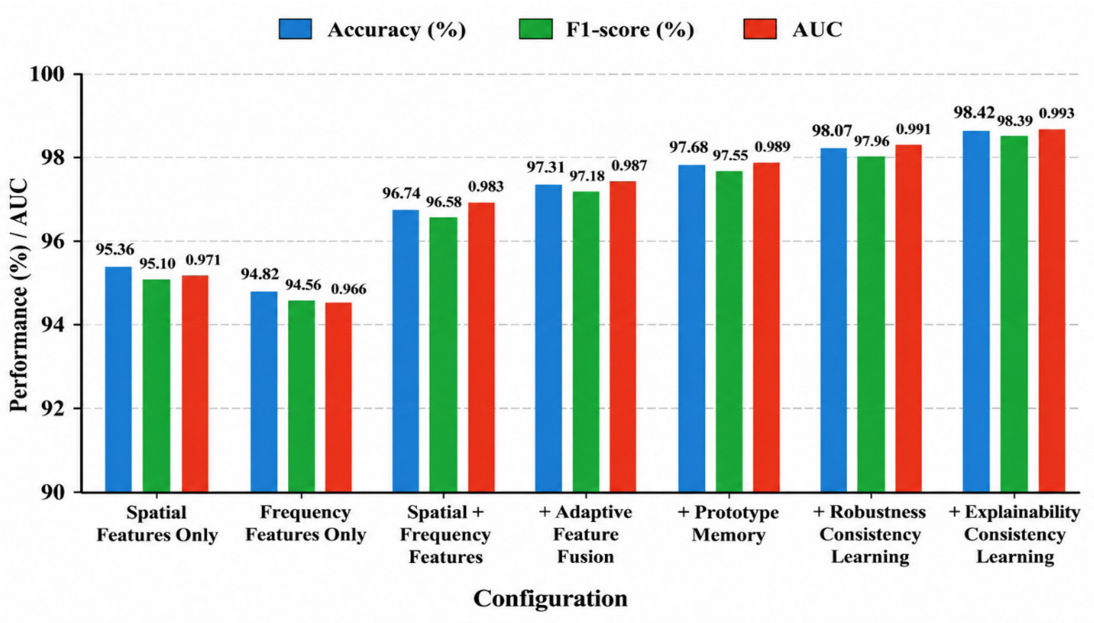


Figure 4: Ablation Study of ECDF-Net Components

Figure 4 presents the contribution of the major components of ECDF-Net in terms of Accuracy, F1-score, and AUC. The individual spatial and frequency feature branches provide strong baseline performance, while their combination improves the results, confirming the complementary value of spatial and frequency-domain information. The successive addition of adaptive feature fusion, prototype memory, robustness consistency learning, and explainability consistency learning further improves performance. The complete ECDF-Net achieves the best results, with 98.42% accuracy, 98.39% F1-score, and 0.993 AUC, demonstrating the positive contribution of each proposed component.

4.4 Continual Adaptation Performance

To evaluate the effectiveness of the proposed continual adaptation strategy, ECDF-Net was assessed at different adaptation stages using both FaceForensics++ and Celeb-DF v2. The evaluation examines whether the model can improve its performance on newly encountered DeepFake data while retaining previously learned forensic knowledge. Table 4 presents the accuracy obtained on both datasets during initial training, initial adaptation, continual adaptation, and the final adapted model stage.

Table 4: Continual Adaptation and Knowledge Retention

Evaluation Stage	FaceForensics++ Accuracy (%)	Celeb-DF v2 Accuracy (%)
Initial Training	98.42	93.76
Initial Adaptation	98.21	95.48
Continual Adaptation	98.08	96.73
Final Adapted Model	**97.96**	**97.21**

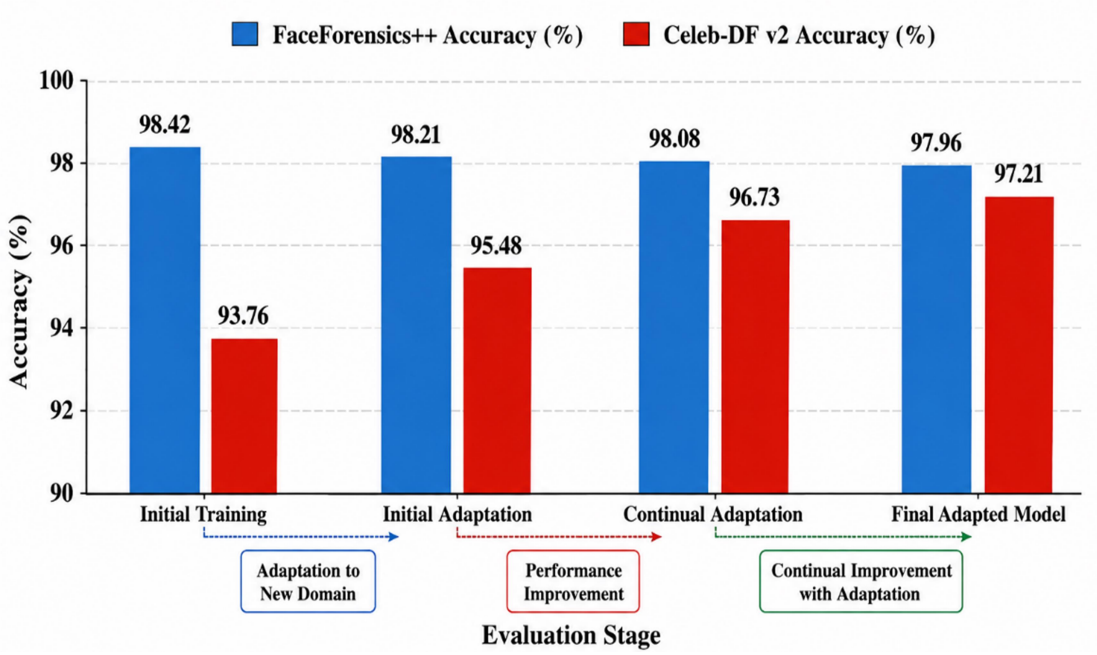


Figure 5: Continual Adaptation Performance of ECDF-Net

Figure 5 illustrates the performance of ECDF-Net across successive adaptation stages. While the accuracy on FaceForensics++ decreases only slightly from 98.42% to 97.96%, the performance on Celeb-DF v2 progressively improves from 93.76% to 97.21%. These results demonstrate that the continual adaptation mechanism effectively learns new DeepFake characteristics while preserving previously acquired forensic knowledge and reducing catastrophic forgetting.

4.5 Forgery Localization Performance

To evaluate the effectiveness of the proposed ECDF-Net in identifying manipulated facial regions, its forgery localization performance was compared with representative CNN-based, attention-based, and spatial-frequency localization methods.

Table 5 presents the comparative results using Intersection over Union (IoU) and Localization Accuracy, providing an assessment of the model's ability to accurately identify and localize DeepFake manipulation regions.

Table 5: Forgery Localization Performance

Method	IoU (%)	Localization Accuracy (%)
CNN-based Localization	78.46	82.15
Attention-based Localization	82.37	85.91
Spatial-Frequency Localization	86.18	89.42
ECDF-Net	90.74	93.28

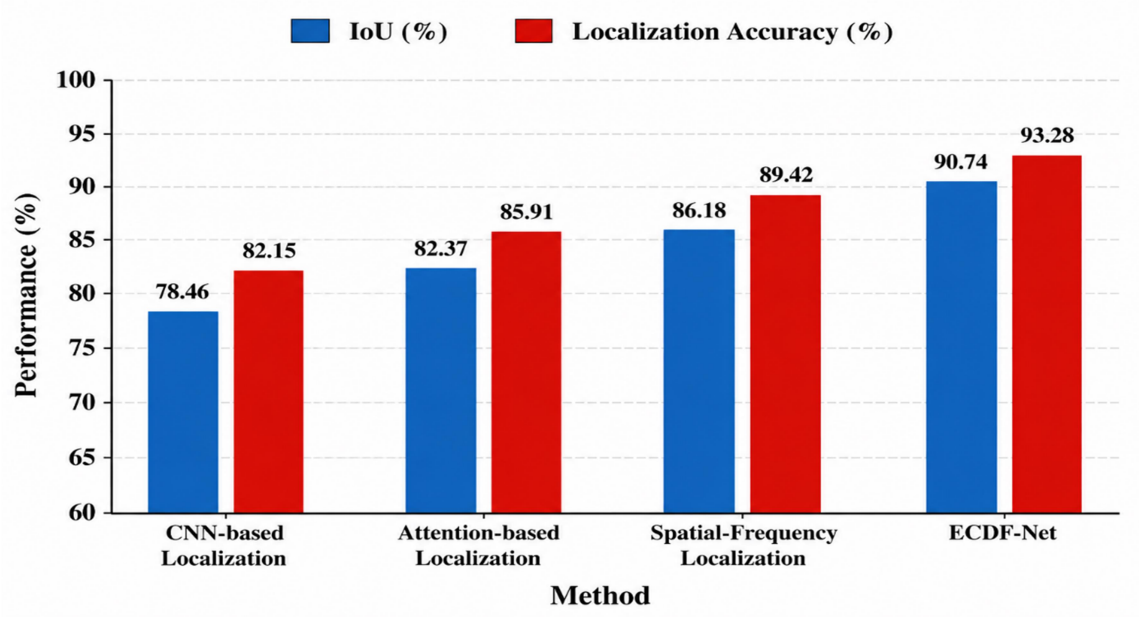


Figure 6: Forgery Localization Performance Comparison

Figure 6 compares the forgery localization performance of CNN-based, attention-based, spatial-frequency, and ECDF-Net methods using IoU and localization accuracy. Performance improves progressively across the methods, with ECDF-Net achieving the best results of 90.74% IoU and 93.28% localization accuracy**.** These results demonstrate the effectiveness of the proposed unified spatial-frequency forensic representation for accurately identifying manipulated facial regions.

4.6 Robustness Under Image Transformations

To evaluate the robustness of the proposed ECDF-Net under common image degradations, experiments were conducted using JPEG compression, Gaussian noise, Gaussian blur, and image resizing. The performance of ECDF-Net was compared with the baseline model using classification accuracy.

Table 6 presents the results, demonstrating the ability of the proposed framework to maintain reliable DeepFake detection performance under different image transformation conditions.

Table 6: Robustness Evaluation Under Different Transformations

Transformation	Baseline Accuracy (%)	ECDF-Net Accuracy (%)
Original Images	96.18	**98.42**
JPEG Compression	89.74	**95.86**
Gaussian Noise	88.92	**95.14**
Gaussian Blur	87.65	**94.62**
Image Resizing	90.38	**96.07**

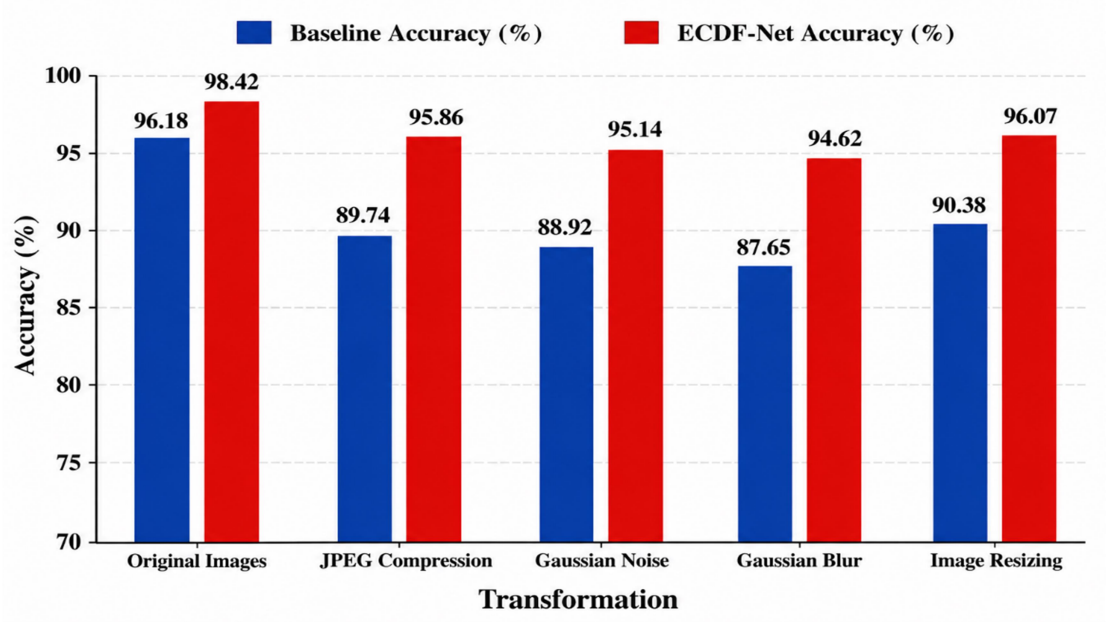


Figure 7: Robustness Evaluation of ECDF-Net under Common Image Transformations

Figure 7 compares the robustness of ECDF-Net with the baseline model under original and transformed images, including JPEG compression, Gaussian noise, Gaussian blur, and image resizing. ECDF-Net consistently achieves higher accuracy across all conditions, maintaining 95.86%, 95.14%, 94.62%, and 96.07% under the respective transformations. These results demonstrate the improved robustness of ECDF-Net against common image degradations and variations encountered in practical DeepFake detection scenarios.

4.7 Explainability and Consistency Analysis

To evaluate the reliability of the proposed ECDF-Net explanations, its performance was compared with Standard Attention, Grad-CAM-based Analysis, and Spatial-Frequency Attention. Table 7 presents the results in terms of Explanation Stability and Consistency Score, assessing the ability of each method to generate stable and reliable forensic evidence for DeepFake detection decisions.

Table 7: Explanation Consistency Performance

Method	Explanation Stability (%)	Consistency Score
Standard Attention	82.16	0.821
Grad-CAM-based Analysis	86.74	0.867
Spatial-Frequency Attention	90.38	0.904
ECDF-Net	94.26	0.943

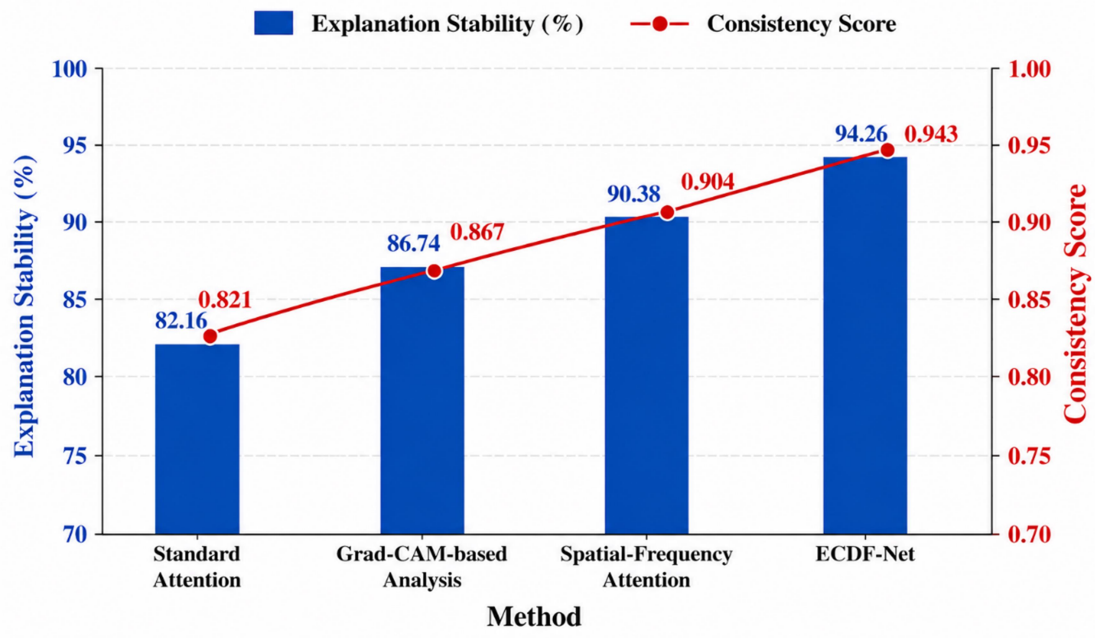


Figure 8: Explanation Consistency Performance

Figure 8 compares the explanation stability and consistency of ECDF-Net with Standard Attention, Grad-CAM-based Analysis, and Spatial-Frequency Attention. Both metrics improve progressively across the methods, with ECDF-Net achieving the highest explanation stability of 94.26% and consistency score of 0.943. These results demonstrate that the proposed framework produces more stable and reliable forensic explanations, supporting consistent interpretation of DeepFake detection decisions.

4.8 Computational Efficiency

To evaluate the computational efficiency of the proposed ECDF-Net, its performance was compared with representative deep learning models, including ResNet50, EfficientNetB0, Vision Transformer, and the Spatial-Frequency Model. Table 8 presents the comparative results in terms of model parameters, inference time, and detection accuracy, highlighting the trade-off between computational complexity, inference efficiency, and DeepFake detection performance.

Table 8: Computational Performance Comparison

Method	Parameters (M)	Inference Time (ms)	Accuracy (%)
ResNet50	25.6	18.7	94.85
EfficientNetB0	5.3	14.2	95.36
Vision Transformer	86.6	28.5	96.18
Spatial-Frequency Model	32.4	24.1	97.06
ECDF-Net	38.7	26.8	98.42

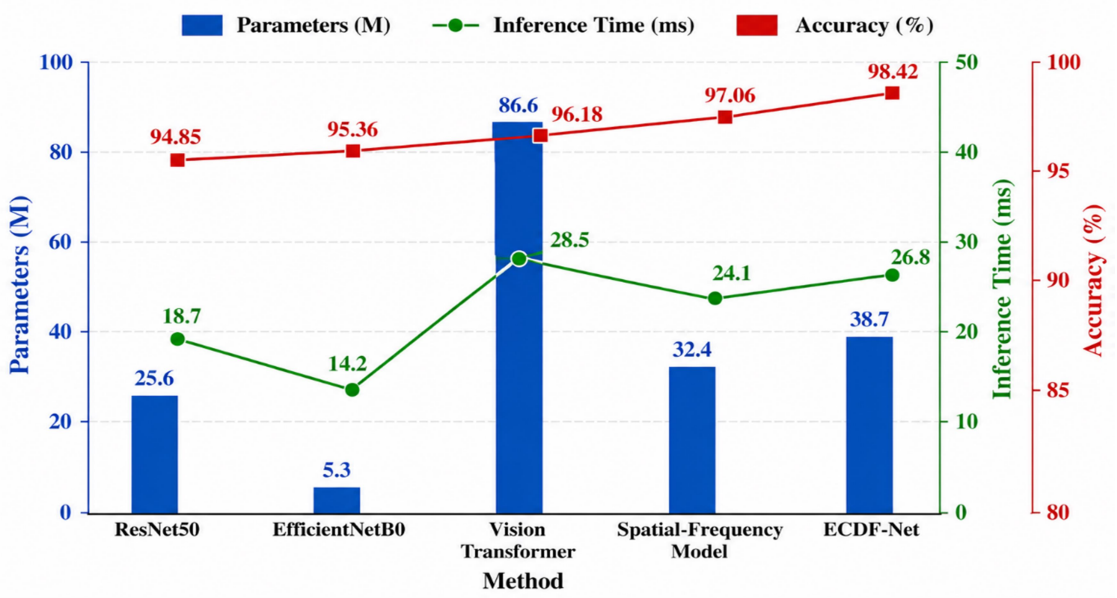


Figure 9: Computational Efficiency and Performance Comparison of ECDF-Net

Figure 9 compares the models in terms of model parameters, inference time, and detection accuracy. ECDF-Net achieves the highest accuracy of 98.42**%** with 38.7 million parameters and an inference time of 26.8 ms**.** Although its computational cost is higher than lightweight models such as EfficientNetB0, it provides a favorable balance between computational efficiency and detection performance, outperforming all compared methods in accuracy.

4.9 Discussion

The results demonstrate that ECDF-Net provides effective DeepFake detection, cross-dataset generalization, continual adaptation, forgery localization, robustness, and explanation consistency. Its strong performance is mainly attributed to the integration of complementary spatial and frequency-domain features through adaptive fusion.

ECDF-Net achieved 98.42% accuracy, 98.39% F1-score, and an AUC of 0.993 on FaceForensics++. The ablation results further confirm that adaptive feature fusion, prototype memory, robustness consistency learning, and explanation consistency learning progressively improve the overall performance.

The cross-dataset results highlight the importance of continual adaptation. Performance on Celeb-DF v2 improved from 93.76% during initial evaluation to 97.21% after adaptation, while maintaining strong performance on FaceForensics++. This demonstrates the effectiveness of prototype-based knowledge retention and reliability-aware updates in adapting to new forgery patterns.

The robustness experiments show that ECDF-Net maintains reliable performance under JPEG compression, noise, blur, and resizing. In addition, the localization and explanation results demonstrate its ability to provide meaningful forensic evidence, achieving 94.26% explanation stability and a 0.943 consistency score.

5. DIFFERENCE FROM PRIOR WORK

This section explains how ECDF-Net differs from the studies reviewed in Section 2, compares its results with those of related methods, analyzes its advantages and disadvantages in light of the literature, and assesses whether the research objectives stated in Section 2.3 were achieved.

5.1. Positioning Against Prior Studies

The related studies can be grouped into four streams. Spatial-frequency detectors [18]-[20], [27]-[29] improve discrimination by combining visual and spectral cues but are trained once and remain static. Continual detectors [11], [17], [30] adapt to new forgeries but produce only a class label. Detection-localization models [21], [26] identify manipulated regions but are not designed to adapt to new domains, and explainable detectors [19], [24] provide visual evidence without enforcing its stability.

ECDF-Net differs from all four streams in that it treats these concerns as one problem and serves them from a single fused representation.

Table 9 summarizes this difference with respect to the most closely related studies.

Table 9: Comparison of the Scope of ECDF-Net with Closely Related Studies

Study	Spatial-frequency fusion	Continual adaptation	Forgery localization	Explainable evidence	Robustness/ explanation consistency
Xu *et al.* [17]	Yes	Yes	No	No	No
Yang *et al.* (SFD-ViT) [18]	Yes	No	No	No	No
Kumar *et al.* [19]	Yes	No	No	Yes	Partial
Zheng *et al.* [21]	Yes	No	Yes	No	No
Al-Imran *et al.* (DeFaX) [24]	Partial	No	No	Yes	Partial
Han *et al.* (HSFF-Net) [26]	Yes	No	Yes	No	No
Wang *et al.* [27]	Yes	No	No	No	No
Tassone *et al.* [30]	No	Yes	No	No	No
ECDF-Net (Proposed)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Note: Yes: explicitly addressed; Partial: addressed in part (e.g., robustness without explanation consistency, or feature fusion that is not specifically spatial-frequency); No: not explicitly addressed, based on the scope reported in each study (Section 2).

Three differences are particularly important. First, Wang *et al.* [27] also fuse spatial and frequency features adaptively, but the fused representation is used only for classification; in ECDF-Net, the same representation additionally forms the prototype memory, the localization map, and the explanation map. Second, compared with the continual detectors of Xu *et al.* [17] and Tassone *et al.* [30], which concentrate on adaptation and knowledge retention for classification, ECDF-Net couples adaptation with a reliability gate that combines prediction confidence, prototype similarity, and uncertainty (Eq. 20), so that only reliable samples update the memory, and it also outputs localization and explanation evidence. Third, whereas Kumar *et al.* [19] and Al-Imran *et al.* [24] generate explanations and SNIPPET [32] evaluates explanations after training, ECDF-Net optimizes explanation stability directly during training through the explanation consistency loss (Eq. 37).

5.2. Comparison of Results with Related Studies

Under the experimental protocol of this study, ECDF-Net achieved 98.42% accuracy and an AUC of 0.993 on FaceForensics++, compared with 97.06% and 0.985 for the adaptive spatial-frequency model [27] and 97.65% and 0.977 for EfficientNet-B4 [14] (Table 1). The main gain, however, lies in the behavior of the model under domain change rather than in in-domain accuracy. The model trained only on FaceForensics++ lost 4.66 percentage points when tested directly on Celeb-DF v2, whereas continual adaptation raised Celeb-DF v2 accuracy to 97.21%, matching the 97.18% obtained by a model trained directly on Celeb-DF v2, with a decrease of only 0.46 points on FaceForensics++ (Tables 2 and 4). For localization, ECDF-Net exceeded spatial-frequency localization by 4.56 IoU points (Table 5), and for explanations it exceeded spatial-frequency attention by 3.88 points in explanation stability (Table 7). These comparisons were obtained under a common protocol; because published studies differ in frame sampling, compression level, and data partitioning, absolute values reported in other papers should be compared with caution.

5.3. Advantages of ECDF-Net

In light of the literature, ECDF-Net offers four main advantages. (i) It provides forensic evidence beyond a binary label: a single forward pass yields a decision, a localization map, prototype similarity, and an explanation, whereas classification-oriented methods such as [18], [20], and [27] yield only a label. (ii) It adapts to a new domain without retraining from scratch, because earlier knowledge is summarized as a compact set of prototypes whose matching cost is O(KD) (Eq. 45) rather than by retraining on the complete source dataset. (iii) Its reliability gate reduces the risk that uncertain samples from the new domain corrupt the learned memory. (iv) Its stability is measured rather than assumed: the average accuracy decrease under four common transformations was 3.00 points, compared with 7.01 points for the baseline (Table 6), and each component made a measurable contribution in the ablation study (Table 3).

5.4. Disadvantages of ECDF-Net

The approach also has disadvantages relative to prior work. (i) It is computationally heavier than lightweight detectors: it requires 38.7 million parameters and 26.8 ms per inference, compared with 5.3 million parameters and 14.2 ms for EfficientNetB0 (Table 8), and it has not been optimized for the edge and UAV settings targeted by [14] and [15]. (ii) Its multi-task objective introduces more hyperparameters than single-objective detectors, namely the loss weights in Eq. (40), the reliability weights in Eq. (20), the threshold δ, and the update coefficient η, which increases tuning effort. (iii) Continual adaptation is not entirely free of forgetting, as shown by the 0.46-point decrease on FaceForensics++. (iv) Because it is frame-based, it cannot exploit the temporal inconsistencies used by spatiotemporal detectors [4], [6]. (v) Pixel-level localization supervision depends on the availability of manipulation masks, which many datasets do not provide.

5.5. Achievement of Research Objectives

Table 10 compares each research objective with the evidence obtained in this study and with the corresponding baseline or related results. All four objectives were achieved within the evaluated setting, and the associated hypotheses are revisited in Section 7.

Table 10: Achievement of Research Objectives Compared with Baseline and Related Results

Objective (Hypothesis)	Evidence from this study	Comparison with baseline / related results	Outcome
O1 (H1): Adaptive spatial-frequency fusion for detection	Spatial only 95.36%, frequency only 94.82%, spatial + frequency 96.74%, with adaptive fusion 97.31%; full model 98.42% accuracy and 0.993 AUC (Tables 1, 3)	+1.36 points accuracy and +0.008 AUC over the spatial-frequency model [27]; +0.77 points over EfficientNet-B4 [14]	Achieved; H1 supported
O2 (H2): Continual adaptation with knowledge retention	Celeb-DF v2 accuracy increased from 93.76% to 97.21%; FaceForensics++ accuracy changed from 98.42% to 97.96% (Table 4)	Cross-dataset gap reduced from 4.66 to 0.75 points; final Celeb-DF v2 accuracy matches a model trained directly on Celeb-DF v2 (97.18%) without retraining from scratch	Achieved; H2 supported
O3 (H3): Joint forgery localization	IoU 90.74% and localization accuracy 93.28% on FaceForensics++ (Table 5)	+4.56 IoU points and +3.86 localization accuracy points over spatial-frequency localization	Achieved; H3 supported
O4 (H4): Robust predictions and stable explanations	Average accuracy decreases of 3.00 points under four transformations; explanation stability 94.26% and consistency 0.943; ablation gains of +0.39 and +0.35 points (Tables 3, 6, 7)	Baseline average decrease of 7.01 points; +3.88 stability points and +0.039 consistency over spatial-frequency attention	Achieved; H4 supported

6. LIMITATIONS OF THE STUDY

Although the results support the proposed framework, several limitations should be considered when interpreting them.

Dataset coverage. The evaluation is limited to two benchmark datasets, FaceForensics++ and Celeb-DF v2, both of which consist mainly of face-swap and facial reenactment videos. ECDF-Net has not yet been evaluated on larger in-the-wild collections such as DFDC, DeeperForensics-1.0, or WildDeepfake, or on fully synthetic and diffusion-generated faces, so its performance on these types of content remains to be established.

Scope of continual adaptation. Continual adaptation was examined for a single new domain introduced in sequential stages. Real deployments may face a long sequence of new generators, in which small amounts of forgetting, such as the 0.46-point decrease observed on FaceForensics++, could accumulate. In addition, the reliability threshold δ, the update coefficient η, and the loss weights were set empirically, and samples that are confidently misclassified can still pass the reliability check and introduce errors into the prototype memory.

Modality and localization. ECDF-Net analyzes individual frames and does not model temporal inconsistencies across frames or audio-visual mismatches, which spatiotemporal methods [4], [6] exploit. Localization was evaluated only on

FaceForensics++, where ground-truth manipulation masks are available; because Celeb-DF v2 does not provide pixel-level masks, localization performance under domain shift was not quantified.

Explanation and robustness assessment. The explanation consistency metrics measure whether explanations remain stable under transformations, not whether they are faithful to the true manipulation or useful to human examiners, and no expert or user study of the kind proposed in SNIPPET [32] was conducted. Robustness was likewise assessed only against benign degradations; adversarial perturbations and adaptive attacks designed to evade the detector were not considered.

Computational and statistical aspects. With 38.7 million parameters and an inference time of 26.8 ms, ECDF-Net is heavier than lightweight detectors intended for edge and UAV platforms [14], [15], and real-time deployment on such devices was not demonstrated. Finally, the results are reported as point estimates; variability across repeated runs and statistical significance tests were not included and will be addressed in future evaluations.

7. CONCLUSION & FUTURE WORK

This study set out to answer the problem stated in Section 2.2: whether a single DeepFake forensic model can remain accurate when a previously unseen forgery domain appears, without retraining from scratch and without forgetting what it has already learned, while also showing where a manipulation lies and keeping that evidence stable when the media are degraded. Four hypotheses were formulated to test this proposition, and the experimental evidence obtained with ECDF-Net supports each of them.

The first premise was that neither spatial nor frequency cues are sufficient on their own (H1). The data confirm this: the spatial and frequency branches alone reached 95.36% and 94.82% accuracy, their combination reached 96.74%, and adaptive fusion raised this to 97.31%. With all components, ECDF-Net reached 98.42% accuracy and an AUC of 0.993 on FaceForensics++, exceeding the adaptive spatial-frequency model [27] (97.06% and 0.985) evaluated under the same protocol.

The central premise of the problem was that static detectors fail when the domain changes, and the results make this failure visible: the model trained only on FaceForensics++ lost 4.66 percentage points when tested directly on Celeb-DF v2. Reliability-aware continual adaptation recovered almost all of this loss, raising Celeb-DF v2 accuracy to 97.21%, which matches the 97.18% obtained by a model trained directly on Celeb-DF v2, while FaceForensics++ accuracy decreased by only 0.46 points. The generalization gap therefore narrowed from 4.66 to 0.75 points, which supports H2 and shows that adaptation and knowledge retention need not be traded against each other when memory updates are gated by reliability.

The remaining premises concerned forensic evidence. ECDF-Net localized manipulated regions with 90.74% IoU, 4.56 points above spatial-frequency localization, supporting H3. Under JPEG compression, Gaussian noise, Gaussian blur, and resizing, its accuracy decreased by an average of 3.00 points, compared with 7.01 points for the baseline, and its explanation stability reached 94.26% with a consistency score of 0.943, compared with 90.38% and 0.904 for spatial-frequency attention, supporting H4. The ablation study further shows that the two consistency objectives increased accuracy by 0.39 and 0.35 points, so stability was obtained without sacrificing detection performance.

Taken together, these findings support the argument that accuracy, adaptability, localization, and explanation stability do not have to be pursued with separate models; a single adaptively fused spatial-frequency representation can serve all of them. This benefit has a measurable cost of 6.3 million additional parameters and 2.7 ms of additional inference time relative to the closest spatial-frequency model, and the conclusion holds within the scope tested, namely two benchmark datasets and a single new forgery domain.

Future work will address the limitations identified in Section 6. The framework will be evaluated on additional benchmarks such as DFDC, DeeperForensics-1.0, and WildDeepfake, on diffusion-generated faces, and over longer continual sequences involving several new generators. Temporal and audio-visual cues will be incorporated, the reliability threshold will be made adaptive rather than fixed, a lightweight variant will be developed through pruning or knowledge distillation for edge deployment, and the explanations will be assessed for faithfulness through expert-centered evaluation and under adversarial perturbations.

REFERENCES:

[1] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5052–5060, doi: 10.1609/aaai.v38i5.28310

[2] B. Yan, C.-T. Li, and X. Lu, "JRC: Deepfake detection via joint reconstruction and classification," *Neurocomputing*, vol. 598, Art. no. 127862, 2024, doi: 10.1016/ j. neucom.2024.127862

[3] Z. Cheng, Y. Wang, Y. Wan, and C. Jiang, "DeepFake detection method based on multi-scale interactive dual-stream network," *Journal of Visual Communication and Image Representation,* vol. 104, Art. no. 104263, 2024, doi: 10.1016/j.jvcir.2024.104263.

[4] B. Liu, M. Ding, and T. Zhu, "MeST-Former: Motion-enhanced Spatiotemporal Transformer for generalizable Deepfake detection," *Neurocomputing,* vol. 610, Art. no. 128588, 2024, doi: 10.1016/j.neucom.2024.128588.

[5] L. Shi, J. Zhang, Z. Ji, J. Bai, and S. Shan, "Real face foundation representation learning for generalized deepfake detection," *Pattern Recognition,* vol. 161, Art. no. 111299, 2025, doi: 10.1016/j.patcog.2024.111299.

[6] R. Zhang, P. He, H. Li, S. Wang, et al., "Temporal Diversified Self-Contrastive Learning for Generalized Face Forgery Detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, doi: 10.1109/TCSVT.2024.3436554.

[7] Q. Gao, B. Zhang, J. Wu, W. Luo, Z. Teng, and J. Fan, "Leveraging facial landmarks improves generalization ability for deepfake detection," *Pattern Recognition*, vol. 164, Art. no. 111528, 2025, doi: 10.1016/j.patcog.2025.111528.

[8] X. Cui, Y. Li, A. Luo, J. Zhou, and J. Dong, "Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19207–19217,2025, doi: 10.1109/CVPR 5273 4.2025.01789.

[9] S. Guo, Q. Li, M. Gao, X. Zhu, and I. Rida, "Generalizable deepfake detection via Spatial Kernel Selection and Halo Attention Network," *Image and Vision Computing*, vol. 160, Art. no. 105582, 2025, doi: 10.1016/j.imavis.2025. 105582.

[10] H. Chen, Y. Jia, Z. Shi, and D. Zhang, "ADNet: Delving into generalizable deepfake detection via adaptive expert selection and discrepancy learning," *Pattern Recognition*, vol. 173, Art. no. 112755, 2026, doi: 10.1016/j.patcog.2025. 112755.

[11] X. Zhang, P. Zhu, C. Zhang, Z. Yan, J. Cheng, M. Lao, S. Cai, and Y. Guo, "Generalization-Preserved Learning: Closing the Backdoor to Catastrophic Forgetting in Continual Deepfake Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 3798–3808,2025, doi: 10.1109/ICCV5170 1.2025.00362.

[12] S. Guo, M. Gao, G. Zhang, C. Liu, and Q. Li, "Deepfake Detection via Spatial-Frequency Attention Network," *IEEE Transactions on Consumer Electronics,* vol. 71, no. 4, 2025, pp. 9832–9841, doi: 10.1109/TCE.2025.3614720.

[13] A. Sardhara, V. Vekariya, A. R. Pathak, and S. Dash, "DeepForgeryNet: A hybrid CNN–LSTM and transfer learning framework for robust image forgery and deepfake detection," *Frontiers in Artificial Intelligence*, vol. 9, Art. no. 1810912, 2026, doi: 10.3389/frai.2026. 1810912.

[14] Aljabri, Jawhara, "Integrating enhanced feature fusion with adaptive hidden semi-Markov model and deep neural network for robust deepfake detection in Internet of Things-based edge computing environments," *Journal of King Saud University: Computer and Information Sciences,* vol. 38, Art. no. 235, 2026, doi: 10.1007/s44443-026-00654-1

[15] N. H. Alshammari and S. A. Alshammari, "DeepLayer-ID: A lightweight multi-domain forensic framework for real-time deepfake detection in resource-constrained UAV sensor platforms," *Sensors*, vol. 26, no. 9, Art. no. 2705, 2026, doi: 10.3390/s26092705

[16] S. Potluri, S. P. Kandagatla, S. N. Mohanty, K. C. Rout, M. Israr, and V. M. Gupta, "Hybrid deep feature integration model for robust deepfake detection using transfer-learned neural networks," *Frontiers in Artificial Intelligence,* vol. 9, Art. no. 1737761, 2026, doi: 10.3389/frai.2026.1737761.

[17] Y. Xu, J. Ren, A. Fang, M. Irfan, and J. Zheng, "Continual face forgery detection based on relation-aware spatial-frequency interaction aggregation and contrastive learning," *Pattern Recognition,* vol. 172, Art. no. 112456, 2026, doi: 10.1016/j.patcog.2025.112456.

[18] Y. Yang, Y. Deng, Q. Yi, and J. Yi, "SFD-ViT: Spatial-Frequency Dual-Stream Transformer with Patch Embedding for Face Forgery Detection," *Knowledge-Based Systems*, vol. 332, Art. no. 114838, 2026, doi: 10.1016/ j.knosys.2025.114838

[19] M. Kumar, A. Kumar, V. Yadav, A. Shankar, and P. Chauhan, "A Hybrid Spatial-Frequency Attention-Based Algorithm Using EfficientNet for Robust and Interpretable Deepfake Detection," *Scientific Reports*, vol. 16, Art. no. 16612, 2026, doi: 10.1038/s41598-026-46086-9.

[20] Y. Kou, P. Li, H. Ma, J. Zhou, Z. Huang, *et al.*, "SFIAD: Deepfake Detection through Spatial-Frequency Feature Integration and Dynamic Margin Optimization," *Artificial Intelligence Review*, vol. 58, Art. no. 217, 2025, doi: 10.1007/s10462-025-11225-7.

[21] J. Zheng, Y. Zhou, N. Zhang, X. Hu, K. Xu, D. Gao, and Z. Tang, "A spatio-frequency cross fusion model for deepfake detection and segmentation," *Neurocomputing*, vol. 628, Art. no. 129683, 2025, doi: 10.1016/j.neucom. 2025.129683

[22] J. Li, Y. Hu, B. Liu, H. She, and C.-T. Li, "Deepfake Detection with Domain Generalization and Mask-Guided Supervision," *Pattern Recognition*, vol. 165, Art. no. 111622, 2025, doi: 10.1016/j.patcog. 2025.111622.

[23] Z. Lai, Y. Zhang, D. Li, and J. Ni, "Leveraging High-Frequency Diversified Augmentation for General Deepfake Detection," *Journal of Information Security and Applications*, vol. 89, Art. no. 103994, 2025, doi: 10.1016/j.jisa. 2025.103994.

[24] M. Al-Imran, M. S. Sheikh, U. Kirtonia, N. T. Arthi, and S. Ripon, "DeFaX: A cross-attention fusion framework for robust and explainable deepfake detection," *IEEE Access*, vol. 13, 2025, pp. 213962–213979, doi: 10.1109/ ACCESS.2025.3645769.

[25] C. Zhao, C. Wang, Z. Song, G. Hu, L. Wang, and D. Miao, "Multi-Definition Deepfake Detection via Semantics Reduction and Cross-Domain Training," *Pattern Recognition*, vol. 163, Art. no. 111469, 2025, doi: 10.1016/ j.patcog.2025.111469.

[26] R. Han, X. Wang, N. Bai, J. Hou, W. Zhang, and J. Li, "HSFF-Net: Hierarchical Spectral-Feature Fusion Network for Deepfake Detection and Localization," *Neural Networks*, vol. 192, Art. no. 107967, 2025, doi: 10.1016/ j.neunet.2025.107967.

[27] F. Wang, Q. Chen, B. Jing, Y. Tang, Z. Song, and B. Wang, "Deepfake detection based on the adaptive fusion of spatial-frequency features," *International Journal of Intelligent Systems*, vol. 2024, no. 1, Art. no. 7578036, 2024, pp. 1–12, doi: 10.1155/2024/7578036.

[28] B. Wang, X. Wu, F. Wang, Y. Zhang, F. Wei, and Z. Song, "Spatial-Frequency Feature Fusion Based Deepfake Detection Through Knowledge Distillation," *Engineering Applications of Artificial Intelligence*, vol. 133, Art. no. 108341, 2024, doi: 10.1016/ j.engappai.2024.108341.

[29] T. Zhai, K. Lu, J. Li, Y. Wang, W. Zhang, P. Yu, and Z. Xia, "Learning spatial-frequency interaction for generalizable deepfake detection," *IET Image Processing*, vol. 18, no. 14, 2024, pp. 4666–4679, doi: 10.1049/ ipr2.13276.

[30] F. Tassone, L. Maiano, and I. Amerini, "Continuous Fake Media Detection: Adapting Deepfake Detectors to New Generative Techniques," *Computer Vision and Image Understanding,* vol. 249, Art. no. 104143, 2024, doi: 10.1016/j.cviu.2024.104143

[31] M. A. Amin, Y. Hu, C.-T. Li, and B. Liu, "Deepfake detection based on cross-domain local characteristic analysis with multi-domain transformer," Alexandria Engineering Journal, vol. 91, 2024, pp. 592–609, doi: 10.1016/ j.aej.2024.02.035.

[32] Y. Yang, B. Joukovsky, J. Oramas Mogrovejo, T. Tuytelaars, and N. Deligiannis, "SNIPPET: A Framework for Subjective Evaluation of Visual Explanations Applied to DeepFake Detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 8, Art. no. 253, 2024, pp. 1-29, doi: 10.1145/3665248.

[33] https://github.com/ondyari/faceforensics

[34] https://github.com/yuezunli/celeb-deepfakeforensics