A LITERATURE REVIEW ON ROBUST DEEPFAKE DETECTION UNDER REAL-WORLD NETWORK CONDITIONS

By

Shreya Bhushappagala

A capstone project submitted for Graduation with University Honors

December 01, 2025

University Honors

University of California, Riverside

Dr. Srikanth V. Krishnamurthy

Department of Computer Science and Engineering

Dr. Amit K. Roy-Chowdhury

Department of Electrical and Computer Engineering

Dr. Begoña Echeverria, Howard H Hays, Jr. Chair

University Honors

Abstract

Deepfakes have become increasingly realistic with the rise of diffusion models, vision transformers, and multimodal generators, posing major risks to security, trust, and authenticity in digital media. While existing detectors achieve strong benchmark accuracy, their performance deteriorates under real-world distortions such as compression, transmission loss, or bitrate fluctuations. This literature review traces the evolution of deepfake detection—from early spatial and temporal CNNs to frequency-, latent-, and transformer-based architectures—and examines robustness-oriented frameworks including ADD, QAD, BZNet, and DPL.

To address the gap between laboratory conditions and deployment environments, this study introduces two complementary research directions. First, an adaptive detection framework, the **MDN-Guided Noise-Aware Modular Ensemble**, models test-time uncertainty using a Mixture Density Network to dynamically select noise-specific fine-tuning modules. Second, a new dataset, **DeFND (Deepfake Forensics under Network Degradation)**, captures authentic Wi-Fi and cellular transmission artifacts by streaming deepfake videos through real WebRTC pipelines.

Together, these contributions provide a foundation for evaluating and improving deepfake detectors under realistic network and corruption conditions, advancing the pursuit of reliable, uncertainty-aware detection systems for practical deployment.

Acknowledgements

This work was supported in part by the National Center for Communications and Computation for Communities (NC4) at the University of California, Riverside, and by the Vision and Learning Laboratory in the Department of Electrical and Computer Engineering. I am deeply grateful for the guidance, discussions, and technical support provided throughout the development of this project. I extend my sincere appreciation to Shashwat Jha, Sayak Nag, Zhutian Liu, Fahim Faisal Niloy, Saketh Bachu, Rohit Kundu, Zhaowei Tan, and Zhiyun Qian for their collaborative efforts and contributions that were instrumental in building and refining the models presented in this work. I am especially thankful to Dr. Srikanth V. Krishnamurthy and Dr. Amit Roy-Chowdhury for their continuous support, valuable insights, and mentorship, which greatly shaped the direction and quality of this research.

Contents

1 Introduction

Recent advances in diffusion-based generators, transformer architectures, and multi-modal foundation models have significantly increased the realism and accessibility of synthetic media [1, 2, 3]. Combined with the global reach of social media platforms, these technologies have enabled the rapid creation and dissemination of highly convincing manipulated content [4]. Such fabricated media—commonly known as *deepfakes*—have been increasingly used for political propaganda, financial deception, and non-consensual explicit material [5, 6, 7]. The growing accessibility of open-source generation tools has blurred the line between authentic and manipulated information, creating profound social, ethical, and security challenges [8, 9]. As a result, the development of robust and generalizable deepfake detection methods has emerged as a critical area of research within computer vision and digital forensics.

Early detection research primarily focused on improving discriminative accuracy under controlled conditions. Convolutional and recurrent neural network models such as XceptionNet, MesoNet, and CNN-LSTM architectures learned spatial inconsistencies, blending boundaries, and short-term temporal irregularities, achieving impressive results on benchmark datasets like FaceForensics++ and Celeb-DF [10, 11, 7, 12, 13, 14]. Although these methods achieved strong performance on curated datasets, their effectiveness decreased sharply when tested on unseen or degraded inputs [15, 16]. Studies show that compression artifacts and real-world degradations—including JPEG compression, motion blur, transmission errors, and sensor noise—can reduce deepfake detection accuracy by 20–60 percentage points, as detectors trained on clean datasets often fail to generalize to compressed or noisy conditions. This highlights that most existing methods implicitly assume clean and stable input distributions [17, 18, 19], revealing a major gap between laboratory accuracy and real-world performance.

To overcome these generalization failures, recent work has shifted toward robustness-oriented frameworks designed to maintain performance under quality degradation and domain shift. Frequency-based models such as ADD [20] employ spectral attention and knowledge distillation to preserve discriminative cues under heavy compression. Quality-Agnostic Detection (QAD) [21]

introduces collaborative learning to align feature representations across multiple quality levels, while Dual Progressive Learning (DPL) [22] progressively bridges high- and low-quality feature spaces. Transformer-based detectors, including UiA-ViT [23, 24], leverage self-attention to capture long-range spatial dependencies and have shown improved cross-dataset generalization. Other recent methods—such as BZNet [25], CrossDF [26], LNCLIP-DF [27], and diffusion-augmented frameworks [28]—extend this direction by improving robustness to compression artifacts and improving cross-domain generalization. Despite these advances, most approaches still rely heavily on supervised data augmentation and struggle to adapt dynamically to unseen corruptions or mixed noise conditions.

This capstone focuses on robustness under real network conditions, which remains one of the main challenges in deepfake forensics. It reviews existing methods that improve the reliability of deepfake detectors against noise, compression, and other real-world distortions, and introduces a probabilistic framework for adaptive detection. The proposed Mixture Density Network (MDN)–guided, noise-aware modular ensemble models test-time corruptions as mixtures of Gaussians, enabling the detector to adjust to new or changing types of degradation. The system is evaluated against Dual Progressive Learning (DPL) under COTS, 5G, and campus network environments. Experimental results show that modeling uncertainty enhances both stability and generalization when videos are degraded by compression, packet loss, or fluctuations in network quality.

Figure 1: Example frames from the *Celeb-DF v2* dataset [29]. The first column shows authentic video frames, while the remaining columns display manipulated (deepfake) versions.

2 Evolution of Deepfake Detection

2.1 Spatial-Based Detectors

Early deepfake detection efforts primarily relied on spatial-domain analysis, focusing on identifying visual inconsistencies in manipulated facial regions using convolutional neural networks (CNNs) [10, 11, 30, 31, 32]. These methods operate frame-by-frame and aim to capture low-level artifacts introduced during synthesis, such as blending boundaries, color mismatches, and texture irregularities.

One of the earliest and most influential models was XceptionNet, introduced by Rössler et al. [10] as part of the FaceForensics++ benchmark. It leveraged depthwise separable convolutions to efficiently extract pixel-level features [7, 33], achieving high detection accuracy on large-scale datasets including DeepFake, Face2Face, FaceSwap, and NeuralTextures. Similarly, MesoNet [11] adopted a lightweight architecture composed of shallow convolutional layers designed to focus on mesoscopic features—intermediate scale manipulation traces that exist between pixel-level distortions and semantic inconsistencies [34]. This balance enabled fast training and good performance on limited data but restricted its ability to generalize beyond known manipulation types. [10, 35]

Bayar and Stamm [30] introduced a constrained convolutional layer that suppresses semantic image content while enhancing forensic noise patterns. This innovation allowed the model to learn subtle pixel-level correlations left behind by manipulation pipelines, independent of scene context [36]. Later studies adopted stronger backbones to improve feature extraction and generalization. The FaceForensics++ benchmark [10] showed that early CNNs like XceptionNet achieved high accuracy on clean data but dropped sharply under compression, revealing limits in robustness. To address this, models began using HRNet and EfficientNet—HRNet preserves high-resolution features critical for fine manipulation detection [37], while EfficientNet [38] balances depth, width, and resolution for efficient scaling. These architectures, used in Face X-Ray and Lips Don't Lie [39, 40], improved cross-dataset generalization on FaceForensics++, DFDC, and Celeb-DF.

Despite these advances, spatial CNN-based detectors often overfit to dataset-specific artifacts and fail to generalize across different generation methods or compression levels. Their single-frame

focus makes them prone to noise and compression, which blur the subtle cues essential for detection. These limitations motivated a shift toward temporal and semantic modeling, where motion dynamics and behavioral consistency offer more robust detection cues. [10, 41, 32] Overall, spatial CNNs excelled at capturing pixel-level artifacts but remained inherently static, motivating researchers to explore temporal coherence and behavioral cues as richer indicators of forgery. [10, 41, 32, 42, 43]

2.2 Temporal and Behavioral Detectors

To capture information beyond static artifacts, later research integrated temporal coherence and behavioral cues into deepfake detection frameworks. Guera and Delp [44] introduced one of the first temporal models for deepfake detection, combining a convolutional feature extractor with a Long Short-Term Memory (LSTM) network to analyze frame sequences. Their approach targeted frame-to-frame inconsistencies—such as flickering and illumination variations—caused by the lack of temporal awareness in deepfake generation, which are often invisible in single-frame analysis. Subsequent works explored explicit behavioral and motion-based cues. Li et al. [45] proposed an eye-blinking–based detector, observing that early generative models failed to reproduce natural blink frequencies due to the lack of temporal modeling in training data.

Sabir et al. [46] proposed a recurrent-convolutional model that combines CNN feature extraction with bidirectional recurrent layers to exploit temporal discrepancies across frames. Their experiments on the FaceForensics++ dataset showed that incorporating temporal information improved detection accuracy over frame-based CNN baselines.

AltFreezing [47] proposed a training strategy that alternately freezes spatial and temporal convolutional weights in a 3D ConvNet, encouraging the model to capture both spatial and temporal artifacts. Combined with video-level data augmentation, this approach significantly improved generalization to unseen manipulations and datasets such as Celeb-DF and DeeperForensics.

More recent multimodal approaches have sought to model fine-grained temporal inconsistencies across audio-visual streams. Astrid et al. [48] proposed an attention-based framework that computes local temporal distances between modalities, enabling the detector to identify subtle desynchronization artifacts and improving cross-dataset robustness on DFDC and FakeAVCeleb.

Although temporal modeling has improved the capture of motion and behavioral cues, these detectors remain sensitive to compression, frame drops, and jitter, which can distort or desynchronize motion patterns. These challenges have motivated the exploration of frequency- and latent-space representations that capture manipulation signals more robustly across visual quality variations. In essence, temporal and behavioral models improved realism-aware detection but introduced new sensitivities to compression and synchronization noise, pushing research toward frequency and representation-based methods that encode manipulation traces more stably. [14, 43]

2.3 Frequency and Latent-Space Detectors

Frequency-domain methods: These approaches transform an image from the spatial domain—defined by pixel intensities—into the spatial frequency domain using the Fourier transform. In this form, an image is represented by its frequency components, each with magnitude and phase [49]. This representation helps reveal spectral irregularities that are often invisible in pixel space, making it useful for detecting synthesis artifacts introduced by generative models.

Le and Woo [50] proposed the Attention-based Deepfake Detection Distiller (ADD), which uses frequency attention and multi-view knowledge distillation to enhance detection on compressed deepfakes. Their model trains a student network to recover high-frequency details lost during compression by learning from a high-quality teacher network. Erukude et al. [51] analyzed GAN-generated faces with a Discrete Fourier Transform and found consistent high-frequency anomalies and periodic patterns in StyleGAN outputs that rarely appear in real images. Tan et al. [52] introduced FreqNet, a frequency-aware framework that jointly models amplitude and phase spectra to improve cross-dataset generalization. Together, these studies show that frequency analysis can expose periodic artifacts overlooked by spatial detectors, although these cues remain fragile—compression or filtering can easily weaken them.

Latent-space methods: Latent-space detectors operate in a model's learned feature space rather than in raw pixels or frequencies, focusing on compact representations that capture semantic structure. As described by Abati et al. [53], this involves modeling the distribution of latent embeddings produced by an encoder or autoencoder, allowing the system to capture meaningful

relationships instead of surface-level details. In deepfake detection, this helps separate real and fake content by comparing their latent representations.

Delmas and Séguier [54] proposed *LatentForensics*, which performs classification directly in the StyleGAN latent space using a small fully connected network. This lightweight design achieves competitive accuracy while requiring far less data and computation. Choi et al. [55] extended this idea with *StyleGRU*, a video-based detector that models how StyleGAN latent features change over time using gated recurrent units and contrastive learning. They observed that fake videos exhibit unusually smooth latent dynamics compared to real ones, reflecting reduced temporal variability in synthetic sequences. By learning these patterns, StyleGRU identifies subtle inconsistencies in facial motion and generalizes well across datasets.

Latent-space methods capture deeper generative characteristics and often generalize better than frequency-based detectors, although their performance can still depend on the datasets and architectures used during training. [53, 52, 49] While frequency and latent-space detectors captured deeper generative fingerprints, their reliance on handcrafted representations limited scalability—prompting a shift toward transformer architectures capable of learning such global dependencies directly from data.

2.4 Transformer-Based Detectors

Transformer-based models have recently gained attention in deepfake detection because of their ability to model long-range dependencies and global relationships within an image [56, 57, 23]. Convolutional neural networks are built on local receptive fields [57], while transformers apply self-attention across image patches to model long-range dependencies [57]. In the context of face forgery detection, this self-attention enables UIA-ViT to capture inconsistency information across patches [23]. This global attention makes them particularly effective for identifying spatially distributed artifacts that CNNs may overlook [58, 23].

One of the first successful transformer-based approaches was UIA-ViT (Zhuang et al., ECCV 2022) [23]. It introduced an unsupervised, inconsistency-aware vision transformer for face forgery detection. The model learns to identify patch-level inconsistencies without relying

on pixel-level annotations through two key modules—Unsupervised Patch Consistency Learning (UPCL) and Progressive Consistency Weighted Assemble (PCWA). By integrating local and global cues, UIA-ViT achieved strong cross-dataset generalization and became a baseline for subsequent transformer-based detectors [23, 57].

Following this, several works have refined transformer-based detection using pre-trained ViT backbones and domain adaptation. Luo et al. (2023) introduced FA-ViT, a framework that freezes the ViT backbone and trains global, local, and fine-grained adaptive modules to detect forgery-specific clues. This design improves generalization to unseen manipulation types, achieving strong cross-dataset performance on Celeb-DF and DFDC. Similarly, Swin-Transformer-based methods have been used to combine hierarchical feature extraction with global attention, achieving better robustness to unseen manipulation types [57]. Collectively, these studies show that attention mechanisms can capture transferable, high-level semantic representations that generalize across manipulation techniques and domains [56, 58].

However, transformer models also have limitations. They require extensive training data and fine-tuning to perform well and may still be sensitive to test-time degradations such as compression, blur, or noise [59, 42]. Despite these challenges, transformer-based architectures mark a major step toward more generalizable deepfake detection systems [23, 58].

Despite improved generalization across datasets, transformer models still faltered under real-world degradations such as compression, packet loss, and transmission noise. Consequently, the focus of recent work has shifted from architectural sophistication to robustness-oriented design, emphasizing resilience to corruptions, domain shift, and unseen perturbations. [57, 60]

3 Robustness-Oriented Detection

3.1 Common Real-World Corruptions

Deepfake detectors that achieve strong in-distribution accuracy often degrade sharply when exposed to realistic distortions such as compression, blur, or noise. In deployment, deepfake videos are rarely pristine—they are transmitted across social platforms, messaging apps, or variable-bandwidth networks that introduce severe quality loss. Common corruptions include *JPEG compression*,

Gaussian and shot noise, *motion blur*, *pixelation*, and *defocus*. These distortions significantly alter local texture and frequency cues that detectors rely upon. [17]

Hendrycks and Dietterich's *ImageNet-C* benchmark [61] systematically quantified the impact of common corruptions on model performance, demonstrating that even minor distortions can lead to substantial drops in classification accuracy. Consistent trends are observed in deepfake forensics: empirical analyses from DeepfakeBench [42] report that detectors trained on high-quality data (e.g., FF++ c23) typically achieve AUC values around 0.95, yet their performance declines to approximately 0.75–0.80 when evaluated under compressed or cross-domain conditions (e.g., FF++ c40, CelebDF-v2, DFDC). This corresponds to an estimated 40–50% reduction in detection reliability. Such brittleness underscores that many existing detectors overfit to their training distributions and fail to capture the invariances necessary for deployment in real-world scenarios.

With the increasing realism of deepfakes, the focus of recent research has shifted toward achieving **robustness across varying visual qualities**, moving beyond improvements confined to dataset-specific accuracy. The following subsection reviews representative approaches developed to address this challenge.

3.2 Cross-Quality and Robust Methods

Several recent methods aim to reduce sensitivity to degradation by training on data of different quality levels or by designing architectures that learn quality-invariant features. The following works illustrate these approaches to robustness-oriented detection.

ADD (AAAI 2022): Frequency-Aware Distillation for Robust Detection. The *Attention-based Deepfake detection Distiller (ADD)* [20] was proposed to address the severe performance drop of existing detectors on low-quality or highly compressed deepfakes. ADD adopts a teacher–student knowledge distillation design that transfers spectral and spatial cues from a high-quality (HQ) teacher to a student trained on low-quality (LQ) inputs. ADD introduces two modules: *Frequency Attention Distillation (FAD)* and *Multi-View Attention Distillation (MAD)*. FAD helps the student model recover details lost by compression by using a frequency-domain attention map. It applies a

Discrete Fourier Transform to feature maps so the student can focus on missing high-frequency cues like textures and edges. [20] MAD complements this by aligning the teacher's and student's feature distributions using the Sliced Wasserstein Distance, creating multiple random views that preserve spatial correlations. Together, FAD and MAD improve the detector's ability to recognize deepfakes even in low-quality, heavily compressed images. During training, the teacher model is first trained on high-quality images, while the student learns from compressed or low-quality versions of the same data. The student is guided by two loss terms that balance frequency and multi-view attention, helping it recover lost details and match the teacher's feature patterns. [20] Because these modules do not depend on any specific network design, they can be easily added to standard CNN-based detectors.

Comprehensive experiments on five FaceForensics++ datasets—NeuralTextures, DeepFakes, Face2Face, FaceSwap, and FaceShifter—demonstrated ADD's effectiveness under compression. On the challenging NeuralTextures (c40) subset, accuracy improved from 60.27% (ResNet50 baseline) to **68.53%**, surpassing existing distillation baselines such as FitNet, Attention Transfer, and Non-Local modules. Across datasets, ADD yielded 1–6% gains under medium compression (c23) and up to 8% under heavy compression (c40). Ablation results confirmed that FAD alone improved accuracy by 6.8%, while integrating MAD contributed an additional 2.5%. By jointly restoring high-frequency information and multi-view feature consistency, ADD achieved superior robustness to low-bitrate distortions compared to prior CNN- and frequency-based detectors. [20] These findings highlight the importance of combining spectral and spatial cues to maintain detection reliability under heavy compression.

QAD (ICCV 2023): Quality-Agnostic Deepfake Detection. Le and Woo [21] proposed *Quality-Agnostic Deepfake Detection (QAD)*, a unified framework designed to handle both high- and low-quality manipulations within a single model. Unlike earlier methods that train separate detectors for each compression level, QAD employs an *intra-model collaborative learning* strategy in which the same image at different quality levels is processed jointly to align their feature representations. To achieve this, QAD maximizes the dependency between high- and low-quality embeddings using the

Hilbert–Schmidt Independence Criterion (HSIC), encouraging geometrical consistency of features across varying qualities. This helps the model learn features that stay consistent across different quality levels, focusing on authenticity instead of appearance. In addition, an *Adversarial Weight Perturbation (AWP)* module is introduced to improve generalization by flattening the weight–loss landscape, making the network less sensitive to compression noise and signal distortions.

The combination of HSIC-based representation alignment and AWP-based regularization enables QAD to disentangle content authenticity from image quality, producing stable predictions even under heavy degradation. Extensive experiments across seven datasets—including FaceForensics++, CelebDF-v2, and FFIW10K—demonstrated the framework's superior robustness. For instance, on the NeuralTextures subset, QAD-E (EfficientNet-B1 backbone) achieved an AUC of 94.9%, outperforming frequency- and texture-based baselines such as ADD [20] and BZNet [25] by more than 5% under strong compression. Ablation studies further showed that combining HSIC and AWP improved the AUC from 88.2% to 91.3%, confirming their complementary roles.

Overall, QAD addresses the quality variance problem that limits most detectors. By jointly learning invariant embeddings and robust optimization behavior, it provides a scalable, architecture-agnostic approach for real-world deepfake detection, where input quality and compression often vary unpredictably.

BZNet (WWW 2022): Branch Zooming Network for Low-Quality Detection. Lee, An, and Woo [25] proposed *BZNet*, a branch zooming network designed specifically to detect deepfakes in low-quality (LQ) and highly compressed videos. Recognizing that compression destroys fine facial cues critical for detection, BZNet reconstructs missing details through an unsupervised *Branch Zooming (BZ)* module followed by a multi-scale classification stage. The BZ module contains one encoder and multiple decoders that generate super-resolved versions of the input face at different resolutions (e.g., 128×128, 192×192, and 256×256). All branches are trained simultaneously, allowing the network to learn diverse spatial representations from multiple zoom levels. Skip connections between the encoder and each decoder, together with residual blocks, help maintain global information while recovering local textures that are degraded by compression. This

design enables BZNet to restore fine facial details—particularly along boundaries and specular regions—that conventional detectors fail to capture under heavy compression.

In the second phase, BZNet performs *multi-scale learning* using two fusion strategies: *best-scale (BS)* and *self-ensemble (SE)*. The BS method selects the scale with the highest confidence score for the final decision, whereas SE averages the confidence scores of all scales. Combining multiple scales improves BZNet's robustness under low-quality and compressed conditions, as averaging helps reduce the influence of corrupted features at specific scales [25].

Quantitative experiments on FaceForensics++ (C40) demonstrated that BZNet significantly outperforms conventional baselines such as Xception, ResNet, and F3-Net [25]. With an Xception backbone, BZNet-SE achieved an average accuracy of 85.33% and 70.78% on the challenging NeuralTextures subset, compared to 82.09% and 66.86% for the baseline, respectively [25, Table 2]. On the Deepfake-in-the-Wild (DW) dataset, BZNet-SE further improved accuracy from 62.78% to 69.33%, demonstrating strong generalization to unseen, low-quality videos [25, Table 2].

Overall, BZNet demonstrated that structural design can play a crucial role in improving robustness against compression, pixelation, and other low-quality distortions. By integrating unsupervised super-resolution with multi-scale feature learning, the model effectively recovered discriminative information that is often lost in heavily compressed videos.

DPL (ACCV 2024): Dual Progressive Learning. Zhang et al. [22] introduced *Dual Progressive Learning (DPL)*, a framework designed to detect deepfakes across different video quality levels. The idea is that low-quality videos hide fake traces more deeply, so the model needs to learn in steps—starting with easier, clearer examples and then moving to harder, more degraded ones. DPL has two main parts: *Visual Quality Guided Progressive Learning (VQPL)* and *Forgery Identifiability Guided Progressive Learning (FIPL)*. Each part focuses on a different challenge: one on handling image quality, and the other on how noticeable the manipulation is. The VQPL branch uses a CLIP-based *Visual Quality Indicator (VQI)* to decide how many progressive steps to run for each video. Lower-quality inputs receive more steps, while higher-quality ones need fewer [22, Sec. 3.1]. A shared *Feature Selection Module (FSM)* helps the model focus on difficult features by masking

easy ones first and refining the rest in later steps [22, Sec. 3.1]. The FIPL branch follows the same idea but uses a CLIP-based *Forgery Identifiability Indicator (FII)* to sort videos from weak to strong forgeries [22, Sec. 3.2]. Both branches share parameters, and their results are merged by adding the outputs together.

Quantitative experiments on FaceForensics++ demonstrated that DPL consistently outperformed recent robustness-oriented detectors such as QAD, UIA-ViT, and ADD [22, Tab. 2]. Trained on FF++(c23) and tested on compressed FF++(c40), DPL achieved an average AUC of 86.36%, surpassing UIA-ViT's 85.75% and QAD's 82.70%. Across individual subsets, it reached 94.22% on DeepFakes, 86.26% on Face2Face, 92.88% on FaceSwap, and 72.09% on NeuralTextures. Cross-dataset evaluations on FSH, CelebDF-v2, and FFIW10K further reported an average AUC of 70.69%, confirming strong generalization under unseen conditions [22, Tab. 4]. Overall, DPL demonstrated that progressive learning across quality and forgery difficulty substantially improves robustness against compression and quality degradation.

These results confirm that progressive learning across both quality and forgery difficulty improves robustness to compression, [22] though its reliance on discrete quality bins limits adaptation to unseen degradations.

3.3 Limitations of Current Robustness Approaches

Despite substantial progress, existing robustness-oriented detectors still face several challenges. Most methods depend on discrete training qualities or predefined augmentation sets, making them vulnerable to unseen degradations such as codec artifacts, transmission noise, or dynamic bitrate changes. [20, 21, 22, 25, 42, 60]

Architectural solutions like BZNet and DPL improve performance within trained ranges but still assume fixed quality levels. [22, 25]

Distillation-based models such as ADD and QAD enhance low-quality robustness but require separate teacher models and large paired datasets, limiting scalability. [20, 21]

Furthermore, none of these frameworks explicitly model uncertainty at test time.

These limitations motivate adaptive probabilistic approaches—such as the Mixture Density

Network (MDN)-guided ensemble proposed in this work—that estimate continuous corruption distributions and provide confidence-aware predictions under real-world conditions.

4 Present Work — MDN-Guided Noise-Aware Modular Ensemble

4.1 Motivation

Deepfake detectors trained only on clean data or limited noise conditions often fail when they encounter distortions not seen during training. These distortions—such as compression loss or transmission noise—can easily mislead models that make rigid, deterministic predictions. Deterministic models produce a single, fixed output for a given input, assuming that the input is clean and reliable. They do not consider uncertainty or the possibility that the input might be corrupted. As a result, even small changes in video quality can cause their predictions to shift dramatically.

To overcome this, our approach treats detection as a probabilistic process rather than a fixed decision. We model uncertainty explicitly by estimating the distribution of possible corruptions that might affect an input video. This allows the system to adapt its predictions based on the level and type of noise present at test time. By combining a Mixture Density Network (MDN) with Parameter-Efficient Fine-Tuning (PEFT) modules, the system learns to recognize and adapt to varying noise levels instead of relying on a fixed model configuration.

4.2 Mixture Density Networks (MDNs)

Traditional deepfake detectors generate a single deterministic output—typically a binary probability of whether a frame is real or fake. While effective in controlled settings, this fixed prediction fails to represent the inconsistencies and distortions present in real-world data. When videos are degraded by compression, blur, or transmission noise, the detector's confidence becomes unreliable, as it cannot model ambiguity in corrupted inputs. To overcome this limitation, our framework incorporates a Mixture Density Network (MDN) that explicitly models prediction uncertainty by learning a probability distribution over possible noise conditions.

An MDN, first proposed by Bishop (1994) [62], replaces fixed outputs with the parameters of a Mixture of Gaussians (MoG)—specifically, the component means (μ), variances (σ), and mixing coefficients (α). Instead of assigning a single prediction, the MDN estimates how likely

each Gaussian component contributes to the observed input, forming a probabilistic representation of the underlying data distribution. This approach has been successfully applied in various computer vision domains to capture ambiguity, such as multi-hypothesis 2D/3D human pose estimation [63] and object detection under uncertainty.

Our work introduces a new use of MDNs for deepfake forensics. We use them to estimate the corruption or noise distribution present at test time. The MDN learns from synthetically corrupted deepfake images, approximating real-world distortions as mixtures of Gaussian components. Each component corresponds to a distinct corruption mode—such as Gaussian noise, motion blur, or compression—captured through its mean and variance.

At inference, the MDN functions as a probabilistic noise estimator, identifying which mixture components best represent the distortions present in the input. The predicted parameters guide the model in selecting and combining the right fine-tuned modules so it can adapt to different or unexpected noise levels. This probabilistic approach forms the core of our adaptive, noise-aware deepfake detection framework.

4.3 Parameter-Efficient Fine-Tuning (PEFT)

Large-scale deep learning models are often computationally expensive to retrain for every new domain or deployment environment. *Parameter-Efficient Fine-Tuning (PEFT)* methods tackle this problem by adding small, trainable modules that let pre-trained networks adapt to new data without changing their main parameters [64]. Among the most widely adopted approaches are **Low-Rank Adaptation (LoRA)** [65], **Weight-Decomposed Low-Rank Adaptation (DoRA)** [66], and **Parameter-Efficient Fine-Tuning with Singular Vectors (SVFT)** [67]. These techniques have demonstrated strong performance across both natural-language and vision-transformer architectures, enabling models to specialize efficiently while preserving generalization and reducing computational cost.

In our framework, PEFT modules form the backbone of modular noise adaptation. Each module is trained on data with a specific level of Gaussian noise, helping it learn precise feature adjustments for that noise level. The base detector (e.g., XceptionNet, UIA-ViT) remains frozen

during this process, while the small adapter modules focus on learning noise-specific adjustments. During inference, the **Mixture Density Network (MDN)** estimates the characteristics of the observed noise and accordingly selects or combines the relevant PEFT modules. This integration allows the detector to adapt dynamically to varying or previously unseen noise environments, enhancing robustness without the need for full retraining.

This design combines efficient fine-tuning with probabilistic control, allowing the detector to adapt to unseen or mixed noise conditions. It achieves robustness similar to full retraining while using much less computation and memory.

4.4 MDN-Guided Modular Ensemble Architecture

The complete framework consists of two training stages followed by MDN-guided inference:

Stage 1 — MDN Pre-training. In the first stage, the Mixture Density Network (MDN) is trained to learn how noise behaves in corrupted deepfake frames. It does this by modeling the differences (residuals) between clean and noisy samples as a combination of several Gaussian distributions. The network predicts three key parameters for each distribution—the mean (μ), variance (σ), and mixing coefficient (α)—which together describe the different types and intensities of corruption found in the data.

Stage 2 — Training PEFT Modules. Next, once the main detector is fixed (its weights are frozen), we attach several small, trainable PEFT modules. Each module is trained on data with a specific amount of added Gaussian noise so that it can adapt for that noise level. This approach allows the model adjust to various noise conditions while preserving what the main network has already learned.

Inference — Dynamic ensemble selection. During inference, the pre-trained MDN first estimates how much noise or distortion is present in the test input. For each noise type it detects, the system selects the PEFT module that was trained on a similar noise level. The final prediction is then made by combining the outputs of these modules, weighted according to the MDN's mixing coefficients. This means that modules trained on noise conditions most similar to the input have a stronger influence on the final result.

This adaptive inference process allows the detector to adjust smoothly across different noise levels, maintaining reliable performance even under severe or unseen distortions. The approach combines probabilistic noise estimation with modular adaptation, achieving robustness similar to full retraining while staying computationally efficient.

Model	Noise-free	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Contrast	Elastic	Pixel	JPEG	Avg. Corr.
XNet-Base	96.11	56.11	63.39	55.16	87.60	64.22	88.56	91.16	84.26	76.57	92.33	64.47	74.89
XNet-Random Ens.	92.83	81.33	87.28	79.90	88.46	87.23	89.24	91.27	84.82	92.02	91.87	86.57	87.27
XNet-MDN Ens.	**96.34**	**89.33**	**93.09**	**86.57**	**92.56**	**91.17**	**94.23**	**94.06**	**87.82**	**93.46**	**95.80**	**89.49**	**91.59**

Table 1: Video-level AUC (%) for deepfake detectors (XceptionNet, UIA-ViT) trained on the *Celeb-DF v1* dataset [29]. The base classifier is trained on clean (noise-free) data, while random and MDN ensembles are trained on data augmented with Gaussian noise.

The XceptionNet results in Table 1 highlight the clear advantage of integrating uncertainty modeling through the MDN-guided ensemble. While the baseline detector shows a severe performance drop under Gaussian and impulse noise, the MDN ensemble consistently maintains an AUC above 90%, showing strong resilience to unseen distortions. Unlike the random ensemble, which offers limited stability across varying conditions, the MDN-guided approach adapts dynamically to unseen corruptions achieving more reliable detection under real-world distortions.

4.5 Discussion and Remaining Gaps

Despite these advances, several limitations remain. First, most current frameworks—including the proposed MDN-guided ensemble—assume that corruption follows a Gaussian distribution, which simplifies modeling but fails to represent complex or adversarial perturbations introduced by modern generative systems [68]. Extending uncertainty modeling to non-Gaussian or adversarial distributions is a promising direction.

Second, emerging multimodal deepfakes that jointly manipulate visual, audio, and textual modalities highlight the need for cross-modal consistency modeling [69]. Current architectures remain largely unimodal and are unable to detect inconsistencies across modalities.

Finally, the computational overhead of probabilistic ensembles can hinder real-time deployment. Techniques such as lightweight MDN inference, online adaptation, and model compression could address scalability and latency constraints [70, 71, 60].

Figure 2: Examples of different corruption types applied to fake frames from the *Celeb-DF v1* test set. The outputs are generated using the MDN-guided XceptionNet ensemble, where each module was trained on *Celeb-DF v1* data with varying levels of Gaussian noise. The base model trained on clean data misclassifies these corrupted samples as real, while the MDN-guided ensemble correctly identifies them as fake.

Addressing these challenges—non-Gaussian modeling, multimodal integration, and real-time efficiency—will be crucial for developing next-generation uncertainty-aware deepfake detection systems capable of robust performance in diverse, real-world conditions.

5 Datasets for Deepfake Detection

5.1 Overview

The progress of deepfake detection has closely followed the development of benchmark datasets. These datasets define the experimental standards that guide model design and comparison. Early benchmarks focused on small, controlled manipulations, while later ones expanded scale, diversity, and realism. More recent efforts introduced synthetic degradations such as compression or blur to study robustness. However, across all publicly available datasets, none capture authentic, time-varying distortions caused by real-world network transmission, leaving detector reliability under live streaming conditions largely unexplored [42, 72, 43].

5.2 Early Controlled Datasets

The first generation of deepfake benchmarks focused on basic manipulation pipelines and fixed acquisition conditions. UADFV introduced a small collection of face-swapped videos generated using early autoencoder-based models, providing one of the earliest testbeds for deepfake detection [29]. DeepfakeTIMIT advanced early autoencoder-based datasets by introducing both low- and high-resolution GAN variants trained on matched subject pairs from VidTIMIT, resulting in 620 face-swapped videos with consistent backgrounds and synchronized audio[73]. These datasets enabled reproducible evaluation but lacked realism: all videos were captured offline with static camera parameters and minimal compression. Consequently, they fail to reflect the distortions.

5.3 Benchmark Datasets

The next major milestone came with **FaceForensics++ (FF++)** [72], which provided 1,000 real and 4,000 forged videos generated using four manipulation techniques—Face2Face, FaceSwap, DeepFakes, and NeuralTextures. To simulate realistic media transmission, FF++ applied H.264 compression at both high (Q=23) and low (Q=40) quality levels, establishing one of the most widely used benchmarks for deepfake training and evaluation. **Celeb-DF v1/v2** [29] improved synthesis realism by increasing face resolution, refining mask boundaries, and applying temporal smoothing to eliminate visible blending artifacts. Additional datasets, such as **FaceShifter** and **Google DFD**, expanded the manipulation variety and identity coverage [74, 75]. While these datasets offered high visual quality and consistency, they were generated under controlled pipelines with static compression settings. Consequently, they do not capture the dynamic variations that occur during real-world video transmission.

5.4 Large-Scale and Multimodal Datasets

To address limitations in diversity and scale, large datasets such as the **DeepFake Detection Challenge (DFDC)** and **FakeAVCeleb** introduced tens of thousands of manipulated videos covering diverse lighting, pose, and speech conditions [76, 77]. DFDC remains one of the largest benchmarks, containing over 128,000 videos generated from 3,426 actors using eight manipulation methods. FakeAVCeleb extends the evaluation scope to audio-visual forgeries by including synchronized

voice and video manipulations. These datasets improved demographic balance and environmental variability but still rely on pre-defined compression pipelines. The degradations remain static and dataset-specific—none reflect the temporal fluctuations, bitrate shifts, or jitter typical of live streaming or mobile communication.

5.5 Robustness-Oriented Datasets

To evaluate detector performance under degraded conditions, several robustness-focused datasets introduced synthetic perturbations. **DeeperForensics-1.0** [78] applied seven controlled distortions—color and contrast shifts, Gaussian blur, white noise, and multiple compression levels—across 60,000 videos to simulate real-world imperfections. **ForgeryNet** [79] provided 2.9 million images and 221,000 videos covering seven image-level and eight video-level manipulation types, augmented with 36 controlled perturbations, for both classification and localization. **WildDeepfake** [80] further advanced realism by collecting real and fake videos directly from the internet, exposing detectors to uncontrolled re-encoding and compression artifacts. Despite these efforts, all manipulations are either synthetically injected or passively observed after upload. None capture authentic, temporally varying artifacts caused by real-time network congestion, packet loss, or adaptive bitrate streaming—conditions that strongly affect deepfake reliability in deployment.

5.6 Identified Gap and Motivation for a New Dataset

Figure 3 illustrates the core limitation of existing deepfake datasets. The left frame shows a clean CelebDF-v1 sample, while the right shows the same frame after being transmitted through our WebRTC testbed under constrained bandwidth, 75 ms delay, and 10 ms jitter. The resulting block artifacts and facial distortions are not the product of synthetic compression or additive noise—they arise from real packet loss, jitter propagation, and encoder adaptation during live transmission. Current datasets cannot reproduce these network-driven degradations.

To address this gap, we construct a new dataset that captures *real network-induced distortions* by streaming deepfake videos over Wi-Fi and cellular links with controlled bandwidth, jitter, and loss profiles. The resulting benchmark—**DeFND (Deepfake Forensics under Network Degradation)**—provides the first controlled platform for evaluating detector robustness under

Clean Noisy (75 ms delay, 10 ms jitter)

Figure 3: Example of real network-induced distortion. The clean CelebDF-v1 frame (left) is compared to the same frame after transmission via RTP over UDP with `tc-tbf` rate limiting and `tc-netem` simulating delay and jitter. The degradation patterns arise from real packet loss rather than synthetic corruption.

authentic transmission conditions, enabling the development of models that are genuinely network-aware and deployment-ready.

6 Present Work — Network Degradation Dataset (DeFND)

6.1 Motivation

While benchmark datasets such as FaceForensics++, Celeb-DF, and DFDC have advanced the field of deepfake detection, they are all captured or synthesized under controlled, offline conditions. None reproduce the temporal distortions and compression noise that occur when videos are transmitted over real network connections. However, in deployment, deepfakes are frequently shared or streamed over Wi-Fi and cellular links where packet loss, bitrate drops, and adaptive encoding severely distort video quality. These network-induced degradations can obscure forgery traces and lead to sharp performance drops in existing detectors. To address this gap, our ongoing work introduces **DeFND — Deepfake Detection under Network Degradation**, the first benchmark specifically designed to evaluate model robustness under realistic streaming conditions.

6.2 Dataset Construction

The DeFND dataset is built using a custom video-call client based on the **WebRTC** framework, which is the standard used by modern conferencing platforms. Instead of live camera feeds, the sender transmits prerecorded deepfake videos from standard datasets (FaceForensics++, FaceShifter, UADFV, Celeb-DF v1/v2), while the receiver records the network-delivered stream. Each experiment is conducted over five real network environments:

- **Wired LAN (L0):** 1 Gbps baseline connection with near-zero packet loss.
- **Office Wi-Fi (W1):** 5 GHz high-quality link with $\sim$0.5% loss.
- **Campus Wi-Fi (W2):** 2.4 GHz link with 3.2% loss and higher latency.
- **5G Cellular (C1):** Outdoor high-bandwidth sub-6 GHz network.
- **4G/LTE (C2):** Indoor low-bandwidth network with $\sim$4% loss.

Each original video is streamed under these conditions, reassembled, and labeled by network type. This results in five corrupted variants per source clip, capturing diverse degradation modes such as lossy encoder compression, decoder error concealment, and bitrate adaptation. Together, the dataset includes approximately **67,650** network-corrupted videos derived from five baseline datasets and covering multiple manipulation types and demographic groups.

6.3 Key Contributions

- **Authentic network degradations:** Captures temporal and spatial artifacts caused by real Wi-Fi and cellular streaming, unlike previous synthetic-noise datasets.
- **Cross-dataset design:** Integrates multiple existing deepfake benchmarks to ensure diversity across manipulation techniques and subjects.
- **Evaluation benchmark:** Enables standardized testing of detectors under realistic streaming conditions, supporting both in-domain and cross-domain analysis.

The DeFND benchmark serves as a bridge between laboratory training and real-world deployment, providing the first reproducible framework to study how network transmission affects deepfake detection reliability.

7 Conclusion

This work presented a comprehensive review of deepfake detection research, emphasizing robustness, generalization, and uncertainty modeling. The study traced the field's evolution from spatial and temporal CNN-based detectors to frequency-, latent-, and transformer-driven architectures, highlighting the persistent performance gap between benchmark datasets and real-world deployment.

To bridge this divide, two research directions were explored. The first, the **MDN-Guided Noise-Aware Modular Ensemble**, advances probabilistic detection by modeling corruption distributions and dynamically re-weighting parameter-efficient modules to adapt to unseen noise. The second, **DeFND**, introduces the field's first dataset capturing real network-induced degradations across Wi-Fi and cellular conditions, enabling reproducible evaluation of detectors under authentic transmission noise.

Together, these contributions address both data and model limitations, offering a framework that unites realistic evaluation with adaptive, uncertainty-aware inference. Future work will extend this foundation toward multimodal deepfakes, non-Gaussian noise modeling, and lightweight real-time inference—key steps toward deploying trustworthy deepfake detection systems in complex, real-world environments.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[2] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162–8171, PMLR, 2021.

[3] O. C. Phukan, Girish, M. M. Akhtar, S. R. Behera, P. Mallick, P. B. Reddy, A. B. Buduru, and R. Sharma, "Towards source attribution of singing voice deepfake with multimodal foundation models," 2025. arXiv preprint arXiv:2505.02412.

[4] C. Drolsbach and N. Pröllochs, "Characterizing AI-generated misinformation on social media," 2025. arXiv preprint arXiv:2505.10266.

[5] S. H. Al-Khazraji, H. H. Saleh, A. I. Khalil, and I. A. Mishkal, "Impact of deepfake technology on social media: Detection, misinformation and societal implications," in *The Eurasia Proceed- ings of Science, Technology, Engineering & Mathematics (EPSTEM)*, vol. 23, pp. 429–441, ISRES Publishing, 2023.

[6] M. S. M. AlShamsi, "The role of ai in digital propaganda and its influence on public political perception: A case study of the uae citizen," Master's thesis, Zayed University, 2025. Master's Thesis, Zayed University, UAE.

[7] A. V., P. Joy, *et al.*, "Deepfake detection using xceptionnet," in *2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE)*, (Kerala, India)

[8] E. P. Boediman, "Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review," *Jurnal Komunikasi*, vol. 19, pp. 313–334, April 2025.

[9] R. Shao, J. Ni, Z. Zhang, J. Zhou, and Z. Liu, “Deepfake detection in the era of diffusion models: A comprehensive survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. Early Access, IEEE TPAMI.

[10] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

[11] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” *arXiv preprint arXiv:1809.00888*, 2018.

[12] A. Heidari *et al.*, “Deepfake detection using deep learning methods: A systematic and compre- hensive review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2024.

[13] M. Singh *et al.*, “Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges,” *Journal of Information Security and Applications*, 2025.

[14] T. Kim *et al.*, “Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection,” *arXiv preprint arXiv:2507.02398*, 2025.

[15] I. Amerini *et al.*, “Deepfake media forensics: Status and future challenges,” *Journal of Imaging*, 2025.

[16] R. Ramanaharan *et al.*, “Deepfake video detection: Insights into model generalisation — a systematic review,” *Journal of Digital Forensics*, 2025.

[17] A. Garg *et al.*, “Beyond the benchmark: Generalization limits of deepfake detectors in the wild,” tech. rep., 2023.

[18] P. Chen *et al.*, “Detecting compressed deepfake images using two-branch convolutional networks with similarity and classifier,” *Symmetry*, vol. 14, no. 12, 2022.

[19] M. Li *et al.*, “Pay less attention to deceptive artifacts: Robust detection of compressed deepfakes on online social networks,” *arXiv preprint arXiv:2506.20548*, 2025.

[20] B. M. Le and S. S. Woo, “Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1163–1171, 2022.

[21] B. M. Le and S. S. Woo, “Quality-agnostic deepfake detection with intra-model collaborative learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22454–22463, 2023.

[22] D. Zhang *et al.*, “Dpl: Cross-quality deepfake detection via dual progressive learning,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 567–582, 2024.

[23] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, and N. Yu, “Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection,” 2022.

[24] D. A. Coccomini *et al.*, “On the generalization of deep learning models in video deepfake detection,” 2023.

[25] S. Lee, J. An, and S. S. Woo, “Bznet: Unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos,” in *Proceedings of the ACM Web Conference (WWW)*, pp. 1532–1541, 2022.

[26] S. Yang, H. Guo, S. Hu, B. Zhu, Y. Fu, S. Lyu, X. Wu, and X. Wang, “Crossdf: Improving cross-domain deepfake detection with deep information decomposition,” 2024.

[27] A. Yermakov, J. Cech, J. Matas, and M. Fritz, “Deepfake detection that generalizes across benchmarks,” 2025.

[28] K. Sun, S. Chen, T. Yao, H. Liu, X. Sun, S. Ding, and R. Ji, “Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion,” 2024.

[29] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” 2020.

[30] B. Bayar and M. C. Stamm, “Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection,” vol. 13, pp. 2691–2706, 2018.

[31] R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions,” 2020.

[32] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake video detection through opti- cal flow based cnn,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1205–1211, IEEE, 2019.

[33] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, IEEE, 2017.

[34] C. Shuai, G. Wang, K. Pan, T. Wu, F. Jin, H. Tan, M. Li, Z. Liu, F. Lin, and K. Ren, “Morphology-optimized multi-scale fusion: Combining local artifacts and mesoscopic seman- tics for deepfake detection and localization,” *arXiv preprint arXiv:2509.13776*, 2025.

[35] S. Tariq, S. Lee, and S. S. Woo, “One detector to rule them all: Towards a general deepfake detection framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12842–12852, 2021.

[36] B. Bayar and M. C. Stamm, “Design principles of convolutional neural networks for multi-media forensics,” in *Proceedings of the IST International Symposium on Electronic Imaging: Media Watermarking, Security, and Forensics*, pp. 1–10, 2017.

[37] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, “Deep high-resolution representation learning for visual recognition,”

IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 43, no. 10, pp. 3349–3364, 2020.

[38] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 2019.

[39] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," 2020.

[40] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," 2021.

[41] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," 2020.

[42] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: A comprehensive benchmark of deepfake detection," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[43] T. Wang, X. Liao, K. P. Chow, X. Lin, and Y. Wang, "Deepfake detection: A comprehensive survey from the reliability perspective," *ACM Comput. Surv.*, vol. 57, Nov. 2024.

[44] D. Güera, P. Bestagini, S. Tubaro, and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of the 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, pp. 1–6, IEEE, 2018.

[45] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," 2018.

[46] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," 2019.

[47] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li, “Altfreezing for more general video face forgery detection,” 2023.

[48] M. Astrid, E. Ghorbel, and D. Aouada, “Audio-visual deepfake detection with local temporal inconsistencies,” *arXiv preprint arXiv:2501.08137*, 2025.

[49] S. J. Sangwine and A. L. Thornton, “Frequency domain methods,” in *The Colour Image Processing Handbook*, pp. 229–258, Chapman & Hall, 1998.

[50] B. M. Le and S. S. Woo, “Add: Frequency attention and multi-view based knowledge distilla- tion to detect low-quality compressed deepfake images,” 2021.

[51] S. T. Erukude, V. C. Marella, and S. R. Velur, “Fourier-based gan fingerprint detection using ResNet50,” 2025. arXiv:2510.19840, License: arXiv.org perpetual non-exclusive license.

[52] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, “Frequency-aware deepfake detection: Im- proving generalizability through frequency space learning,” *arXiv preprint arXiv:2403.07240*, 2024.

[53] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, “Latent space autoregression for novelty detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 481–490, 2019.

[54] M. Delmas and R. Seguier, “Latentforensics: Towards frugal deepfake detection in the stylegan latent space,” 2024.

[55] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, “Exploiting style latent flows for generalizing deepfake video detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1133–1143, 2024.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” 2021.

[58] A. Luo, R. Cai, C. Kong, Y. Ju, X. Kang, J. Huang, and A. C. Kot, “Generalized face forgery detection via adaptive learning for pre-trained vision transformer,” 2024.

[59] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, “A timely survey on vision transformer for deepfake detection,” 2024.

[60] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” 2019.

[61] C. M. Bishop, “Mixture density networks,” Tech. Rep. NCRG/94/004, Neural Computing Research Group, Aston University, Birmingham, U.K., 1994.

[62] C. Li and G. H. Lee, “Generating multiple hypotheses for 3d human pose estimation with mixture density network,” 2019.

[63] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, “Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment,” 2023.

[64] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021.

[65] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, “Dora: Weight-decomposed low-rank adaptation,” 2024.

[66] V. Lingam, A. Tejaswi, A. Vavre, A. Shetty, G. K. Gudur, J. Ghosh, A. Dimakis, E. Choi, A. Bojchevski, and S. Sanghavi, “Svft: Parameter-efficient fine-tuning with singular vectors,” 2024.

[67] A. Gandhi and S. Jain, “Adversarial perturbations fool deepfake detectors,” 2020.

[68] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, "A multimodal framework for deepfake detection," in *Proceedings of the International Conference on Intelli- gent Systems, Smart and Green Technologies (ICISGT)*, pp. 1122–1133, IEEE, 2023. Equal authorship: Kashish Gandhi and Prutha Kulkarni.

[69] T. Anusha and A. Srinagesh, "Deepfake video detection: A comprehensive survey of advanced machine learning and deep learning techniques to combat synthetic video manipulation," in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, pp. 1033–1041, 2025.

[70] S. Banerjee, S. K. Yadav, A. Dhara, and M. Ajij, "A survey: Deepfake and current technologies for solutions," in *Proceedings of the International Conference on Recent Trends in Informa- tion Technology and Computer Science (ICRTITCS 2023)*, vol. 3900 of *CEUR Workshop Proceedings*, pp. 79–86, CEUR-WS.org, 2023.

[71] I. Balafrej and M. Dahmane, "Enhancing practicality and efficiency of deepfake detection," *Scientific Reports*, vol. 14, no. 1, p. 31084, 2024.

[72] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," 2019.

[73] P. Korshunov and S. Marcel, "Deepfaketimitn," in *2018 International Conference on Biomet- rics (ICB)*, 2018.

[74] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," 2020.

[75] G. Research, "Contributing data to deepfake detection research." https://research.google/blog/contributing-data-to-deepfake-detection-research/, Sept. 2019.

[76] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge (dfdc) dataset,” 2020.

[77] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, “Fakeavceleb: A novel audio-video multimodal deepfake dataset,” in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets Benchmarks Track*, 2021.

[78] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, “Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection,” 2020.

[79] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, “Forgerynet: A versatile benchmark for comprehensive forgery analysis,” 2021.

[80] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “Wilddeepfake: A challenging real-world dataset for deepfake detection,” 2024.