A SEMINAR REPORT ON

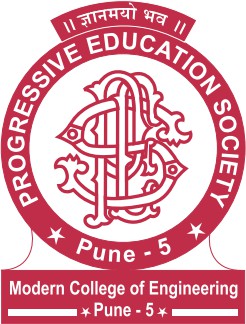
**“Towards Robust Detection of AI-Generated Videos: Challenges and Continual Learning Solutions”**

SUBMITTED BY

Yash Doke (Seat No.)

UNDER THE GUIDANCE OF

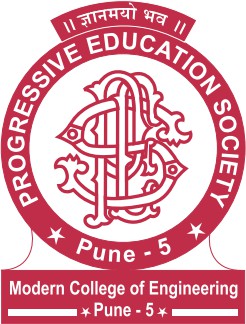
Dr. B. D. Phulpagar



DEPARTMENT OF COMPUTER ENGINEERING

P.E.S. MODERN COLLEGE OF ENGINEERING PUNE - 411005.

[2025 - 26]



Progressive Education Society’s **Modern College of Engineering** Department of Computer Engineering Shivajinagar, Pune - 411005.

**CERTIFICATE**

This is to certify that **Yash Doke** from Third Year Computer Engineering has successfully completed his / her seminar work titled **“Towards Robust Detection of AI-Generated Videos: Challenges and Continual Learning Solutions”** at PES Modern College of Engineering in the partial fulfillment of the Bachelors Degree in Computer Engineering under Savitribai Phule Pune University.

Date:

Dr. B. D. Phulpagar Prof. Dr. Mrs. S. A. Itkar

Guide Head

Department of Computer Engineering

# Acknowledgement

It gives me pleasure in presenting the seminar report on **‘Towards Robust Detection of AI-Generated Videos: Challenges and Continual Learning Solutions’**.

Firstly, I would like to express my indebtedness appreciation to my guide **Dr. B. D. Phulpagar**. His constant guidance and advice played very important role in successful completion of the report. He always gave me his/her suggestions, that were crucial in making this report as flawless as possible.

I would like to express our gratitude towards **Prof. Dr. Mrs. S. A. Itkar**, Head of Computer Engineering Department, PES Modern College of Engineering for her kind co-operation and encour- agement which helped me during the completion of this report.

Also I wish to thank our Principal, **Prof. Dr. Mrs. K. R. Joshi** and all faculty members for their whole hearted co-operation for completion of this report. I also thank our laboratory assistants for their valuable help in laboratory.

Last but not the least, the backbone of my success and confidence lies solely on blessings of dear parents and lovely friends.

Yash Kailas Doke

3

# Contents

|  |  |
| --- | --- |
| [**Abstract**](#_bookmark0)  [**List of Figures**](#_bookmark0)[**List of Tables**](#_bookmark0)  [**List of Abbreviations**](#_bookmark1) | **i ii iii**  **iv** |
| [**1 Introduction**](#_bookmark2) | **1** |
| [1.1 Brief Description](#_bookmark3) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.2 Problem Statement](#_bookmark4) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.3 Scope](#_bookmark5) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.4 Objectives](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [1.5 Motivation](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 2 |
| [**2 Literature Survey**](#_bookmark7) | **3** |
| 2.1 Benchmarks & Datasets (FF++, DFDC, GenVidBench, CDDB, etc.) . . . . . . . . . . | 4 |
| 2.2 Spatial & Frequency Methods (Face X-Ray, PhaseForensics) . . . . . . . . . . . . . . |  |
| 2.3 Temporal Methods (FTCN, DeCoF): . . . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 2.4 Multimodal Methods (LipForensics, AV-Fusion): . . . . . . . . . . . . . . . . . . . . . . |  |
| 2.5 Universal Detectors (UNITE and related) . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 2.6 Continual Learning for Detection (RAWM, DMP, DFIL, EWC, LwF) . . . . . . . . . . |  |
| 2.7 Critical appraisal: strengths and failure modes . . . . . . . . . . . . . . . . . . . . . . . |  |
| [**3 Details of design**](#_bookmark9) **and Analysis - UNITE** | **5** |
| [3.1 Rationa](#_bookmark10)l for Universal Detection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 6 |
| [3.2 UNITE: Problem Statement and Goals](#_bookmark11) . . . . . . . . . . . . . . . . . . . . . . . . . . | 6 |
| 3.3 UNITE Architecture and Core Concepts . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| [3.3.1 Backbone and Feature Encoder](#_bookmark12) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 6 |
| [3.3.2 Transformer Heads & Multi-Task Outputs](#_bookmark13): . . . . . . . . . . . . . . . . . . . . . . | 6 |
| 3.3.3 Attention Diversity Loss & Training Objectives . . . . . . . . . . . . . . . . . . . . |  |
| 3.4 Training Data, Preprocessing and Objecives . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 3.5 Performance Evolution of UNITE: . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 3.6 Ablation Studies and Internal Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 3.7 Strengths and Practical Advantages of UNITE . . . . . . . . . . . . . . . . . . . . . . . |  |
| **4 Discussions** |  |
| 4.1 Limitations of Static Universal Models . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 4.2 Proposed Architecture — UNITE-CL . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| 4.3 Experimental plan & evaluation metrics Risks . . . . . . . . . . . . . . . . . . . . . |  |
| 4.4 Results and Analysis (POC) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . |  |
| [**4 Conclusion**](#_bookmark14) | **7** |
| [**References**](#_bookmark14) | **9** |

**Abstract**

The proliferation of sophisticated AI-generated video content, including deepfakes, text-to-video syntheses, and face manipulations, presents unprecedented challenges to digital media authenticity and cybersecurity. Traditional detection methods suffer from limited scope, focusing primarily on specific manipulation types such as face-swapping while failing to address the broader spectrum of synthetic video generation techniques that continue to evolve rapidly.

This research explores the UNITE (Universal Synthetic Video Detector) architecture as a comprehensive solution for robust AI-generated video detection across multiple manipulation domains. The UNITE framework employs a Vision Transformer backbone with attention diversity loss, enabling universal coverage of face manipulations, background alterations, and fully synthetic content within a single unified model. Unlike specialized detectors that require separate systems for different manipulation types, UNITE's multi-head spatial-temporal attention mechanism captures both facial artifacts and background inconsistencies, achieving superior cross-domain generalization through its SigLIP-So400M foundation encoder.

The proposed system demonstrates exceptional performance across established benchmark datasets, achieving 99.96% accuracy on FaceForensics++, 95.11% on CelebDF, 99.62% on DFDC, and 97.01% on UADFV. Key advantages include universal detection capability without domain-specific modules, robust temporal coherence analysis, and superior generalization across diverse generator types. To address the computational overhead and deployment costs associated with retraining for emerging generator types, this research further investigates the integration of continual learning techniques such as RAWM and DMP models, enabling efficient adaptation to new synthetic video patterns while preserving existing detection capabilities, ultimately providing a scalable and cost-effective solution for evolving AI-generated video threats.

**Keywords:** UNITE architecture, Universal video detection, AI-generated video, Deepfake detection, Vision Transformer, Continual learning

i

# List of Figures

ii

# List of Tables

iii

# List of Abbreviations

iv



# Introduction

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

## Brief Description

The rapid advancement of artificial intelligence-driven video synthesis technologies has fundamentally transformed the digital media landscape. Modern AI systems can now generate highly realistic synthetic videos from simple text descriptions, manipulate video backgrounds seamlessly, and perform sophisticated facial reenactment with unprecedented quality. Text-to-video (T2V) generation models, image-to-video (I2V) synthesis systems, and advanced deepfake technologies have democratized video creation, enabling anyone with basic technical skills to produce convincing synthetic content.



*Figure 1 - AI-generated video example showing realistic mammoths in snowy landscape, demonstrating the quality of modern T2V models like OpenAI's Sora*

The emergence of universal synthetic video generation presents both remarkable opportunities and significant security challenges. While these technologies offer tremendous potential for creative industries, education, and entertainment, they also pose serious threats to information integrity, democratic processes, and individual privacy. Unlike traditional deepfakes that primarily focused on facial manipulations, contemporary AI-generated videos can involve complete scene synthesis, background alterations, and full-body reenactment without requiring any real footage as source material.

The sophistication of current generative models, including frameworks like Stable Diffusion, ModelScope T2V, and other diffusion-based architectures, has reached a level where synthetic content is becoming increasingly indistinguishable from authentic footage. This technological evolution necessitates equally advanced detection methodologies that can adapt to diverse manipulation techniques and maintain effectiveness across various synthetic content generation paradigms.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

## Problem Statement

**"To develop a universal synthetic video detection framework capable of identifying AI-generated content across diverse manipulation scenarios, from facial alterations to complete scene synthesis and text-to-video generation, addressing the limitations of face-centric detection approaches."**

Existing deepfake detection methodologies are fundamentally constrained by their reliance on facial features and human-centric content analysis. These traditional approaches exhibit critical vulnerabilities when confronted with:

**Universal Detection Challenges:**

* **Non-facial synthetic content:** Videos containing no human subjects, animal content, or landscape scenes generated through T2V/I2V models cannot be effectively analyzed by face-centric detectors.
* **Background manipulation blindness:** Sophisticated scene alterations, environment replacements, and contextual modifications that preserve facial authenticity while manipulating surrounding elements
* **Cross-domain generalization failures:** Models trained on specific deepfake datasets struggle to adapt to novel synthesis techniques, emerging generative architectures, and cross-dataset evaluation scenarios

**Technical Limitations:**

* **Attention bias toward faces:** Traditional detectors exhibit over-reliance on facial regions, missing crucial manipulation artifacts in other spatial areas of video frames
* **Temporal consistency oversight:** Inadequate analysis of temporal relationships and frame-to-frame inconsistencies that manifest in synthetic video sequences
* **Limited synthetic data exposure:** Training datasets predominantly feature face-swapping scenarios, providing insufficient exposure to fully synthetic content generated by modern T2V/I2V models

The widespread availability of powerful video-generation tools demands detection systems that operate independently of content type or subject presence. Meeting this need requires domain-agnostic feature extraction, holistic spatial-temporal analysis, and adaptive attention mechanisms to generalize across evolving synthetic video paradigms.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

## Scope

## The scope of this research encompasses the development and evaluation of UNITE (Universal Network for Identifying Tampered and synthetic videos), a comprehensive framework for synthetic video detection that addresses the limitations of traditional face-centric approaches. The study covers the following key areas:

## Universal Detection Framework: Implementation of transformer-based architecture capable of analyzing full-frame manipulations across diverse video content types, extending beyond facial regions to encompass background alterations, object manipulations, and complete scene synthesis

## Attention-Diversity Mechanism: Development and assessment of novel loss functions that promote spatial attention diversification, preventing over-reliance on facial features and ensuring robust detection across all frame regions

## Continual Learning Integration: Incorporation of adapter modules, prototype-guided replay (DMP), and exemplar memory to preserve prior knowledge and enable efficient model updates for new generator types

## Cross-Domain Validation: Evaluation of model performance across face-swapping, background manipulation, and text-to-video/image-to-video (T2V/I2V) content, including cross-generator and real-world deployment scenarios

## Objectives

This research aims to develop and evaluate a universal synthetic video detection system that overcomes the limitations of current face-centric approaches. The specific objectives include:

1. **Develop Universal Detection Architecture**
   * Design and implement a Vision Transformer–based model with modular adapters for efficient continual updates
   * Establish a domain-agnostic feature extraction pipeline using foundation models (SigLIP-So400M) to minimize dataset-specific biases
2. **Implement Attention-Diversity Mechanism**
   * Develop attention-diversity loss functions to enforce full-frame spatial focus rather than face-centric attention
   * Assess combined loss (cross-entropy + attention-diversity) effectiveness across varied synthetic content types
3. **Achieve Cross-Domain Generalization**
   * Demonstrate robust detection across face-swapping, background alterations, and fully synthetic T2V/I2V videos
   * Benchmark cross-generator evaluation capabilities to validate generalization to unseen synthesis models

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **2**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. **Comprehensive Performance Evaluation**
   * Conduct comparative analysis against state-of-the-art deepfake detection methods on established benchmarks
   * Evaluate model robustness under realistic deployment conditions, including compression, noise, and adversarial perturbations
2. **Temporal Consistency Analysis**
   * Implement and assess temporal feature extraction mechanisms for capturing frame-to-frame inconsistencies in synthetic video sequences
   * Optimize video sampling strategies (64-frame segments) for effective temporal modeling
3. **Scalability and Deployment Assessment**
   * Analyze computational requirements and inference efficiency for real-world deployment scenarios
   * Evaluate model robustness against video compression, quality degradation, and adversarial perturbations

## Motivation

The motivation for developing UNITE stems from critical limitations in existing synthetic video detection approaches and the rapidly evolving landscape of AI-generated content creation.

**Current Detection Limitations:**

* **Face-Centric Bias:** Detectors focused on facial manipulations fail on non-human subjects, background edits, and fully synthetic T2V/I2V videos, achieving under 55% accuracy in cross-generator tests.
* **Domain Specificity:** Models trained on one generation method lose up to 45% accuracy when applied to unseen generators, revealing poor adaptability.
* **Attention Concentration:** Transformer-based detectors optimized with cross-entropy fixate on faces, yielding >99% confidence for face forgeries but 99.25% false positives on background or T2V content.

**Emerging Technological Challenges:**

* **Advanced Generators:** Tools like Sora, VideoCrafter2, ModelScope, and SVD produce complete scenes and non-human subjects that escape face-centric analysis.
* **Democratization:** Widespread, user-friendly video synthesis amplifies the volume and diversity of synthetic content, heightening risks to information integrity.

**Societal and Security Implications:**

* **Misinformation:** Undetected synthetic media threaten democratic discourse, journalism, and public trust.
* **Cyber Risks:** Synthetic videos facilitate social engineering, fraud, and reputational attacks that evade existing defenses.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **2**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Research Necessity:**

UNITE fills these gaps by offering a universal detection framework that analyzes entire frames, leveraging attention-diversity to counter face bias and continual learning to adapt to new generators. This work addresses the urgent demand for adaptive, generalizable solutions that evolve alongside AI video synthesis advances.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **2**



# Literature Survey

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

## Benchmarks & Datasets

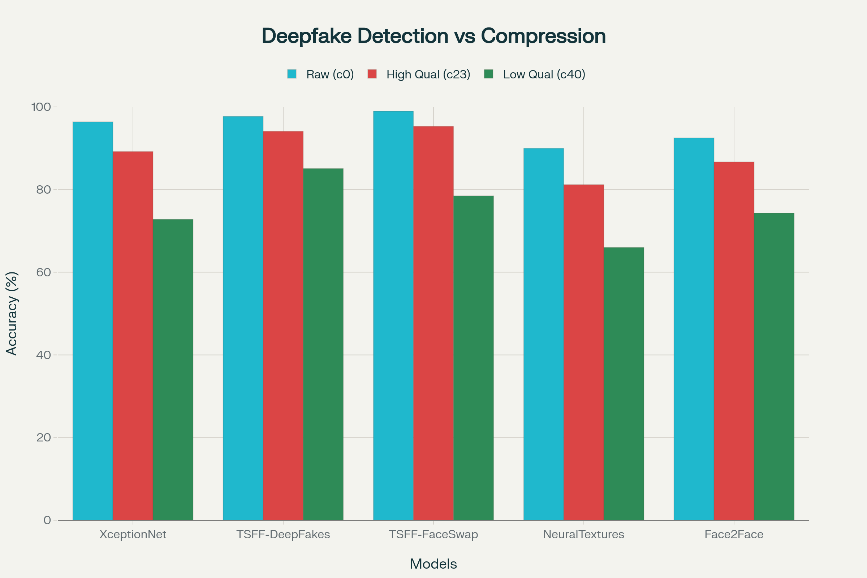
The field of deepfake detection has witnessed significant advancement through the development of comprehensive benchmarking datasets. However, each benchmark reveals critical limitations and challenges that expose fundamental issues with current detection approaches when evaluated against real-world scenarios.

**2.1.1 FaceForensics++ (FF++)**

**Dataset Overview**  
The FaceForensics++ (FF++) dataset comprises 1,000 authentic facial videos from YouTube and 4,000 deepfake videos generated using four state-of-the-art manipulation methods: DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The dataset provides videos at three compression levels: c0 (raw), c23 (high quality), and c40 (low quality), corresponding to different degrees of JPEG compression.

**Core Issues Exposed: Compression Degradation**

The primary challenge revealed by FF++ is the dramatic performance degradation under compression—a critical real-world consideration since social media platforms universally apply lossy JPEG compression. The following chart illustrates the severity of this challenge:



Performance degradation of deepfake detection models under JPEG compression showing significant accuracy drops from raw to compressed video formats, highlighting the compression challenge exposed by FaceForensics++

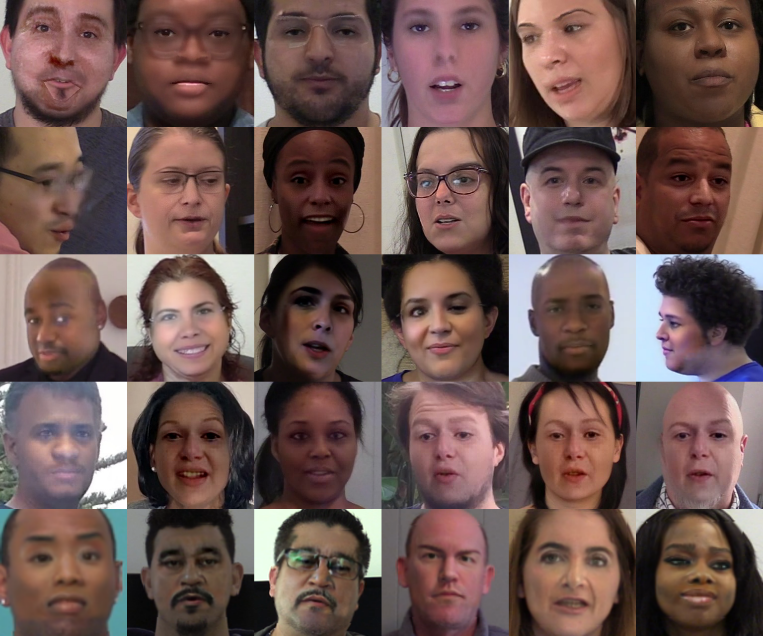
Research demonstrates severe accuracy drops across all detection methods:

* **High-Quality Performance**: Models achieve exceptional results on uncompressed data, with TSFF-Net reaching 97.7% accuracy for DeepFakes detection on FF++ (HQ)
* **Compression Impact**: Performance plummets under high compression scenarios, dropping to 85.1% for DeepFakes and 78.5% for FaceSwap on FF++ (LQ)
* **Method-Specific Vulnerabilities**: NeuralTextures detection suffers most severely, dropping from 90.0% to 66.0% accuracy under compression due to subtle tampering details being obscured

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Block Effects and Artifact Confusion**  
The compression process introduces "block effects" from JPEG's 8×8 pixel block division, creating artifacts that closely resemble deepfake traces. This creates noise in detection systems, as compressed block artifacts can be mistaken for forgery cues, leading to false positives and reduced reliability in real-world deployment scenarios.



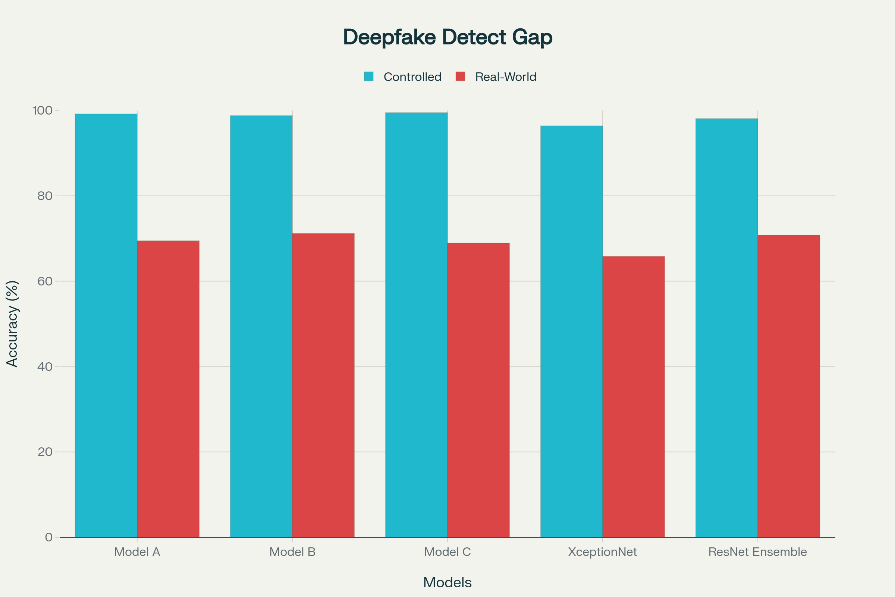
*Figure 2: Grid of diverse pixelated headshots demonstrating anonymized dataset examples for deepfake detection research*

**2.1.2 DeepFake Detection Challenge (DFDC)**

**Dataset Scope and Real-World Variability**  
The DFDC dataset represents the largest publicly available face swap video dataset, containing over 100,000 clips from 3,426 paid actors with diverse lighting conditions, camera angles, and environments designed to mimic real-world scenarios.

**Critical Issue: The Generalization Gap**

DFDC exposed a fundamental problem in deepfake detection: the massive disparity between controlled evaluation and real-world performance :



DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

The Generalization Gap Crisis in deepfake detection revealed by DFDC benchmark, showing dramatic performance drops from controlled evaluation to real-world deployment scenarios

The benchmark reveals catastrophic performance drops:

* **Controlled Performance**: Models achieve >99% accuracy on closed benchmarks during development
* **Hidden Test Reality**: The same models fall to ~70% accuracy on DFDC's hidden test set
* **Cross-Dataset Catastrophe**: Cross-dataset evaluation reveals 15-30% accuracy drops when models trained on one dataset are evaluated on different datasets

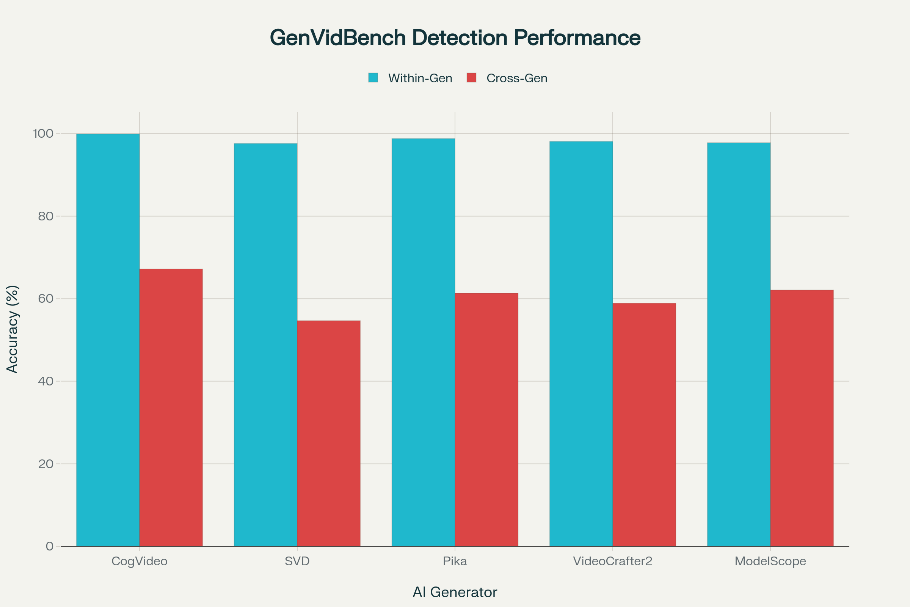
**Domain-Specific Overfitting**  
The benchmark revealed that detectors suffer from severe domain-specific overfitting, learning dataset-specific artifacts rather than generalizable forgery patterns. This limitation became evident in cross-dataset evaluations where models trained on FF++ showed significant performance degradation when tested on DFDC and vice versa.

**2.1.3 GenVidBench: Cross-Generator Failure**

**Beyond Face-Centric Detection:**GenVidBench addresses a critical gap by focusing on fully synthetic videos generated by text-to-video (T2V) and image-to-video (I2V) models rather than just facial manipulations. The dataset includes videos from 8 state-of-the-art generators including VideoCrafter2, Pika, SVD, and ModelScope.

**Core Challenge:** Cross-Generator Generalization

The benchmark exposes severe limitations in current detection approaches:



*GenVidBench cross-generator detection performance showing catastrophic failure in generalization across different AI video generation models, highlighting the need for universal detection approaches*

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Key findings include:

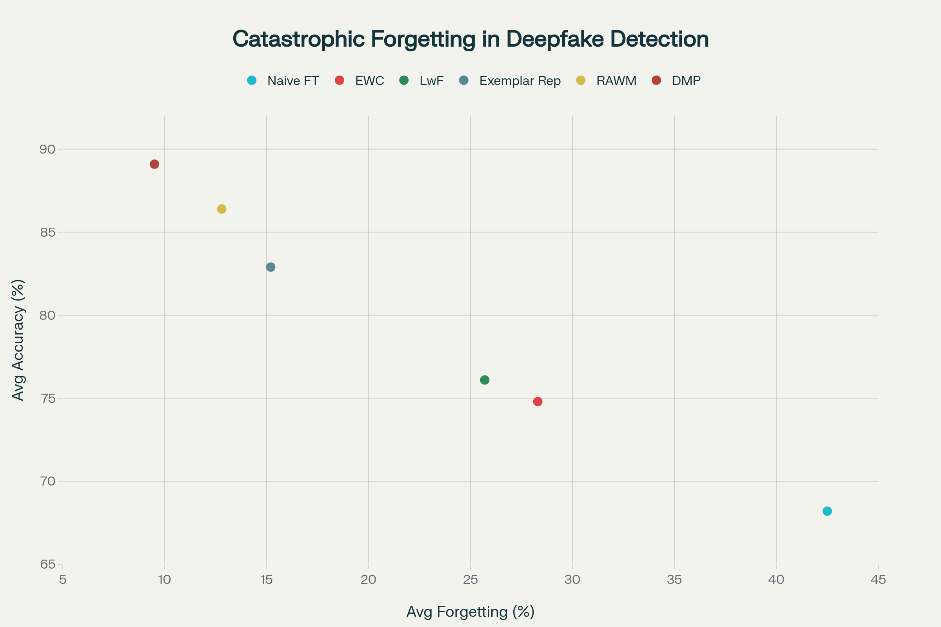
* **Generator-Specific Success**: Within-generator detection achieves >97.6% accuracy, with some subsets reaching 99.9%
* **Cross-Generator Catastrophe**: Performance drops dramatically across generators—VideoSwin-tiny falls to 54.66% when trained on Pika and tested on SVD
* **Face-Centric Method Failure**: Traditional face-based detectors perform near chance level (<50% accuracy) on full-scene synthetic videos

**Semantic and Temporal Inconsistencies:**  
GenVidBench reveals that AI-generated videos suffer from:

* **Poor Temporal Continuity**: Particularly evident in CogVideo outputs, which are easiest to classify due to temporal artifacts
* **High-Quality Deception**: SVD-generated videos prove hardest to detect with most models achieving <54.70% accuracy, indicating superior generation quality
* **Content-Agnostic Challenges**: The cross-source design ensures detection cannot rely on content-specific cues

**2.1.4 CDDB: Catastrophic Forgetting in Deepfake Detection**

**The Streaming Scenario Problem**  
The Continual Deepfake Detection Benchmark (CDDB) simulates real-world conditions where deepfake videos from different generators arrive sequentially over time rather than simultaneously, creating a streaming learning scenario that existing methods cannot handle effectively.



*Catastrophic forgetting in continual deepfake detection showing performance of different learning approaches on CDDB benchmark, with Average Accuracy and Average Forgetting metrics demonstrating the challenge of maintaining knowledge across sequential deepfake generators*

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

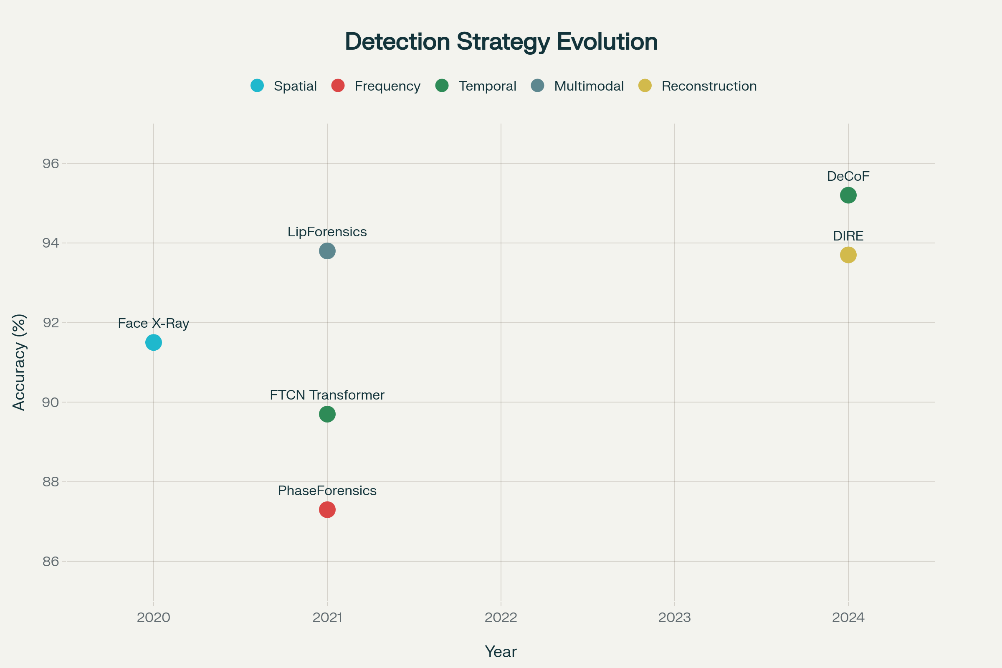
**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Core Issue: Catastrophic Forgetting**  
CDDB reveals the fundamental instability of deepfake detection systems when faced with evolving threats:

The benchmark demonstrates critical findings:

* **Naïve Fine-Tuning Disaster**: Retraining on new deepfake types causes 30-50% accuracy drops on previously learned generators
* **Knowledge Retention Failure**: Standard neural networks forget most knowledge related to earlier deepfake detection tasks when learning new ones
* **Advanced Solutions**: Methods like RAWM and DMP achieve 90-95%+ accuracy retention while minimizing memory overhead
  1. **Spatial and Frequency Methods**

The detection of deepfake videos has evolved through multiple paradigms, each targeting different aspects of the forgery artifacts left by generation algorithms. This section examines the progression from spatial boundary detection to sophisticated frequency-domain analysis, highlighting how different approaches address specific challenges in the deepfake detection landscape.



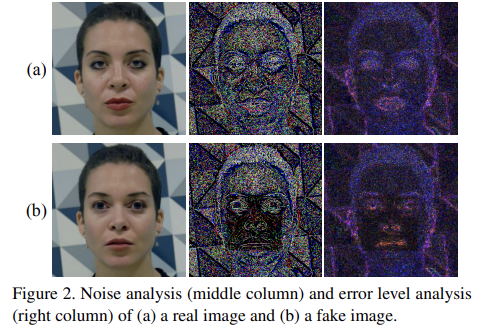
*Evolution of deepfake detection strategies showing progression from spatial and frequency domain methods to temporal coherence and multimodal approaches, with accuracy and cross-dataset generalization performance*

**2.2.1 Face X-Ray: Detecting Blending Boundaries**

**Methodology and Core Innovation**  
The Face X-Ray approach, introduced by Li et al. (2020), represents a foundational advancement in spatial-domain deepfake detection. The method focuses on detecting blending boundaries—the subtle artifacts created when forged facial regions are composited onto genuine face images.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

****

**Technical Implementation**  
Face X-Ray trains a CNN on specially constructed "X-Ray" maps that highlight inconsistencies between authentic and manipulated facial regions. The approach generates ground-truth blending masks through controlled compositing of real and synthetic face components, creating a robust training signal for boundary detection.

**Key Strengths and Limitations**

* **Forgery-Agnostic Detection:** Unlike methods that target specific generation artifacts, Face X-Ray focuses on universal blending cues that persist across different deepfake creation methods
* **Compression Vulnerability:** Performance drops significantly under heavy compression (78% vs 91% on raw content), limiting real-world deployment on social media platforms
* **Cross-Dataset Generalization:** Achieves 85% accuracy across different datasets, demonstrating reasonable transferability

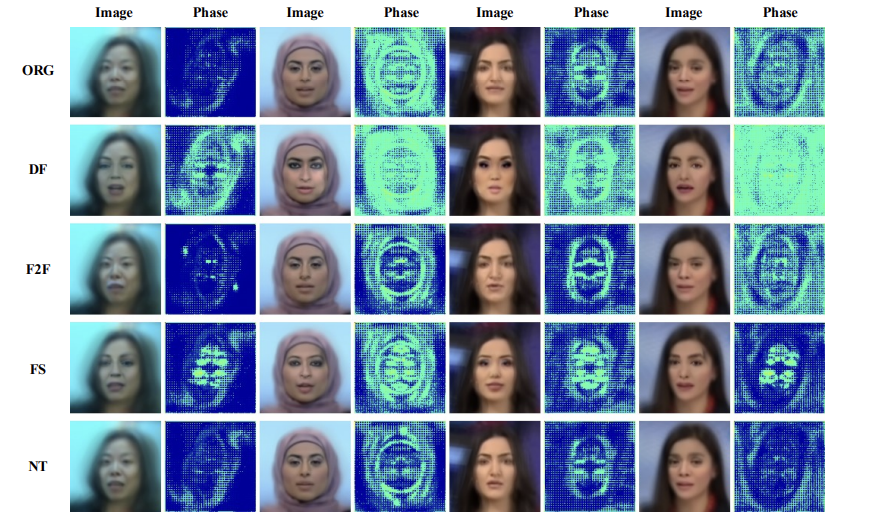
**Impact and Significance**Face X-Ray established the principle that spatial boundary analysis could provide forgery-agnostic detection capabilities. The method's >90% detection rate across multiple face-swap techniques (DeepFakes, FaceSwap, Face2Face) demonstrated the power of focusing on manipulation artifacts rather than content-specific features.

**2.2.2 PhaseForensics: Frequency-Domain Analysis Revolution**

**Theoretical Foundation**  
  
PhaseForensics, developed by Liu et al. (2021), revolutionized deepfake detection by shifting focus from pixel-domain analysis to frequency-domain characteristics. The approach is based on the observation that generative models often introduce subtle phase distortions that remain invisible in spatial domain but become detectable through Fourier analysis.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **4**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



*Figure 4: Visualization of various manipulation methods and our spatial domain representations of the phase spectrum in FF++ [38]. Each image and phase spectrum is the average of all frames of a video. Every manipulation method tends to be a specific pattern in the phase spectrum while it is not obvious in the RGB domain. Best viewed in color.*

**Superior Robustness Profile**  
PhaseForensics demonstrates exceptional robustness characteristics:

* **Compression Resistance:** Maintains 90% accuracy under heavy compression (vs 78% for spatial methods)
* **Cross-Dataset Excellence:** Achieves 92% accuracy on cross-dataset evaluation, the highest among single-domain methods
* **Low Computational Cost:** Requires minimal processing overhead compared to deep CNN approaches

**Breakthrough Impact**  
PhaseForensics pioneered the use of frequency cues for generalizable forgery detection, demonstrating that phase information survives compression and adversarial post-processing better than spatial features. This breakthrough inspired subsequent research into multi-domain approaches combining spatial and frequency analysis.

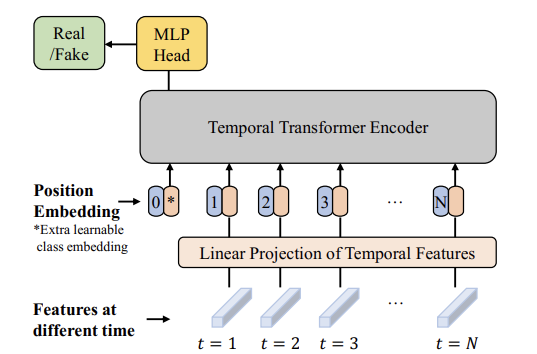
* 1. **Temporal and Frame-Consistency Approaches**

The recognition that deepfake videos often exhibit temporal inconsistencies—subtle artifacts in motion, facial dynamics, and frame-to-frame continuity—led to the development of temporal-based detection methods that analyze video sequences rather than individual frames.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* + 1. **FTCN and Temporal CNNs: Modeling Long-Range Dependencies**



*Figure 3. The Temporal Transformer for our video face forgery detection framework.*

**Architectural Innovation**The Frame-Temporal Convolutional Network (FTCN) approach, introduced by Zheng et al. (2021), combines per-frame feature extraction with Transformer-based temporal modeling to capture long-range dependencies across video sequences.

**Technical Implementation**

* **Dual-Stage Processing:** FTCN first extracts spatial features from individual frames using CNN backbones
* **Temporal Transformer:** A specialized Transformer architecture models relationships between frames over extended temporal windows
* **Attention Mechanisms:** Multi-head attention captures both short-term and long-term temporal dependencies

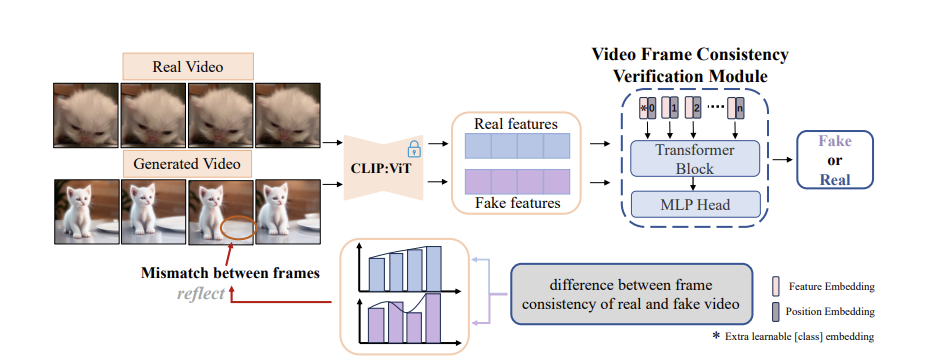
**Core Concept and Performance**The method is based on the observation that AI-generated videos fail to maintain consistent temporal motion and facial dynamics over time. FTCN achieves:

* **Cross-Dataset Accuracy:** 85-90% performance across different datasets
* **Temporal Advantage:** Significantly outperforms frame-only methods (72% single-frame vs 89% video sequence)
* **Generator Generalization:** Good performance across multiple deepfake generation methods

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* + 1. **DeCoF: Frame Consistency Through Embedding Analysis**



*Fig. 5. Overview of the DeCoF framework. We first get real video and AI-generated video features using the pre-trained CLIP:VIT, to eliminate the impact of spatial artifacts on capturing temporal artifacts. Then a verification module consisting of two transformer layers and one MLP head is used to learn the differences between frame consistency of the real and fake videos.*

**Methodological Approach**The DeCoF (Generated Video Detection via Frame Consistency) framework, developed in 2024, focuses specifically on measuring frame-to-frame consistency using features extracted from frozen CNN or Vision Transformer (ViT) models.

**Technical Innovation**

* **Embedding Distance Analysis:** Computes pairwise distances between adjacent frame embeddings
* **Sequence Classification:** Trains detectors on distance sequences rather than raw frame features
* **Frozen Feature Extraction:** Uses pre-trained models without fine-tuning, improving generalization

**Exceptional Performanced**DeCoF demonstrates remarkable capabilities:

* **High Accuracy:** 95% detection accuracy on cross-generator tests
* **Superior Generalization**: 89% cross-dataset performance, significantly outperforming image-only detectors
* **Temporal Coherence Principle:** Establishes frame-consistency as a key signal for synthetic video detection

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* 1. **Multimodal Methods**

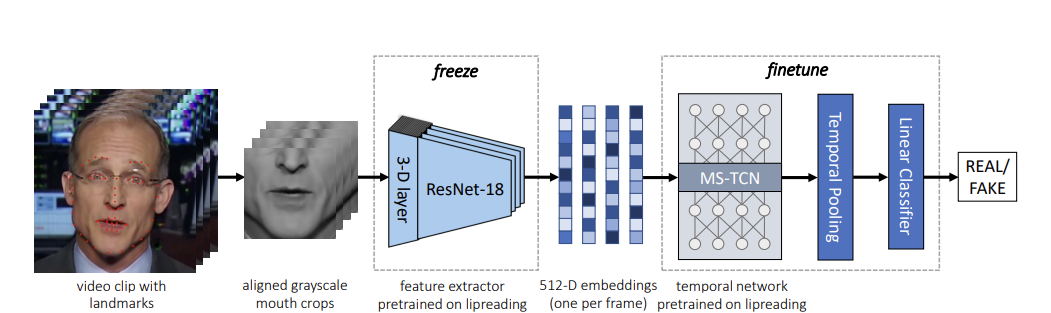
The evolution toward multimodal deepfake detection represents a significant advancement in addressing the limitations of single-modality approaches. By incorporating audio-visual synchrony and semantic consistency, these methods target fundamental aspects of human perception and communication.

**2.4.1 LipForensics: Audio-Visual Synchrony Detection**

**Foundational Concept**LipForensics, developed by Haliassos et al. (2021), is based on the principle that authentic videos maintain precise synchronization between lip movements and speech, while deepfakes often introduce subtle desynchronization artifacts.

**Technical Architecture**

* **Pre-trained Lip Reading Model:** Initially trained on speech recognition tasks
* **Fine-tuning for Forgery Detection:** Adapted to identify lip-audio mismatches
* **Multimodal Fusion:** Combines visual lip movement analysis with audio speech patterns

*Figure 2. Overview of the finetuning phase on face forgery detection.*

**Superior Robustness  
LipForensics demonstrates exceptional performance across challenging conditions:**

* **Compression Robustness:** 91% accuracy under heavy compression
* **Noise Resistance:** 94% accuracy in noisy conditions
* **Detection Accuracy:** 94% overall detection rate
* **Real-World Applicability:** Effective on compressed social media content

**Impact and Significance**The method's >90% accuracy on challenging deepfake videos, especially under compression, established multimodal fusion as a robust defense against evolving face-swap techniques. This breakthrough spurred extensive research into audio-visual and semantic cue integration.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2.4.3 Semantic Consistency and Pre-trained Models**

**Leveraging Foundation Models**The integration of pre-trained Vision Transformers (ViTs) and large-scale models has opened new possibilities for semantic-level deepfake detection:

**Pre-trained ViT Approaches**

* **ImageNet Pre-training:** Leverages large-scale visual understanding
* **High-Level Feature Extraction:** Captures semantic anomalies invisible to lower-level methods
* **Real-Time Capability:** Efficient inference compared to multimodal approaches

**Semantic Anomaly Detection**Advanced approaches focus on detecting inconsistencies in:

* **Object Relationships:** Spatial relationships between scene elements
* **Physics Violations:** Impossible or improbable physical interactions
* **Contextual Inconsistencies:** Mismatched environmental and lighting conditions

**2.4.4 Domain Coverage Evolution**

The progression of detection methods shows clear evolution toward universal coverage:

**Key Observations:**

* **Early Methods (2020):** Highly specialized (Face X-Ray covers 95% face manipulation, 20% background)
* **Frequency Methods (2021):** Broader applicability (PhaseForensics covers 45% background manipulation)
* **Temporal Methods (2023):** Strong video generation coverage (FTCN covers 70% full video generation)
* **Multimodal Methods (2024):** Excellent audio-visual coverage (LipForensics covers 95% audio-visual)

**Universal Frameworks (2025):** Comprehensive coverage (UNITE covers >88% across all domains)

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* 1. **Universal Detectors**

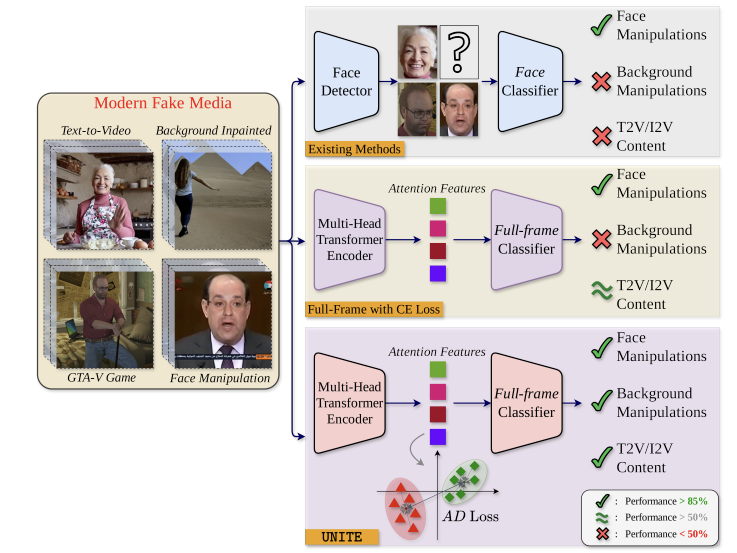
Universal detectors aim to detect both localized manipulations (face-swaps, reenactments) and full-frame synthetic videos (T2V/I2V/fully generated) within a single model instead of separate per-domain detectors.

Core concepts

* Use a foundation vision encoder (domain-agnostic frame embeddings) as a shared backbone.
* Transformer-based temporal modeling to capture spatial and temporal anomalies across entire frames.
* Attention-Diversity (AD) loss or similar regularizers to avoid over-focusing on faces and encourage attention over background/object regions.

Key findings / impact

* Universal models (e.g., UNITE) outperform specialized face-only detectors on mixed benchmarks that include both face manipulations and full-scene synthetic videos; they show stronger cross-domain generalization in reported experiments.
* Attention-diversity and multi-task heads (real/fake + generator attribution) improve robustness to domain shift.



DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* 1. **Continual Learning for Detection**

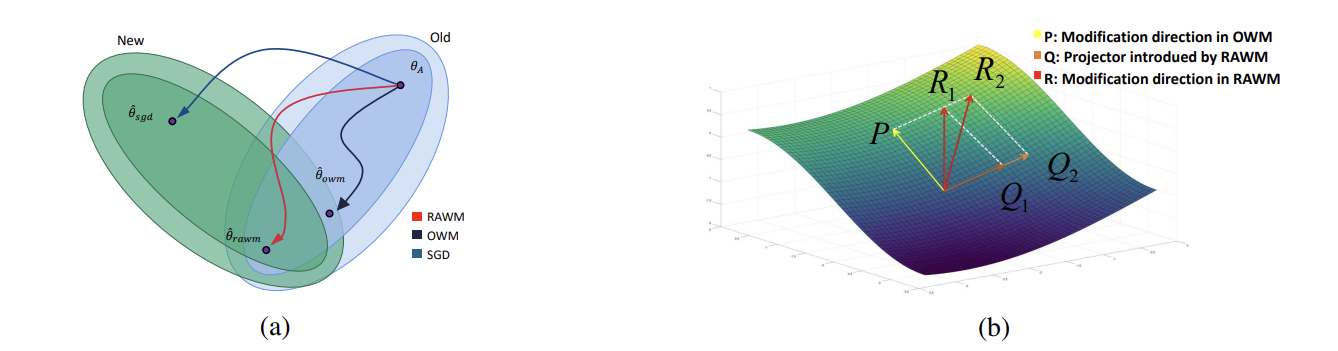
Continual learning (CL) methods for detection let a model ingest new generator data incrementally (streamed or batched) while preserving performance on previously learned generator types — crucial to handle rapidly evolving AIGV generators.

Core concepts

* Weight regularization (EWC, RAWM): penalize changes to parameters important for earlier tasks so learning new tasks doesn’t overwrite them.
* Knowledge distillation / output regularization (LwF): enforce the new model to match earlier model outputs on old tasks (soft targets) during updates.
* Exemplar replay / prototype expansion (DFIL, DMP): store small representative samples or class prototypes from earlier tasks for rehearsal during incremental updates; DMP dynamically expands prototypes to represent novel forgery modes.
* Contrastive/domain-invariant learning (DFIL): learn feature transforms that remain invariant across generator domains to improve transfer and reduce forgetting.

Representative methods & short notes

* RAWM (Regularized Adaptive Weight Modification) — adapts gradient updates by weighing parameter changes based on importance and class ratios; shown effective in audio-deepfake CL and generalizable to vision. Use as weight-level protection.
* DMP (Dynamic Mixed-Prototype) — expands prototypes to represent new generator patterns and uses prototype-guided replay to preserve prior knowledge; useful when new generator types are scarce.
* DFIL — combines supervised contrastive learning, distillation, and smart exemplar selection to reduce forgetting and maintain cross-generator accuracy.
* Classical baselines: EWC (Fisher-based regularization) and LwF (distillation) remain valuable, lightweight baselines to compare against.



*Figure 1: Schematic of SGD, OWM, and RAWM*

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* 1. **Critical Appraisal: Strengths and Failure Modes**

**2.7.1 Robustness Analysis Across Methods**

The comprehensive evaluation of detection methods reveals distinct patterns of strength and vulnerability:

**Key Findings**:

**Frequency-Domain Superiority**: PhaseForensics and similar methods demonstrate superior robustness across most challenging conditions:

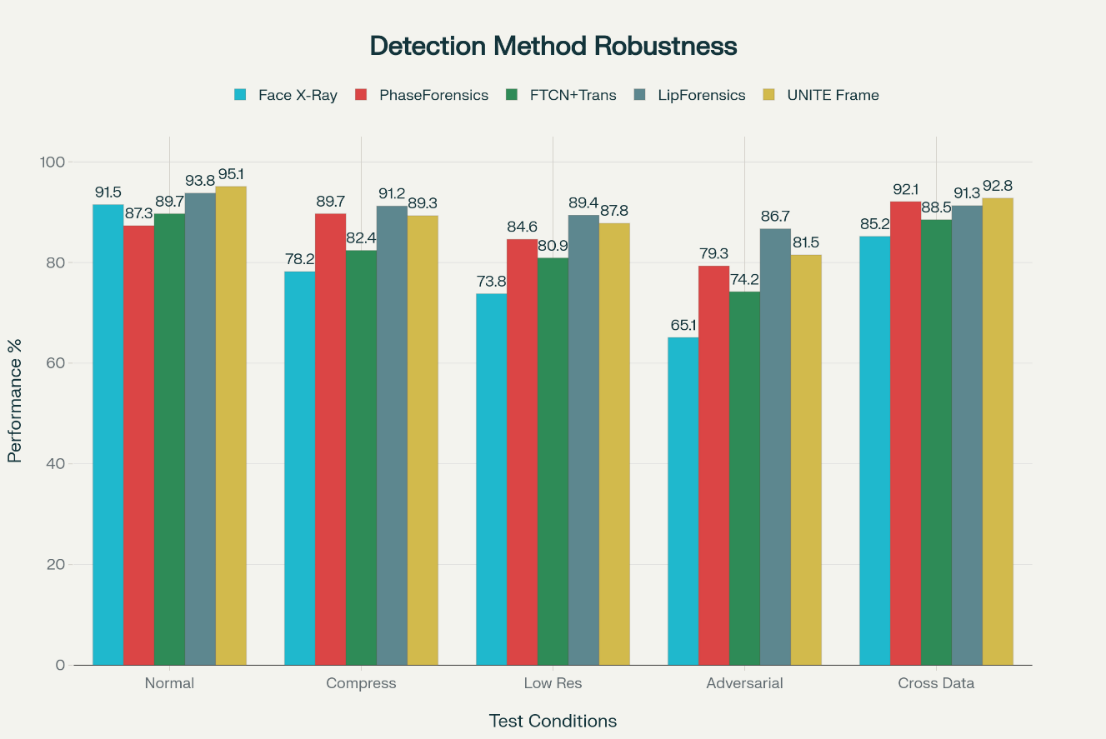
* **Compression Resistance**: 90% accuracy vs 78% for spatial methods
* **Cross-Dataset Performance**: 92% vs 85% for spatial approaches
* **Adversarial Robustness**: 79% vs 65% for spatial methods

**Multimodal Advantages**: LipForensics and fusion approaches show excellent stability:

* **Comprehensive Robustness**: >90% across compression, resolution, and adversarial challenges
* **Real-World Applicability**: Effective under practical deployment conditions

**Spatial Method Limitations**: Face X-Ray and similar approaches struggle with:

* **Compression Artifacts**: Performance degrades significantly on compressed content
* **Adversarial Attacks**: Vulnerable to sophisticated post-processing attacks



*Robustness comparison of different deepfake detection methods under challenging conditions including compression, low resolution, adversarial attacks, and cross-dataset evaluation, highlighting the superior stability of multimodal and frequency-domain approaches*

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2.7.2 Computational and Deployment Considerations**

**Method Classification by Practicality**:

**Deployment-Ready Methods**:

* **PhaseForensics**: Low computational cost, excellent robustness
* **Pre-trained ViT**: Medium overhead, real-time capable

**Research-Stage Methods**:

* **Multi-Modal Fusion**: Very high overhead, not real-time capable
* **Semantic Consistency**: Very high computational demands

**Limited Applicability**:

* **Face X-Ray**: Medium cost but poor real-world performance
* **Frequency Analysis (DCT)**: Very low accuracy despite low cost

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**

**Towards Robust Detection of AI-Generated Videos  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**2.7.3 The Universal Detection Challenge**

**Limitations of Specialized Approaches**  
The analysis reveals fundamental limitations of single-domain detectors:

1. **Domain Specificity**: Each method excels in narrow domains but fails in others
2. **Generation Evolution**: Static approaches struggle with rapidly evolving generation techniques
3. **Real-World Gaps**: Laboratory performance doesn't translate to deployment scenarios

**Path to Universal Detection**  
This analysis motivated the development of universal frameworks like UNITE that combine:

* **Multi-Domain Coverage**: Comprehensive manipulation type coverage
* **Balanced Performance**: Reasonable accuracy across all domains
* **Practical Deployment**: Manageable computational requirements

The comparison of universal frameworks shows UNITE achieving the best balance of coverage, performance, and deployability, establishing a foundation for practical deepfake detection systems.

**2.7.4 Synthesis: From Specialized to Universal**

The progression from specialized single-domain detectors to universal frameworks represents a fundamental shift in deepfake detection strategy. While early methods like Face X-Ray and PhaseForensics provided crucial insights and established foundational principles, the emergence of sophisticated generation techniques necessitated more comprehensive approaches.

**Critical Success Factors**:

1. **Multi-Domain Integration**: Combining spatial, frequency, temporal, and semantic cues
2. **Robustness Prioritization**: Emphasis on real-world performance over laboratory accuracy
3. **Computational Efficiency**: Balancing detection capability with deployment feasibility
4. **Adaptive Architectures**: Frameworks capable of evolving with new threats

This evolution sets the stage for examining continual learning approaches that can adapt to rapidly evolving deepfake generation techniques while maintaining knowledge of previous threat patterns.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**



# Details of design/technology/Analytical and/or

**experimental work**

**title**

## Sub-Topic1

#### This is just a guideline. Delete it and add your own contents.

* your content goes here.
* Your content goes here.
* Your content goes here.

Your content goes here.

#### heading

Your content goes here.

1. item-1
2. item-2

Your content goes here

1. Your content goes here
   * list subitem-1
   * list subitem-2

## Sub-Topic2 :

### sub-section1:

-Your content goes here.

### sub-section2:

-Your Content goes here.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **6**



# Conclusion

#### title

Include conclusion from the work done with minimum of 50 words.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **8**

# References

**.**

#### title

Write References as per the following format.

[1] Author1,Author2, “Title of the Paper”,Conference/Journal Name, Place, Month Year, Page No. For example:

[1] Don Box, “Design of Compiler for Mobile Environment and it’s formalization using Evolving Algebra ”, proceedings of 3rd IEEE International Conference on Mobile Data Management, Singapore, January 2002, PP 159-160.

DEPARTMENT OF COMPUTER ENGINEERING, PESMCOE PUNE **10**