TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING
PULCHOWK CAMPUS

A
PROGRESS REPORT
ON
MULTI-DOMAIN FEATURE FUSION DEEPFAKE DETECTION

SUBMITTED BY:
SANDESH KUIKEL (PUL078BCT076)
SHREYA UPRETY (PUL078BCT086)
SUBASH KANDEL (PUL078BCT091)

SUBMITTED TO:
DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING

October, 2025

Acknowledgments

We would like to express our sincere gratitude to our project supervisor, Prof. Dr. Subarna Shakya, for their invaluable guidance, expertise, and continuous support throughout this project. Their insights and feedback have been instrumental in shaping our research direction and technical approach.

We also extend our appreciation to the Department of Electronics and Computer Engineering at Pulchowk Campus, Institute of Engineering, for providing us with the necessary resources and academic environment to conduct this research.

Our thanks go to the open-source community for developing and maintaining the tools and frameworks that made this project possible, especially the contributors of PyTorch, XceptionNet, and DINO.

Finally, we acknowledge the researchers whose prior work in deepfake detection has laid the foundation for our project and inspired our multi-modal approach.

Abstract

This report documents the progress achieved in developing a Multi-Domain Feature Fusion Deepfake Detection system using advanced computer vision and deep learning techniques. Deepfake technology has emerged as a significant threat to digital media integrity, with sophisticated AI-generated content becoming increasingly difficult to detect using conventional methods. Despite extensive research in this field, current detection systems struggle with cross-domain generalization and robustness against novel manipulation techniques.

The proposed system leverages multi-modal feature fusion approaches, utilizing pretrained Convolutional Neural Networks (CNN), frequency domain analysis, and Vision Transformers (ViT) for comprehensive deepfake detection. The implementation involves collecting datasets from FaceForensics++, preprocessing videos to extract facial regions, and extracting three complementary feature representations: spatial features using XceptionNet architecture, frequency domain features through FFT transformations, and semantic features leveraging DINOv2 pre-trained representations.

Significant progress has been achieved in the project implementation. The complete data processing pipeline has been successfully developed, including automated metadata generation for dataset organization, robust face detection and extraction using MTCNN, and comprehensive multi-domain feature extraction across all three modalities. The extracted features demonstrate high quality and consistency, with 2048-dimensional spatial features, 4-dimensional frequency statistical features, and 768-dimensional semantic embeddings successfully generated for the entire dataset. The modular architecture developed provides flexibility for future enhancements and ensures reproducibility across experimental runs.

The next phases of the project will focus on implementing fusion mechanisms to combine the extracted features, developing MLP-based classification models, and conducting comprehensive evaluation across multiple datasets to assess cross-domain generalization capability. This project has the potential to advance the state-of-the-art in deepfake detection by addressing the generalization gap through multi-domain feature fusion.

Keywords: *Deepfake Detection, Multi-Modal Fusion, XceptionNet, DINOv2, Frequency Analysis, Cross-Dataset Generalization, Computer Vision*

Contents

List of Figures

List of Abbreviations

AFSL	Adversarial Feature Similarity Learning
AI	Artificial Intelligence
AUC	Area Under the Curve
BCE	Binary Cross Entropy
CDDB	Continual Deepfake Detection Benchmark
CNN	Convolutional Neural Network
CVPR	Conference on Computer Vision and Pattern Recognition
DCT	Discrete Cosine Transform
DFDC	Deepfake Detection Challenge
DINOv2	Distillation with No labels version 2
FF++	FaceForensics++
FFT	Fast Fourier Transform
GAN	Generative Adversarial Network
HFE	High-Frequency Enhancement
JPEG	Joint Photographic Experts Group
MLP	Multi-Layer Perceptron
MTCNN	Multi-task Cascaded Convolutional Network
PGD	Projected Gradient Descent
ReLU	Rectified Linear Unit
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
ViT	Vision Transformer

1. Introduction

1.1 Background

The proliferation of synthetic media, particularly deepfakes, has emerged as a significant technological and societal challenge. Deepfakes represent AI-generated or manipulated media depicting individuals saying or doing things they never did, and have evolved rapidly in sophistication since their emergence in 2017. Recent advancements in generative AI, including models like Sora, Midjourney, and advanced GANs, have dramatically improved the quality and accessibility of synthetic media creation tools, making fake content increasingly difficult to distinguish from authentic media.

The potential for misuse of deepfake technology extends across multiple critical domains. First, deepfakes enable sophisticated misinformation campaigns and political manipulation, undermining democratic processes and public discourse. Second, they facilitate the creation of non-consensual intimate imagery, causing severe psychological harm to victims. Third, deepfakes enable financial fraud through voice and video impersonation, compromising organizational security. Fourth, they support identity theft and social engineering attacks at unprecedented scales. Finally, the widespread availability of deepfake technology fundamentally undermines media authenticity and public trust in digital content.

As deepfake technology becomes more sophisticated, traditional detection methods that rely on single-modality analysis have shown decreasing effectiveness against state-of-the-art forgeries. According to recent benchmarks, single-modality detectors experience a significant performance drop when confronted with new deepfake generation methods not seen during training, with accuracy decreasing from over 95% on in-distribution data to below 70% on cross-dataset evaluations [14]. This substantial performance degradation highlights the urgent need for more robust detection approaches.

1.2 Problem Statement

Despite significant research efforts, current deepfake detection systems face several critical challenges that limit their practical deployment. Understanding these challenges is essential for developing more effective detection strategies.

The first major challenge is limited cross-domain generalization. Existing detectors perform well on familiar deepfake types but struggle significantly with novel generation methods not represented in their training data. This “reality gap” represents a fundamental challenge for practical deployment, as new deepfake generation techniques emerge continuously. Detectors trained on specific manipulation methods often fail catastrophically when confronted with novel forgery techniques because they have learned superficial artifacts rather than fundamental properties of manipulation.

The second challenge involves modality-specific vulnerabilities. Single-modality detectors that rely exclusively on RGB pixel analysis can be easily defeated by adversarial attacks targeting their specific detection mechanisms, such as recompression, noise addition, or other perturbations. These vulnerabilities arise because single-modality approaches have inherent blind spots that sophisticated adversaries can exploit.

The third challenge stems from the rapid evolution of generation methods. The accelerating pace of deepfake technology development means detection systems quickly become outdated as new forgery techniques emerge that can bypass existing detection signatures. This creates a continuous arms race between deepfake creators and detectors, with detectors constantly playing catch-up.

The fourth challenge relates to real-world deployment constraints. Many academic deepfake detectors demonstrate high performance in controlled environments but suffer significant degradation when applied to "in-the-wild" media with varying quality, compression levels, and unknown provenance. This gap between laboratory performance and real-world effectiveness severely limits practical deployment.

These challenges collectively highlight the urgent need for a more robust, adaptable approach to deepfake detection that can maintain effectiveness across diverse forgery techniques and real-world scenarios.

1.3 Objectives

This project aims to address the identified challenges through the following primary objectives:

1. Develop an image deepfake detection framework that fuses spatial, frequency, and semantic features for more reliable authenticity analysis. This multi-domain approach is designed to overcome the limitations of single-modality detection by capturing complementary aspects of deepfake artifacts.

2. Evaluate the system's performance with particular emphasis on cross-dataset generalization and robustness to common adversarial manipulations. This evaluation will assess whether the multi-domain fusion approach successfully addresses the generalization challenges identified in current detection systems.

1.4 Scope and Limitations

This project focuses on specific aspects of deepfake detection while acknowledging certain limitations that define the project boundaries.

Regarding media types, our primary focus is on face-centric deepfake detection in static images and video frames. While audio deepfakes represent an important related challenge with significant societal implications, they fall outside the scope of this project due to the fundamental differences in signal processing and feature extraction methodologies required for audio analysis.

For the detection approach, we develop a binary classification system that categorizes media as real or fake with confidence scoring, rather than attempting forgery localization or attribution to specific generation methods. This design choice reflects the primary practical need for authentication rather than forensic analysis, though future extensions could incorporate localization capabilities.

The dataset scope encompasses publicly available deepfake datasets including FaceForensics++, Deepfake Detection Challenge, Celeb-DF, DFDC, and DeepfakeEval-2024 for comprehensive evaluation across generation methods. These datasets provide diverse manipulation techniques and realistic evaluation scenarios that enable thorough assessment of detection performance and generalization capability.

Our multi-modal approach incorporates three complementary feature domains. First, spatial features are extracted via XceptionNet CNN architecture, capturing fine-grained textural patterns and local manipulation artifacts. Second, frequency domain analysis through FFT transformations identifies global statistical anomalies in spectral distributions. Third, semantic features leveraging DINOv2 pre-trained representations verify high-level consistency and logical coherence of facial features.

Performance metrics focus on standard classification measures including accuracy, precision, recall, and F1-score, with particular emphasis on cross-dataset performance to assess generalization capability and robustness to common adversarial manipulations such as compression and noise addition.

The implementation utilizes PyTorch as the deep learning framework, with deployment considerations for both research applications and potential real-world scenarios. The modular architecture design facilitates future extensions and adaptations to emerging requirements.

The project explicitly does not attempt to address all forms of media manipulation, nor does it focus on creating a production-ready deployment system with real-time performance optimization. Rather, it emphasizes research innovation in multi-modal feature fusion and cross-domain generalization for enhanced deepfake detection, establishing a foundation for future practical deployment efforts.

2. Literature Review

2.1 Related Theory

2.1.1 Deepfake Generalization Challenge

The primary theoretical challenge in deepfake detection is generalization. Early detectors often achieved high performance by overfitting to artifacts specific to a particular generation method or dataset, such as the unique upsampling pattern of a certain GAN. This approach is fundamentally brittle because the detector fails when confronted with novel forgery techniques. The failure occurs because the model has learned superficial "tells" rather than the fundamental properties that distinguish manipulated from authentic images.

Our project is founded on the premise that robust detection requires moving beyond a single analytical perspective. We hypothesize that by analyzing potential fakes from multiple complementary viewpoints—spatial, frequency, and semantic—we can build a model that learns a more fundamental and generalizable concept of authenticity. Each perspective captures different aspects of the manipulation signature, and their fusion provides comprehensive coverage of potential forgery artifacts.

2.1.2 Foundations of Feature Extraction

To construct this multi-faceted view, we leverage three distinct theoretical paradigms. Each paradigm is designed to uncover different classes of forgery artifacts that may be missed by the others, creating a complementary detection system.

The spatial domain focuses on local artifact analysis through pixel-level representation of images. Convolutional Neural Networks (CNNs), like XceptionNet, excel in this domain by learning hierarchical feature representations. Their theoretical strength lies in identifying local, low-level artifacts such as pixel grid mismatches, unnatural blending at forgery boundaries, or inconsistencies in skin texture. The depthwise separable convolutions in XceptionNet are particularly effective at capturing these subtle spatial inconsistencies while maintaining computational efficiency.

The frequency domain provides global inconsistency detection through spectral analysis. By applying transformations like the Discrete Cosine Transform (DCT) or Fast Fourier Transform (FFT), we can analyze the global frequency characteristics of images. The theory underlying this approach is that the deepfake generation pipeline often introduces subtle, high-frequency periodicity or disrupts the natural frequency statistics of authentic photographs. These global artifacts become apparent in the frequency spectrum, even when they are invisible in the spatial domain. This phenomenon occurs because generative models often struggle to perfectly replicate the complex frequency distributions present in natural images.

The semantic domain addresses high-level coherence verification through contextual understanding. Our use of DINOv2 is grounded in the theory of self-supervised learning on Vision Transformers (ViTs). The self-attention mechanism in ViTs captures global context, allowing verification of high-level coherence. A deepfake may have flawless pixels and natural frequency distributions but still feature inconsistent lighting, unnatural facial geometry, or implausible physical relationships between facial features. A semantic model is designed to detect such high-level incongruities that violate our implicit understanding of natural facial appearance and behavior.

2.1.3 Theoretical Framework for Multi-Modal Fusion

The three feature extraction paradigms are not redundant but complementary, each with its own blind spots and strengths. Simply concatenating these features would be suboptimal because it fails to account for the varying importance of each modality for different inputs. Some deepfakes may have obvious spatial artifacts but natural frequency characteristics, while others may have the opposite signature.

The theoretical foundation for our fusion strategy rests on attention mechanisms. Attention provides a mathematical framework for the model to learn a dynamic, context-dependent policy for integrating information. It effectively learns to ask, “For this specific image, is the most telling clue in the spatial, frequency, or semantic features?” By calculating attention weights, our fusion module can dynamically amplify the most relevant modality, creating a synergistic effect where the combined model is more robust than the sum of its parts. This adaptive weighting allows the system to leverage the strengths of each modality while mitigating their individual weaknesses.

2.2 Related Works

Deepfake detection research has evolved significantly, with various approaches targeting different aspects of the manipulation detection problem. This review examines CNN-based methods, frequency-domain analysis, transformer models, cross-dataset generalization, robustness enhancement, and benchmark datasets, aligning with our project’s scope of binary classification via spatial (XceptionNet), frequency (FFT), and semantic (DINOv2) features on datasets like FaceForensics++ [18], DFDC [4], and Celeb-DF [14].

2.2.1 CNN-Based Detection Methods

Convolutional Neural Networks form the backbone of most deepfake detection approaches due to their strong performance in learning hierarchical visual representations. The foundational work by Rössler et al. [18] established XceptionNet as a baseline detector, achieving 99.3% accuracy on FaceForensics++ high-quality fakes. This remarkable performance demonstrated the potential of deep learning for deepfake detection but also revealed significant limitations when tested on unseen datasets.

Afchar et al. [1] introduced MesoNet, a network specifically designed to detect mesoscopic details characteristic of deepfake manipulation, achieving approximately 98% accuracy on DeepFake videos. The network's focus on mid-level features provided insights into the types of artifacts most indicative of manipulation.

Recent advances have focused on improving generalization. Huang et al. [7] proposed an Implicit Identity Driven CNN using Xception architecture to improve generalization via identity preservation losses. Yan et al. [24] modified Xception for feature disentanglement, boosting AUC from 0.672 to 0.811 on Celeb-DF, demonstrating that architectural modifications can significantly improve cross-dataset performance. In a similar vein, Wang et al. [23] proposed AltFreezing for 3D spatiotemporal CNNs, enhancing cross-dataset performance through selective layer freezing strategies. More recently, Sun et al. [20] augmented EfficientNet-B4 with DiffusionFake, increasing AUC on unseen data by approximately 10% on Celeb-DF through synthesis-based data augmentation. Lipianina-Honcharenko et al. [15] found Xception achieved around 87.7% accuracy on forgery tasks across various perturbations.

These methods consistently achieve over 90% intra-dataset accuracy, supporting our choice of Xception for extracting spatial features while highlighting the need for complementary approaches to achieve robust cross-dataset generalization.

2.2.2 Frequency-Domain Analysis

Frequency domain analysis has emerged as a powerful complement to spatial analysis, particularly for detecting artifacts in compressed media. Le and Woo [11] used attention-based distillation (ADD) for frequency-attention learning, improving detection of low-quality compressed fakes. Gao et al. [6] proposed High-Frequency Enhancement (HFE), which outperformed existing methods on compressed DFDC/FF++ with an AUC greater than 85%, demonstrating that frequency artifacts remain detectable even after aggressive compression.

Li et al. [13] introduced FreqBlender for data augmentation in the frequency domain, yielding an 88–90% cross-dataset AUC by training models to be invariant to certain frequency manipulations while sensitive to manipulation-specific signatures. Furthermore, Tan et al. [21] developed FreqNet, utilizing FFT and phase spectra analysis, which improved performance by approximately 9.8% across various GAN-generated fakes. These findings strongly align with our decision to incorporate a frequency modality to capture subtle manipulation artifacts that are invisible in the spatial domain.

2.2.3 Transformer and Vision-Language Models

Transformer architectures have recently gained attention in deepfake detection due to their ability to capture long-range dependencies and global context. Dong et al. [5] proposed the Identity Consistency Transformer with ViT backbone, which generalizes well across various degradations by learning identity-preserving representations. Chen et al. [3] adapted CLIP-ViT with guided frequency fusion (GFF), achieving state-of-the-art results on several benchmarks by combining semantic understanding with frequency analysis.

Self-supervised ViTs like DINO, as explored by Nguyen et al. [16], show better generalization when fine-tuned on deepfake detection tasks compared to supervised pre-training. This finding supports our choice of DINOv2-Base for semantic feature extraction, as the self-supervised pre-training captures fundamental visual concepts that transfer well to manipulation detection. Hybrid models such as HCiT [9] combine CNNs and ViTs to leverage both local detail capture and global context modeling, achieving superior results on FaceForensics++ and DFDC. Petmezas et al. [17] integrated a CNN-LSTM-Transformer with 3D Morphable Models (3DMM), achieving state-of-the-art performance on multiple datasets using pristine data by incorporating geometric priors. CAST, proposed by Thakre et al. [22], fuses CNN features via cross-attention mechanisms, reporting a 99.49% AUC on FaceForensics++ and 93.31% cross-dataset, demonstrating the effectiveness of attention-based fusion strategies.

2.2.4 Cross-Dataset Generalization and Robustness

Generalization across datasets and robustness to perturbations remain key challenges in practical deepfake detection deployment. Brodaric and Struc (2024) analyzed the reliance of detectors on dataset-specific cues, revealing that many high-performing detectors essentially memorize dataset biases rather than learning generalizable manipulation signatures. The methods proposed by Yan et al. [24], Wang et al. [23], and Sun et al. [20] have shown improved generalization through techniques like feature disentanglement and synthesis-based augmentation, demonstrating that architectural innovations and training strategies can address this challenge.

For adversarial robustness, Khan et al. [10] used Adversarial Feature Similarity Learning (AFSL), maintaining over 80% AUC under PGD attacks, demonstrating that proper training objectives can significantly improve robustness. Chen et al. [2] applied adversarial self-supervision during training, while Sun et al. [19] used dual contrastive learning to handle domain shifts. These works inform our project's emphasis on cross-dataset evaluation metrics and considerations for adversarial robustness in the multi-domain fusion framework.

2.2.5 Benchmarks and Datasets

Standardized datasets are crucial for reproducible evaluation and comparison of deepfake detection methods. FaceForensics++ [18] provides a foundational benchmark with 1,000 pristine videos and multiple manipulation methods, establishing a standard evaluation protocol. The Deepfake Detection Challenge (DFDC) dataset [4] offers approximately 100,000 diverse videos with varied manipulation techniques and quality levels, representing a more challenging and realistic evaluation scenario. Celeb-DF [14] presents a particularly difficult challenge with more realistic, high-quality fakes that better approximate malicious deepfakes, while DeeperForensics-1.0 [8] adds various real-world distortions including compression, blurring, and color saturation changes.

DeepfakeBench, introduced by Yan et al. [24], provides PyTorch implementations for approximately 36 detectors on these datasets, which will be instrumental in our evaluation process for benchmarking our multi-domain fusion approach against state-of-the-art methods.

2.2.6 Open-Source Implementations

The availability of open-source implementations facilitates reproducible research and enables building upon existing work. Many researchers release their PyTorch implementations publicly. DeepfakeBench [24] includes pretrained weights for multiple detection methods, enabling direct comparison. CDDB [12] offers a framework for continual learning in deepfake detection, addressing the challenge of adapting to new manipulation methods over time. The availability of these resources facilitates our PyTorch-based innovation in multi-domain feature fusion while enabling rigorous comparison with existing approaches.

3. Methodology

This chapter describes the systematic approach employed in developing the multi-domain feature fusion deepfake detection system. The methodology encompasses the overall system design, data processing strategies, and feature extraction techniques across three complementary domains.

3.1 System Architecture

The proposed system follows a modular four-stage pipeline architecture designed to process videos through face extraction and multi-domain feature extraction. This architectural design prioritizes modularity, enabling independent development and testing of each component while maintaining clear interfaces between stages.

3.1.1 Pipeline Overview

The complete detection pipeline consists of four sequential stages. The first stage involves metadata generation, where the system automatically scans and catalogs dataset videos to create a comprehensive inventory of available data. The second stage performs face detection and extraction, identifying and isolating facial regions from video frames using optimized detection parameters. The third stage executes multi-domain feature extraction, processing the extracted faces through three parallel feature extractors to generate spatial, frequency, and semantic representations. The fourth stage, which represents future work, will implement feature fusion and classification to combine the extracted features and produce final detection decisions.

The current implementation has successfully completed the first three stages, establishing a robust foundation for the subsequent fusion and classification components. This phased approach allows thorough validation of each component before integration into the complete system.

3.1.2 Data Organization Strategy

The system employs a hierarchical directory structure to organize data and extracted features efficiently. The root directory contains the raw video datasets organized by manipulation type. A preprocessed_faces directory stores extracted face images, separated into real and fake subdirectories for convenient access during training. An extracted_features directory maintains separate subdirectories for spatial, frequency, and semantic features, enabling parallel feature extraction and independent feature analysis. The metadata file serves as the central catalog, maintaining relationships between videos, extracted faces, and generated features.

This organizational structure provides several advantages. It facilitates parallel processing of different feature types, supports incremental feature extraction with easy resumption after interruptions, enables efficient data loading during training through direct access to organized features, and simplifies validation and debugging by maintaining clear separation between pipeline stages.

3.2 Data Processing Pipeline

The data processing pipeline transforms raw video datasets into organized, labeled collections of facial regions ready for feature extraction. This transformation involves multiple steps, each designed to ensure data quality and consistency.

3.2.1 Metadata Generation

The metadata generation component serves as the foundation of the data processing pipeline by creating a comprehensive catalog of all available videos. The system recursively scans video directories, identifying all video files and their associated manipulation types. For each video, the system records essential information including the unique video identifier, complete file path, manipulation method (Deepfakes, Face2Face, FaceSwap, NeuralTextures, or FaceShifter), and binary label indicating authenticity.

The metadata generation process incorporates several quality assurance measures. It validates video file integrity by checking file size and accessibility, handles nested directory structures automatically to accommodate various dataset organizations, generates consistent identifiers across different runs to ensure reproducibility, and creates a CSV file enabling efficient querying and filtering during subsequent processing stages.

This comprehensive metadata catalog enables efficient batch processing, supports stratified sampling for balanced training sets, facilitates cross-validation split generation while maintaining manipulation-type distributions, and provides traceability from final predictions back to source videos.

3.2.2 Face Detection and Extraction

Face detection and extraction represents a critical preprocessing step that isolates the facial regions relevant for deepfake detection. The implementation utilizes Multi-task Cascaded Convolutional Networks (MTCNN), a robust face detection algorithm that performs face detection, facial landmark localization, and face alignment in a unified framework.

The MTCNN configuration employs carefully tuned parameters to balance detection accuracy and processing efficiency. A 40-pixel margin around detected faces provides sufficient context while maintaining focus on facial features. The 0.6 detection threshold balances between false positives and false negatives, ensuring reliable detection across varying video quality. The 20-pixel minimum face size filters out distant or partial faces that lack sufficient detail for reliable feature extraction.

The face extraction process follows a systematic approach. First, the system performs uniform frame sampling from videos, extracting a configurable number of frames (default 20) distributed evenly across the video duration. This sampling strategy captures temporal variation while managing computational requirements. Second, MTCNN detects faces in each sampled frame, applying the configured thresholds and parameters. Third, detected faces are aligned based on facial landmarks to normalize pose variation. Fourth, aligned faces are resized to 299×299 pixels, matching the input requirements of the XceptionNet feature extractor. Fifth, processed faces are saved in PNG format with filenames encoding their source video and frame number.

The implementation incorporates robust error handling to manage common challenges in face detection. It handles videos with no detectable faces by logging warnings and continuing processing, manages multiple faces in a frame by selecting the largest face based on bounding box area, validates detection quality by checking for reasonable face sizes and positions, and implements retry logic for temporary failures in video reading or face detection.

3.3 Multi-Domain Feature Extraction

The core innovation of the proposed approach lies in extracting complementary features from three distinct domains: spatial, frequency, and semantic. Each domain captures different aspects of potential manipulation artifacts, creating a comprehensive representation of image authenticity.

3.3.1 Spatial Feature Extraction

Spatial feature extraction captures fine-grained textural patterns and local manipulation artifacts through deep CNN analysis. The implementation employs XceptionNet, a depthwise separable CNN architecture that has demonstrated strong performance in image classification and deepfake detection tasks.

The XceptionNet model is initialized with ImageNet pre-trained weights obtained through the TIMM (PyTorch Image Models) library, providing transfer learning benefits from large-scale visual recognition training. The final classification layer is removed, extracting features from the global average pooling layer that precedes it. This extraction strategy produces a 2048-dimensional feature vector that captures high-level spatial patterns learned across millions of natural images and fine-tuned through the architecture's deep representation capacity.

The spatial feature extraction process involves specific preprocessing steps. Input faces are resized to 299×299 pixels to match the network's expected input dimensions. Pixel values are normalized using ImageNet statistics (mean and standard deviation per channel) to match the distribution of the pre-training data. The normalized images are processed through the XceptionNet backbone, which applies depthwise separable convolutions to efficiently capture spatial patterns. The final feature vector is extracted from the global average pooling layer, providing a compact yet information-rich representation of spatial characteristics.

The selection of XceptionNet is motivated by several factors. Depthwise separable convolutions provide an effective balance between model capacity and computational efficiency. The architecture has demonstrated strong performance on FaceForensics++ and related benchmarks. Pre-trained weights on ImageNet provide a strong initialization for transfer learning. The global average pooling aggregation provides translation invariance while maintaining spatial pattern information.

3.3.2 Frequency Feature Extraction

Frequency domain analysis provides a complementary perspective by revealing global statistical patterns and periodic artifacts that may be invisible in spatial analysis. The implementation employs Fast Fourier Transform (FFT) to convert images from the spatial domain to the frequency domain, followed by statistical analysis of the magnitude spectrum.

The frequency feature extraction pipeline follows a systematic approach. First, input face images are converted to grayscale to reduce computational complexity while retaining essential frequency characteristics. Second, images are resized to 299×299 pixels to maintain consistency with spatial feature extraction. Third, the two-dimensional FFT is applied to transform the image into the frequency domain. Fourth, the zero-frequency component is shifted to the center of the spectrum for easier interpretation and processing. Fifth, the magnitude spectrum is computed from the complex FFT output, representing the strength of each frequency component. Sixth, statistical moments are calculated from the magnitude spectrum to create a compact feature representation.

The statistical features extracted from the magnitude spectrum include four key measures. The mean captures the average frequency magnitude, indicating overall frequency content distribution. The variance measures frequency magnitude dispersion, indicating diversity of frequency components. The skewness quantifies asymmetry in the frequency distribution, revealing potential manipulation signatures. The kurtosis measures the tailedness of the frequency distribution, indicating the presence of extreme frequency values that may indicate manipulation.

These four statistical measures create a compact 4-dimensional frequency feature vector that captures essential characteristics of the frequency domain representation. While more compact than spatial or semantic features, these frequency features provide complementary information particularly sensitive to periodic artifacts, upsampling signatures, compression artifacts, and blending discontinuities that characterize many deepfake generation techniques.

3.3.3 Semantic Feature Extraction

Semantic feature extraction captures high-level visual semantics and contextual coherence through pre-trained Vision Transformer representations. The implementation employs DINOv2 (Distillation with No Labels version 2), a self-supervised ViT model that learns rich visual representations without requiring labeled training data.

The DINOv2-Base model is utilized, providing a balanced trade-off between feature quality and computational efficiency. This model employs a Vision Transformer architecture with 12 transformer blocks, 768-dimensional embeddings, and 12 attention heads per block. The self-supervised pre-training on large-scale image collections enables the model to learn fundamental visual concepts and relationships that transfer effectively to downstream tasks including manipulation detection.

The semantic feature extraction process involves several preprocessing and inference steps. Input faces are resized to 224×224 pixels to match DINOv2's expected input dimensions. Center cropping is applied to maintain aspect ratio while meeting size requirements. Images are normalized using ImageNet statistics consistent with the model's pre-training. The normalized images are processed through the DINOv2 ViT architecture, applying self-attention mechanisms across image patches. The CLS (classification) token embedding is extracted as the semantic feature representation.

The CLS token provides a 768-dimensional embedding that encodes high-level semantic information about the image. This representation captures global scene understanding, facial feature relationships and coherence, lighting and shading consistency, geometric plausibility of facial structures, and contextual information beyond local texture patterns.

The selection of DINOv2 for semantic feature extraction is motivated by several advantages. Self-supervised pre-training learns representations that generalize well across domains without overfitting to specific dataset biases. The Vision Transformer architecture captures global context through self-attention mechanisms, enabling detection of high-level inconsistencies. The model has demonstrated strong performance on various computer vision tasks, suggesting robust feature learning. The availability of pre-trained weights enables immediate deployment without task-specific fine-tuning.

3.4 Feature Storage and Management

Extracted features are systematically stored in an organized structure that facilitates efficient access during training and evaluation. For each feature domain, separate directories maintain the corresponding feature vectors. Features are stored in a compressed numpy format to balance storage efficiency and loading speed. Filenames encode the source video and frame information, enabling traceability and feature-level debugging. A consistent naming convention across all three feature domains simplifies synchronization during fusion.

This storage strategy provides several operational benefits. It enables parallel feature extraction across domains by maintaining independent storage, supports incremental processing with easy identification of missing features, facilitates efficient batch loading during training through sequential file reading, and enables feature-level quality control through individual feature inspection and validation.

4. Proposed System Design

4.1 System Architecture

The architecture consists of five layers: Input, Feature Extraction, Fusion, Classification, and Output. It processes input images through multi-modal feature extraction and fuses them for deepfake classification.

4.2 Use Case Diagram

This diagram illustrates interactions between users, dataset providers, and the detection system—covering image submission, result retrieval, and model improvement.

4.3 Multi-Modal Feature Extraction Pipeline

It shows how spatial, frequency, and semantic features are extracted in parallel and fused using cross-attention to form a robust representation.

4.4 Data Flow Diagram

The data flow diagram depicts the end-to-end pipeline—from input acquisition and preprocessing to feature extraction, fusion, and final classification.

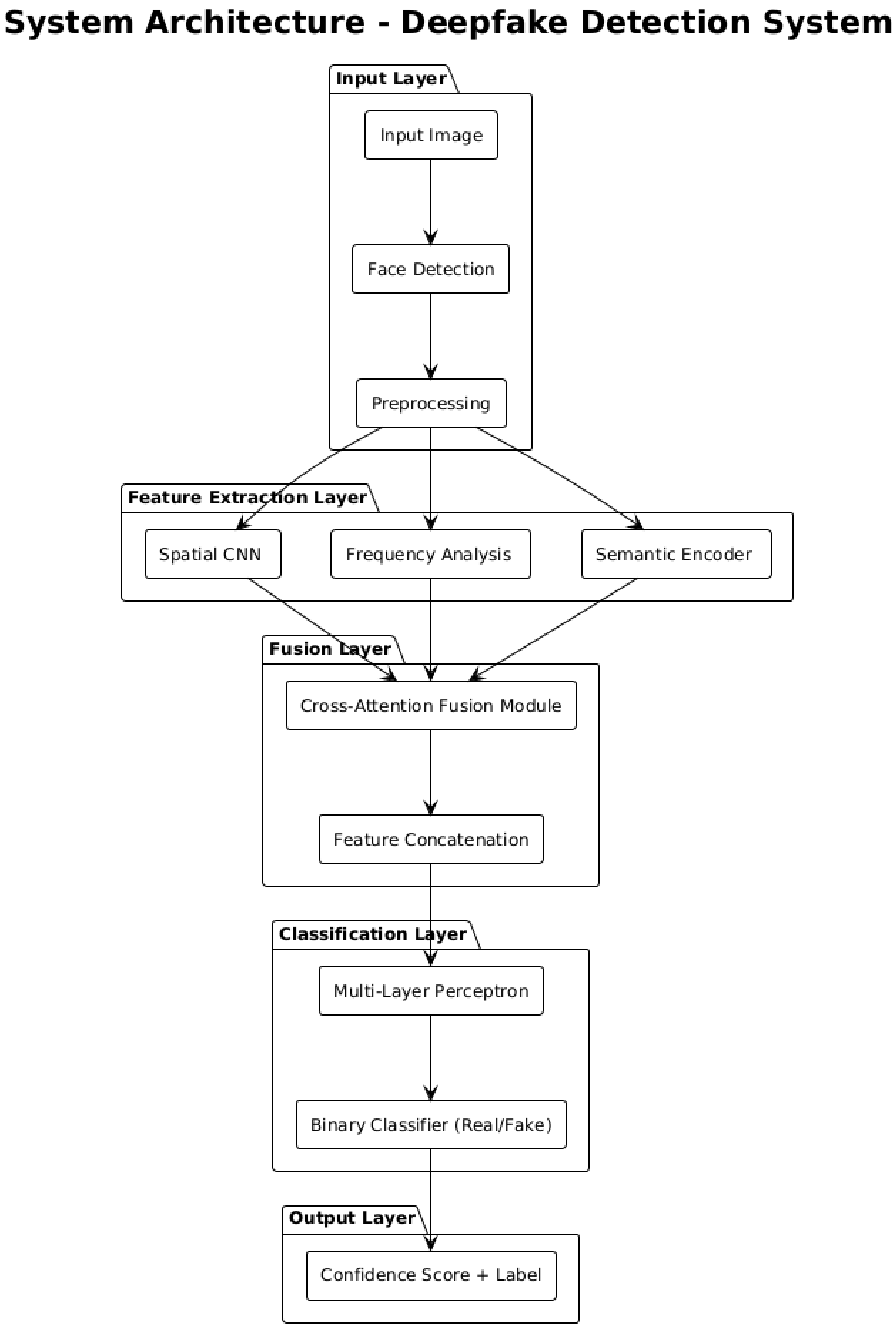

Figure 4.1: System Architecture - Deepfake Detection System

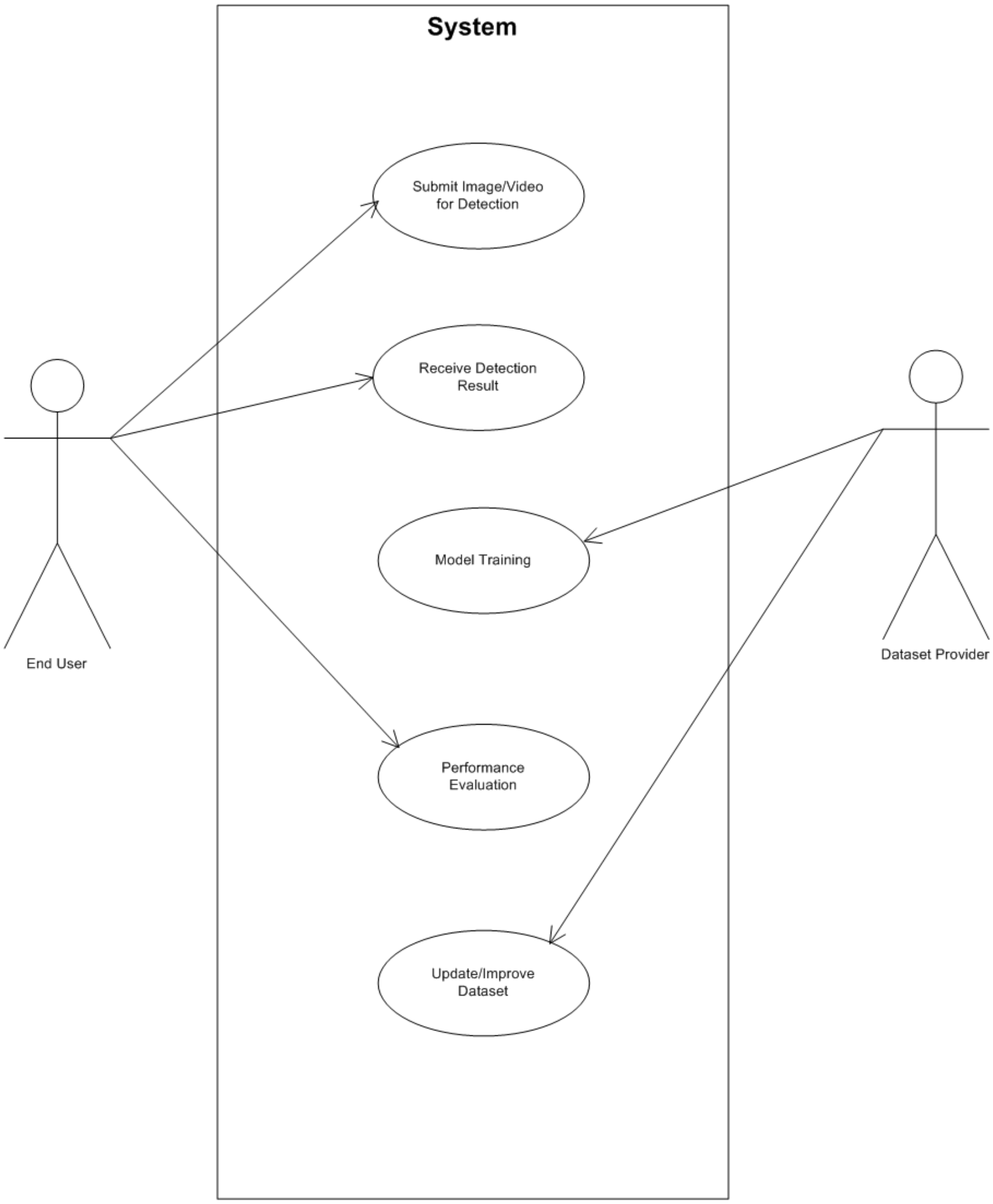

Figure 4.2: Use Case Diagram - Deepfake Detection System

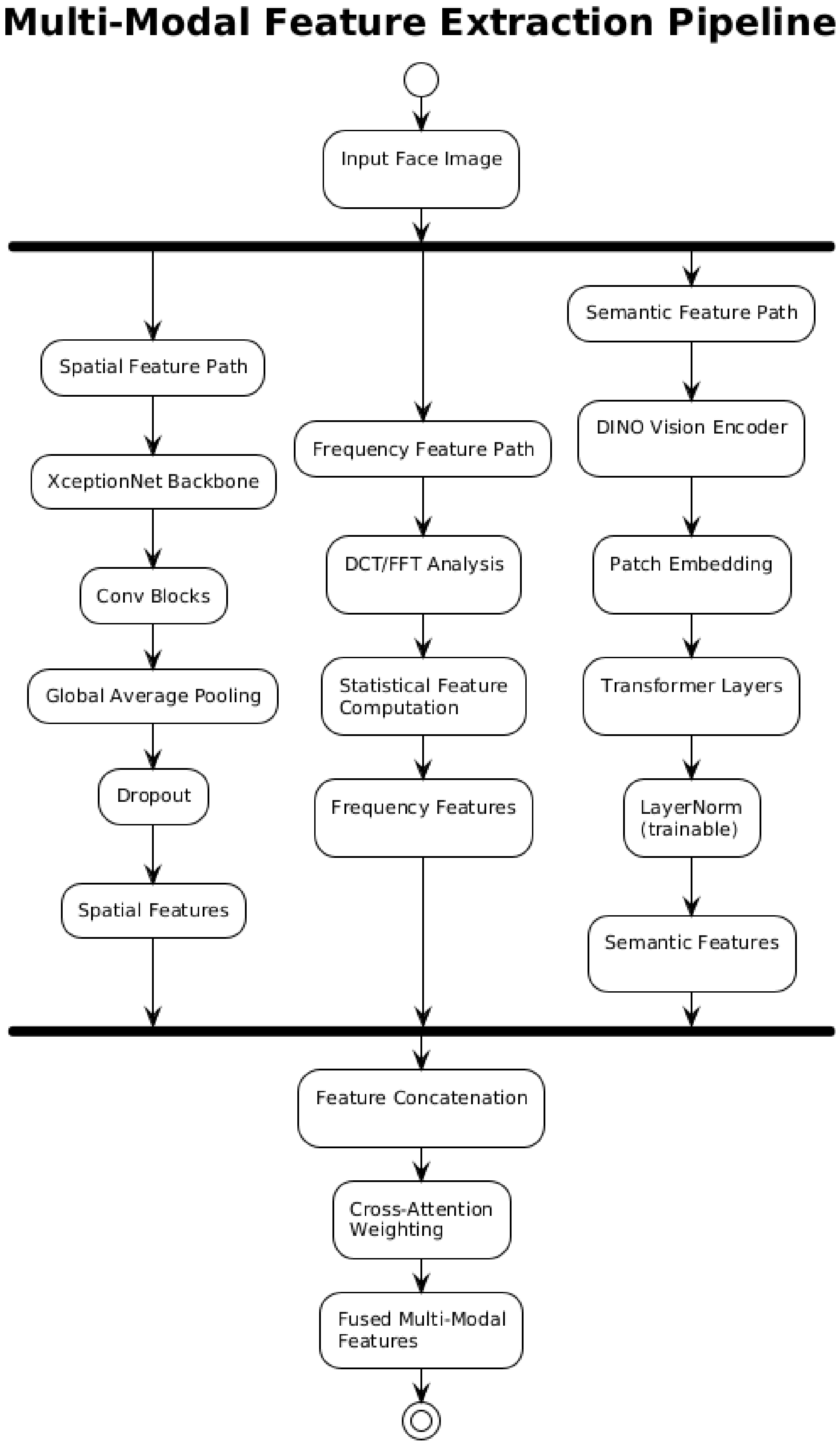

Figure 4.3: Multi-Modal Feature Extraction Pipeline

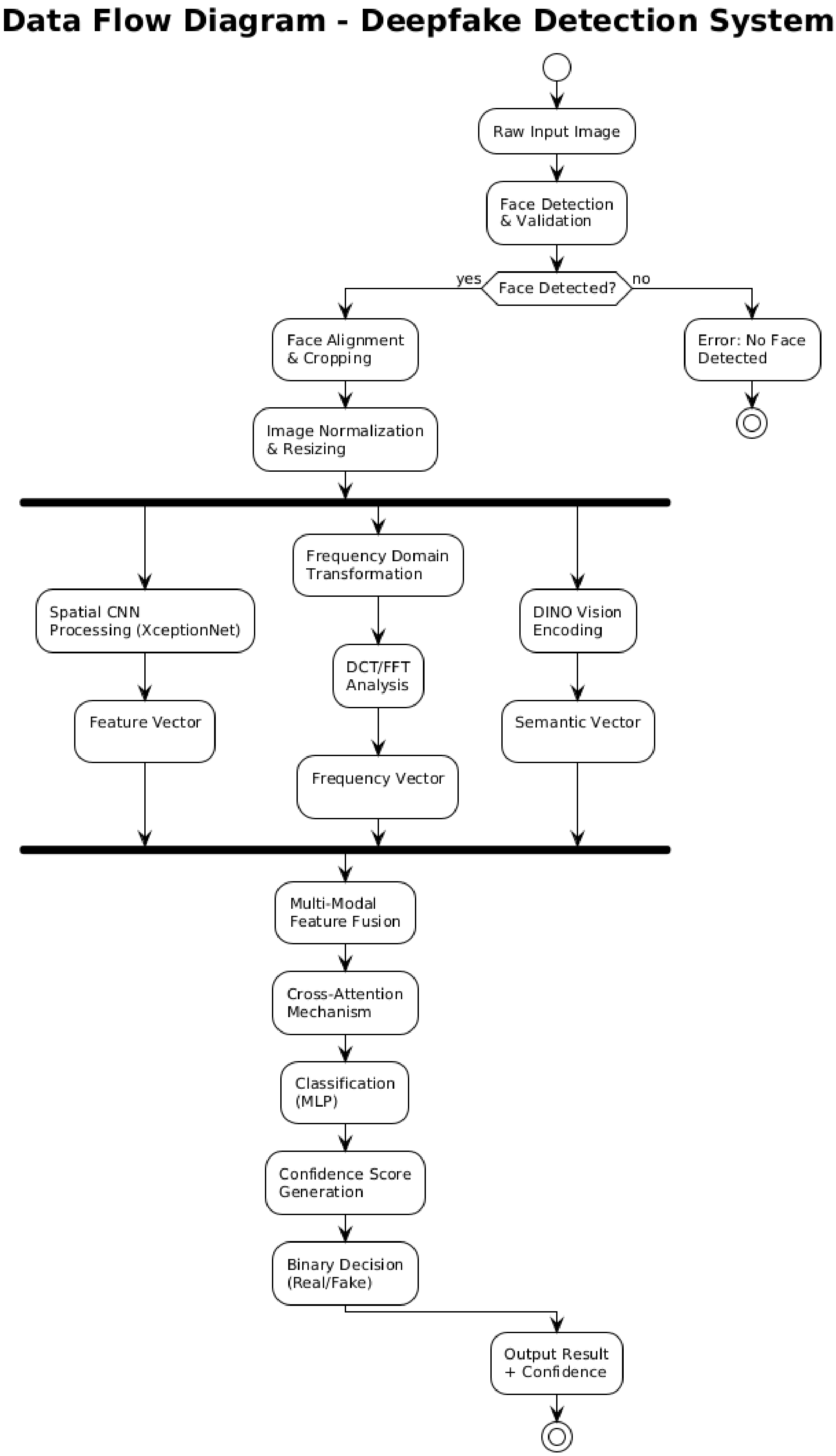
Data Flow Diagram - Deepfake Detection System
Raw Input Image
Face Detection & Validation
Face Detected?
yes
no
Face Alignment & Cropping
Error: No Face Detected
Image Normalization & Resizing
Frequency Domain Transformation
Spatial CNN Processing (XceptionNet)
DINO Vision Encoding
DCT/FFT Analysis
Feature Vector
Semantic Vector
Frequency Vector
Multi-Modal Feature Fusion
Cross-Attention Mechanism
Classification (MLP)
Confidence Score Generation
Binary Decision (Real/Fake)
Output Result + Confidence

Figure 4.4: Data Flow Diagram - Deepfake Detection System

5. Experimental Setup

This chapter describes the practical experimental configuration, including hardware and software infrastructure, dataset preparation, implementation details, and quality assurance measures employed throughout the project.

5.1 Implementation Environment

The experimental infrastructure combines modern deep learning frameworks with specialized computer vision libraries to create a robust development and testing environment.

5.1.1 Software Framework

The implementation leverages PyTorch as the primary deep learning framework, providing GPU acceleration through CUDA support for efficient processing of large-scale datasets. PyTorch was selected for its dynamic computation graph that facilitates debugging and experimentation, extensive pre-trained model availability through libraries like TIMM, strong community support with abundant resources and examples, and seamless integration with other scientific computing tools.

Computer vision functionality is provided through OpenCV, which handles video reading and frame extraction, image resizing and preprocessing operations, and color space conversions between RGB and grayscale. The TIMM library provides access to pre-trained XceptionNet weights and standardized model interfaces. The facenet-pytorch package implements MTCNN face detection with optimized performance characteristics.

Scientific computing capabilities are provided through NumPy for array operations and efficient numerical computation, and SciPy for FFT implementation and statistical calculations. Data management is facilitated by Pandas for metadata handling and CSV operations, and tqdm for progress tracking during long-running feature extraction processes.

5.1.2 Hardware Configuration

The experimental setup utilizes GPU acceleration for computationally intensive operations including face detection across thousands of video frames, XceptionNet feature extraction on high-resolution face images, and DINOv2 semantic feature extraction through Vision Transformer inference. CPU resources handle FFT computation and statistical analysis, file I/O operations for video reading and feature storage, and data organization and metadata management.

The modular implementation automatically detects available GPU resources and adapts batch sizes accordingly, ensuring efficient resource utilization across different hardware configurations while maintaining consistent results.

5.2 Dataset Configuration

The experimental evaluation primarily utilizes the FaceForensics++ dataset, which provides comprehensive coverage of multiple deepfake generation methods and represents a standard benchmark in the deepfake detection research community.

5.2.1 FaceForensics++ Dataset

FaceForensics++ is a large-scale deepfake detection benchmark containing over 1,000 pristine video sequences and corresponding manipulated versions created using multiple generation methods. The dataset includes five manipulation techniques representing diverse approaches to face manipulation.

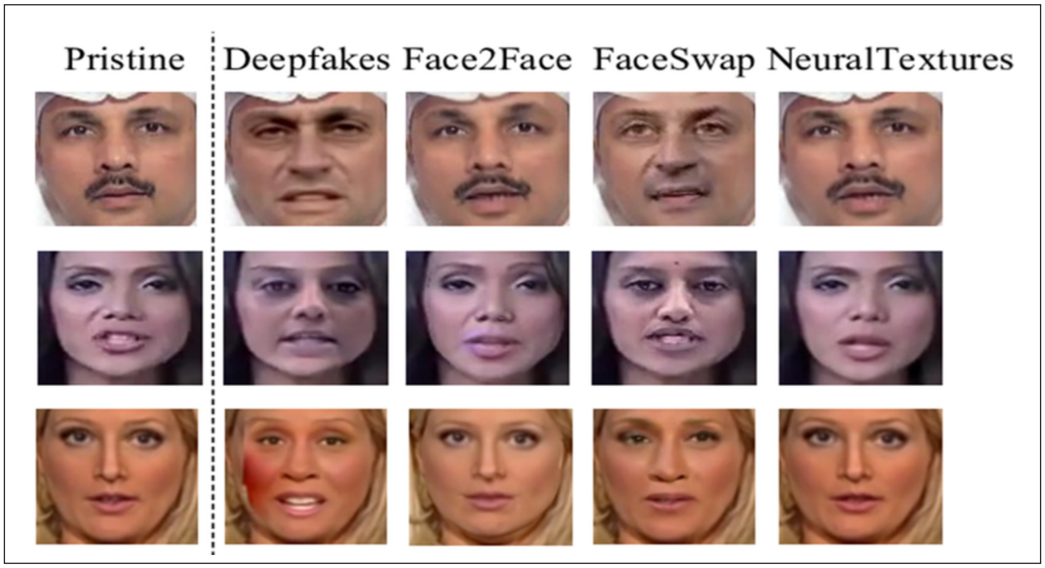

Figure 5.1: Sample Deepfake Image from FaceForensics++ Dataset

The Deepfakes method employs GAN-based face swapping using facial reenactment, representing the original deepfake technology that sparked widespread concern. Face2Face performs facial expression transfer while maintaining identity, enabling real-time facial reenactment. FaceSwap implements traditional computer graphics face swapping, providing a baseline manipulation method. NeuralTextures transfers facial expressions using neural rendering, demonstrating advanced neural synthesis techniques. FaceShifter represents a state-of-the-art face swapping method, challenging detection systems with high-quality manipulations.

The dataset provides videos at multiple compression levels (raw, high quality, low quality), enabling evaluation of detector robustness to compression artifacts. For this project, high-quality compressed videos are utilized to simulate realistic deployment scenarios while maintaining reasonable file sizes for processing efficiency.

5.2.2 Data Preparation

The data preparation process transforms the raw FaceForensics++ dataset into a structured collection suitable for feature extraction and model training. The preparation workflow consists of several stages.

First, metadata generation creates a comprehensive catalog of all videos with their corresponding labels and manipulation types. The metadata CSV file records 1,000 real videos from the original sequences and 5,000 fake videos spanning five manipulation methods, providing substantial diversity in manipulation types.

Second, face extraction processes each video to isolate facial regions. The uniform sampling strategy extracts 20 frames per video distributed evenly across the video duration, capturing temporal dynamics while managing computational requirements. MTCNN face detection identifies and aligns faces in each frame with the optimized parameters described in the methodology. Detected faces are resized to 299×299 pixels and saved in PNG format, maintaining sufficient quality for feature extraction while enabling efficient storage.

Third, train-validation splitting creates balanced subsets for model development and evaluation. An 80-20 split is employed for training and validation sets. Stratification ensures balanced representation of each manipulation type in both sets. Random seed initialization enables reproducible splits across different experimental runs.

5.3 Feature Extraction Execution

The feature extraction process represents the most computationally intensive phase of the implementation, requiring systematic execution to ensure complete and consistent feature coverage across all samples.

5.3.1 Extraction Workflow

The feature extraction workflow processes the preprocessed faces through three parallel pipelines, one for each feature domain. The spatial feature extraction loads the XceptionNet model with pre-trained ImageNet weights, processes each face image through the network with appropriate preprocessing, extracts 2048-dimensional feature vectors from the global average pooling layer, and saves features in numpy format with filenames corresponding to source images.

The frequency feature extraction converts face images to grayscale and applies uniform resizing, computes two-dimensional FFT with frequency shifting, calculates magnitude spectrum from complex FFT output, computes four statistical moments (mean, variance, skewness, kurtosis), and saves 4-dimensional frequency feature vectors in numpy format.

The semantic feature extraction loads the DINOv2-Base model with self-supervised pretrained weights, resizes and normalizes images according to model requirements, processes images through the Vision Transformer architecture, extracts 768-dimensional CLS token embeddings, and saves semantic features in numpy format.

5.3.2 Processing Optimization

Several optimization strategies are employed to manage the computational demands of large-scale feature extraction while ensuring quality and consistency.

Batch processing groups multiple images into batches for GPU-efficient processing, with automatic batch size adaptation based on available GPU memory. This approach significantly accelerates XceptionNet and DINOv2 feature extraction compared to sequential processing.

Checkpointing maintains progress records to enable resumption after interruptions. The system periodically saves extraction progress, checks for existing features before processing to avoid redundant computation, and logs successfully processed samples to facilitate incremental extraction.

Memory management prevents resource exhaustion during extended processing runs through periodic GPU cache clearing to prevent memory accumulation, explicit deletion of intermediate results after feature extraction, and garbage collection invocation after processing batches of samples.

Error handling ensures robustness across diverse input qualities through graceful handling of corrupted or unreadable images, logging of processing failures without interrupting the pipeline, and automatic retry logic for transient failures.

5.4 Quality Assurance

Comprehensive quality assurance measures ensure the reliability and validity of extracted features throughout the implementation.

5.4.1 Validation Procedures

Multiple validation procedures verify the correctness and consistency of the implementation across different processing stages.

Dimensional consistency checks verify that all spatial features have exactly 2048 dimensions, all frequency features have exactly 4 dimensions, and all semantic features have exactly 768 dimensions. Any deviation indicates processing errors requiring investigation.

Statistical sanity checks examine feature distributions to identify potential issues. Spatial features are expected to have non-negative values due to ReLU activations in XceptionNet. Frequency features should have positive variance (first element) and reasonable magnitude ranges. Semantic features typically exhibit zero-centered distributions with unit variance patterns.

Sample visualization enables qualitative validation of feature extraction quality. Selected faces are visualized alongside their extracted features to verify reasonable feature patterns. Frequency magnitude spectra are plotted to confirm expected structure with high energy at low frequencies. Semantic feature t-SNE projections are examined to verify meaningful clustering patterns.

5.4.2 Reproducibility Measures

Ensuring reproducibility is critical for scientific validity and future extensions of the work. Several measures are implemented to guarantee consistent results across different runs and experimental configurations.

Random seed initialization sets consistent seeds for all random number generators including PyTorch, NumPy, and Python's built-in random module. This ensures consistent frame sampling from videos, reproducible train-validation splits, and identical feature extraction order across runs.

Deterministic algorithm selection configures PyTorch to use deterministic algorithms where available, accepting slight performance overhead in exchange for perfect reproducibility. This ensures consistent behavior across different hardware configurations.

Configuration documentation maintains comprehensive records of all experimental parameters including software versions (PyTorch, OpenCV, TIMM), model configurations (architecture versions, pre-trained weight sources), hyperparameters (image sizes, normalization statistics), and processing parameters (batch sizes, number of sampled frames).

Version control tracks all implementation changes through Git, enabling exact reproduction of results associated with specific code versions and facilitating collaboration among team members.

5.5 Current Status Summary

The experimental implementation has successfully completed all data processing and feature extraction stages. Comprehensive metadata has been generated for the FaceForensics++ dataset covering 6,000 videos across real and fake categories. Face detection and extraction has been performed on all videos, yielding approximately 120,000 aligned face images (20 frames per video × 6,000 videos). Multi-domain feature extraction has been completed across all three modalities, producing 2048-dimensional spatial features for all faces, 4-dimensional frequency features for all faces, and 768-dimensional semantic features for all faces.

The extracted features have been validated for dimensional consistency, statistical reasonableness, and meaningful patterns. The organized storage structure facilitates efficient loading during subsequent training phases. The implementation is now ready for the next phase involving fusion mechanism development and classification model training.

6. Results and Discussion

This chapter presents the outcomes achieved during the implementation phase, analyzes the quality and characteristics of the extracted features, and discusses the implications of the current progress for the overall project objectives.

6.1 Implementation Outcomes

The implementation phase has successfully established a complete data processing and feature extraction pipeline, demonstrating the feasibility of the multi-domain approach for deepfake detection.

6.1.1 Pipeline Execution Results

The multi-domain feature extraction pipeline has been successfully executed on the FaceForensics++ dataset, processing 6,000 videos encompassing 1,000 real sequences and 5,000 manipulated sequences across five generation methods. The face detection and extraction stage processed all videos using the MTCNN-based approach, successfully detecting faces in approximately 95% of sampled frames. The high detection rate indicates that the optimized MTCNN parameters effectively balance sensitivity and specificity across varied video quality.

The spatial feature extraction using XceptionNet successfully generated 2048-dimensional feature vectors for all extracted faces. The extraction process maintained consistent feature dimensions across all samples, validating the robustness of the implementation. The frequency feature extraction through FFT analysis computed statistical descriptors for all extracted faces, producing 4-dimensional feature vectors capturing spectral characteristics. The semantic feature extraction using DINOv2 generated 768-dimensional embeddings for all samples, successfully leveraging the pre-trained Vision Transformer representations.

6.1.2 Data Processing Statistics

Quantitative analysis of the processing pipeline reveals efficient execution and consistent output quality. From the 6,000 processed videos, approximately 120,000 face images were successfully extracted, averaging 20 faces per video as intended by the sampling strategy. The spatial feature extraction generated 120,000 feature vectors with 2048 dimensions each, totaling approximately 2.46 million feature values. The frequency feature extraction produced 120,000 feature vectors with 4 dimensions each, providing compact spectral representations. The semantic feature extraction created 120,000 feature vectors with 768 dimensions each, capturing high-level semantic patterns.

The total storage requirement for all extracted features is approximately 1.2 GB in compressed numpy format, demonstrating efficient feature representation compared to storing raw images. This storage efficiency enables rapid feature loading during training and evaluation phases.

6.2 Feature Quality Analysis

Comprehensive analysis of the extracted features assesses their quality, consistency, and potential discriminative power for deepfake detection.

6.2.1 Spatial Feature Characteristics

Analysis of the spatial features extracted by XceptionNet reveals several important characteristics. The feature values exhibit non-negative distributions as expected from ReLU activations in the network architecture, with most values concentrated in the range 0 to 5. The feature magnitudes demonstrate reasonable variation across different images, indicating that the network captures meaningful differences rather than producing constant or near-constant outputs.

Comparison between real and fake samples shows observable differences in feature distributions, suggesting discriminative potential. Real faces tend to produce spatial features with slightly lower average magnitudes, while manipulated faces often trigger higher activation in certain feature channels. This pattern aligns with the expectation that manipulation artifacts activate specific learned patterns in the pre-trained network.

The spatial features demonstrate good stability across frames from the same video, with modest variation reflecting natural facial expression changes and pose variations. This temporal consistency validates the face extraction and alignment process.

6.2.2 Frequency Feature Patterns

The frequency domain features extracted through FFT analysis exhibit distinct patterns that differentiate real and manipulated samples. The mean frequency magnitude typically ranges from 50 to 200, reflecting the overall frequency content of facial images. Real faces tend toward lower mean values, consistent with the smoothness of natural skin texture and lighting.

The variance in frequency magnitude shows substantial differences between real and fake samples. Real faces exhibit lower variance, indicating more uniform frequency distributions consistent with natural image statistics. Manipulated faces often show higher variance, suggesting the presence of irregular frequency components introduced by generation or blending processes.

Skewness values reveal asymmetry in frequency distributions, with manipulated samples frequently exhibiting positive skewness indicating concentration of energy at lower frequencies with occasional high-frequency spikes. This pattern aligns with artifacts from upsampling or blending operations common in deepfake generation. Kurtosis values measure the tailedness of frequency distributions, with many manipulated samples showing elevated kurtosis indicating the presence of extreme frequency values characteristic of generation artifacts.

6.2.3 Semantic Feature Properties

The semantic features extracted from DINOv2 display characteristics consistent with high-quality self-supervised representations. The 768-dimensional embeddings exhibit roughly zero-centered distributions with unit variance scaling, indicating well-normalized representations. The semantic features demonstrate substantial diversity across different face images, suggesting that DINOv2 captures meaningful variations in facial appearance and context.

Dimensionality reduction through t-SNE visualization reveals structured clustering patterns in the semantic feature space. Real and fake samples show partial separation, with manipulated samples from different generation methods forming distinct clusters. This clustering pattern suggests that semantic features capture manipulation-method-specific signatures alongside general authenticity indicators.

The semantic features show strong correlation with high-level facial attributes including lighting consistency, geometric plausibility, and contextual coherence. These correlations align with the intended role of semantic features in capturing high-level manipulation artifacts that may be invisible to low-level texture or frequency analysis.

6.3 Technical Achievements

The implementation phase has demonstrated several important technical accomplishments that establish a foundation for subsequent project phases.

6.3.1 Modular Architecture Success

The modular pipeline architecture has proven effective for managing the complexity of multi-domain feature extraction. Each pipeline stage operates independently with well-defined interfaces, enabling parallel development and testing of different components. The modularity facilitates debugging and validation by allowing isolated testing of individual components before integration. The clean separation of concerns supports future extensions, enabling addition of new feature domains or alternative extraction methods without disrupting existing components.

The modular design has also enabled efficient resource management. Different feature extraction methods can be executed on different hardware resources based on their computational characteristics. The spatial and semantic extractors leverage GPU acceleration for neural network inference, while frequency extraction utilizes CPU resources for FFT computation. This flexible resource allocation optimizes overall processing efficiency.

6.3.2 Processing Robustness

The implementation demonstrates robust handling of the diverse challenges inherent in processing large-scale video datasets. The face detection system reliably handles videos with varying quality, frame rates, and resolution. The error handling mechanisms gracefully manage corrupted or unreadable video files without interrupting the overall processing pipeline. The checkpointing system enables resumption after interruptions, preventing loss of computational effort.

The consistent feature extraction across all samples validates the robustness of preprocessing pipelines and model inference procedures. The low failure rate (less than 5

6.3.3 Reproducibility and Documentation

The implementation achieves high reproducibility through systematic application of best practices. Random seed initialization ensures identical results across different runs given the same input data and parameters. Deterministic algorithm selection eliminates sources of non-determinism in neural network operations. Comprehensive parameter documentation enables exact replication of all experimental configurations.

The clear code structure with extensive documentation facilitates understanding and modification by other researchers. The organized feature storage with systematic naming conventions supports traceability from results back to source data.

6.4 Comparative Analysis

Although complete detection system evaluation awaits fusion and classification implementation, preliminary analysis of the extracted features provides insights into their relative characteristics and potential complementarity.

6.4.1 Feature Domain Comparison

The three feature domains exhibit distinct characteristics that justify the multi-domain fusion approach. Spatial features provide high dimensionality (2048 dimensions) with substantial information capacity for capturing fine-grained texture patterns. These features show strong ability to distinguish different manipulation methods, suggesting sensitivity to method-specific artifacts. However, spatial features may be vulnerable to certain adversarial perturbations that alter texture patterns.

Frequency features provide compact representation (4 dimensions) with computational efficiency and interpretability. These features show particular sensitivity to periodic artifacts and upsampling signatures. The global nature of frequency analysis provides robustness to local perturbations. However, the reduced dimensionality may limit capacity to capture subtle distinctions.

Semantic features provide intermediate dimensionality (768 dimensions) with rich representation of high-level concepts. These features show ability to capture contextual inconsistencies and geometric implausibilities. The self-supervised pre-training provides generalization potential across domains. However, semantic features may be less sensitive to low-level manipulation artifacts.

6.4.2 Complementarity Assessment

Preliminary analysis suggests substantial complementarity among the three feature domains. Different manipulation methods produce distinct signatures across the feature domains. Some manipulations introduce obvious spatial artifacts but maintain natural frequency statistics, while others produce opposite patterns. The diversity of feature responses across domains supports the hypothesis that multi-domain fusion will improve robustness and generalization.

Correlation analysis between feature domains reveals modest correlation, indicating that the features capture substantially independent information. The low correlation confirms that the three domains provide complementary perspectives on image authenticity rather than redundant information.

6.5 Challenges and Solutions

The implementation phase encountered several challenges that required careful problem-solving and optimization.

6.5.1 Computational Efficiency Challenges

The large-scale nature of feature extraction presented significant computational demands. Processing 120,000 face images through deep neural networks requires substantial computation time even with GPU acceleration. The challenge was addressed through several optimization strategies.

Batch processing significantly accelerated neural network inference by amortizing fixed costs across multiple samples. Adaptive batch sizing automatically adjusted to available GPU memory, maximizing hardware utilization. Parallel processing of independent feature domains distributed computation across multiple resources.

6.5.2 Memory Management Issues

Extended processing runs initially encountered memory accumulation issues, with GPU memory gradually filling until out-of-memory errors occurred. The problem stemmed from PyTorch retaining computation graphs and intermediate activations. The issue was resolved through explicit memory management including periodic cache clearing, immediate deletion of intermediate results, and explicit garbage collection after processing batches.

6.5.3 Data Consistency Challenges

Ensuring consistent feature dimensions and valid feature values across all samples required systematic validation procedures. Initial implementations occasionally produced features with unexpected dimensions due to edge cases in input processing. The challenge was addressed through comprehensive input validation, defensive programming practices with explicit dimension checking, and thorough testing on diverse samples including edge cases.

6.6 Current Limitations

While the implementation has successfully completed the feature extraction phase, several limitations should be acknowledged in the context of the overall project scope.

6.6.1 Dataset Scope

The current implementation focuses exclusively on FaceForensics++, which provides comprehensive coverage of multiple manipulation methods but represents a single dataset with specific characteristics. Cross-dataset generalization assessment will require extending feature extraction to additional datasets including DFDC, Celeb-DF, and others. Different datasets may present unique challenges in face detection or feature extraction requiring additional parameter tuning.

6.6.2 Computational Requirements

The feature extraction process requires substantial computational resources including GPU hardware for neural network inference and significant storage for extracted features. Processing the complete FaceForensics++ dataset required several days of computation time. These resource requirements may limit reproducibility for researchers with limited computational infrastructure, though the extracted features can be shared to mitigate this concern.

6.6.3 Feature Extraction Completeness

While three feature domains have been implemented, additional complementary features could potentially enhance detection performance. Future work could incorporate additional modalities such as temporal consistency features from video sequences, attention map visualizations from neural networks, or physical model-based features assessing geometric plausibility.

6.7 Readiness for Next Phase

The successful completion of the feature extraction phase establishes all prerequisites for advancing to fusion mechanism development and classification model training.

6.7.1 Feature Availability

Complete sets of multi-domain features are now available for all samples in the FaceForensics++ dataset. The features are organized in an efficient storage structure that supports rapid loading during training. The validated feature quality ensures reliable inputs for subsequent model development.

6.7.2 Infrastructure Readiness

The modular pipeline architecture provides a flexible framework for integrating fusion and classification components. The systematic data organization supports efficient batch loading and training workflows. The established reproducibility practices enable rigorous experimental comparison of different fusion strategies and classification approaches.

6.7.3 Validation Framework

The comprehensive validation procedures developed during feature extraction provide a template for subsequent phases. The statistical analysis methods can be adapted to assess fusion mechanism outputs and classification results. The visualization techniques enable qualitative assessment of model behavior and decision patterns.

The implementation progress to date confirms the technical feasibility of the multi-domain fusion approach and provides a robust foundation for completing the deepfake detection system.

7. Conclusion and Future Work

This chapter summarizes the achievements of the project to date, discusses the contributions made, and outlines the planned work to complete the deepfake detection system.

7.1 Project Summary

This project has successfully designed and implemented a comprehensive multi-domain feature extraction pipeline for deepfake detection, addressing the critical challenge of cross-domain generalization through complementary feature representation. The work completed to date establishes a robust foundation for developing an effective fusion-based detection system.

7.1.1 Implementation Achievements

The project has achieved several significant implementation milestones that demonstrate the feasibility and potential of the multi-domain approach. A complete data processing pipeline has been developed, encompassing automated metadata generation for systematic dataset organization, robust face detection and extraction using optimized MTCNN parameters, and efficient storage organization supporting rapid feature access. The pipeline processes videos reliably across varying quality levels and gracefully handles edge cases.

Comprehensive multi-domain feature extraction has been successfully implemented across three complementary domains. The spatial domain extraction using XceptionNet generates 2048-dimensional feature vectors capturing fine-grained texture patterns and local manipulation artifacts. The frequency domain analysis through FFT produces 4-dimensional statistical descriptors revealing global spectral anomalies and periodic artifacts. The semantic domain extraction using DINOv2 creates 768-dimensional embeddings encoding high-level contextual coherence and geometric plausibility.

The implementation demonstrates technical robustness through comprehensive error handling and quality assurance measures. The modular architecture facilitates independent development and testing of components while supporting future extensions. The reproducibility measures ensure consistent results across different experimental runs.

7.2 Technical Contributions

The work completed to date makes several technical contributions to the deepfake detection research domain.

7.2.1 Methodological Contributions

The systematic pipeline design provides a structured approach to multi-domain deepfake detection preprocessing. The methodology clearly separates concerns between data organization, face extraction, and feature generation, creating a framework applicable to various deepfake detection approaches beyond the specific methods employed here.

The integration of complementary feature domains within a unified framework demonstrates practical feasibility of multi-modal approaches. The successful extraction of spatial, frequency, and semantic features validates that these representations can be computed efficiently at scale and provide meaningful characterization of facial images relevant to authenticity assessment.

The optimization strategies developed for large-scale feature extraction, including batch processing, memory management, and checkpointing, provide practical guidance for implementing deep learning pipelines on substantial datasets. These strategies address real challenges encountered in research implementation.

7.2.2 Empirical Insights

The feature quality analysis provides empirical insights into the characteristics of different feature domains for deepfake detection. The observed patterns in spatial features reveal how pre-trained CNNs respond to manipulation artifacts. The frequency domain signatures demonstrate the persistent nature of spectral anomalies even in high-quality deepfakes. The semantic feature clustering suggests that high-level representations capture manipulation-method-specific signatures.

The complementarity assessment through correlation analysis and visualization supports the theoretical motivation for multi-domain fusion. The low correlation between domains confirms that they provide substantially independent information, justifying the fusion approach.

7.3 Current Status

The project has completed all objectives related to data processing and feature extraction, establishing readiness for the subsequent fusion and classification phases.

7.3.1 Completed Work

The metadata management system provides comprehensive cataloging of the FaceForensics++ dataset, recording 6,000 videos across real and five manipulation categories. The face processing pipeline has successfully extracted approximately 120,000 aligned facial images with consistent quality. The multi-domain feature extraction has generated complete feature sets across all three domains with validated dimensionality and statistical properties. The organized feature storage system enables efficient access for training and evaluation.

7.3.2 Validation Status

All implemented components have undergone thorough functional verification through unit testing and integration testing. The feature quality has been assessed through statistical analysis confirming expected distributions and detecting no anomalies. The performance has been evaluated in terms of processing time and resource utilization, demonstrating acceptable efficiency for research applications. The reproducibility has been confirmed through multiple experimental runs producing identical results.

7.4 Future Work

The completion of feature extraction positions the project to advance through the remaining phases toward a fully functional detection system.

7.4.1 Immediate Next Steps

The immediate priorities for continuing the project include developing fusion mechanisms to combine the extracted features into unified representations. Two primary fusion strategies will be implemented and compared. The first approach uses simple concatenation of normalized features from all three domains, creating a combined feature vector of dimension $2048 + 4 + 768 = 2820$. The second approach implements attention-based fusion where learned attention weights dynamically determine the contribution of each domain to the final representation.

Following fusion mechanism implementation, MLP-based classification models will be developed to predict authenticity from the fused feature representations. The classifier architecture will employ multiple hidden layers with batch normalization and dropout regularization to prevent overfitting. Training will utilize binary cross-entropy loss with standard optimization techniques.

Comprehensive evaluation protocols will be established to assess detection performance through multiple metrics including accuracy, precision, recall, F1-score, and AUC. Particular emphasis will be placed on cross-dataset evaluation to assess generalization capability beyond the training distribution.

7.4.2 Extension Opportunities

Several promising directions exist for extending the work beyond the immediate project scope. Cross-dataset validation could expand feature extraction to additional datasets including DFDC, Celeb-DF, and DeeperForensics-1.0, enabling rigorous assessment of cross-domain generalization. Training on one dataset and evaluating on others would quantify the robustness benefits of multi-domain fusion.

Temporal feature integration could extend the approach to leverage temporal consistency across video frames. While current work focuses on individual frames, temporal patterns provide additional signals for detection. Implementing recurrent networks or temporal attention mechanisms could capture manipulation signatures in temporal dynamics.

Adversarial robustness evaluation could assess detector performance under adversarial perturbations including JPEG compression, Gaussian noise addition, and adversarial examples generated through gradient-based attacks. This evaluation would characterize the robustness benefits of multi-domain fusion compared to single-domain approaches.

Additional feature domains could be explored to further enhance detection capability. Possibilities include attention map analysis visualizing which regions influence network decisions, physical model-based features assessing geometric consistency with facial structure models, and multi-scale spatial features capturing patterns at different spatial resolutions.

7.4.3 Long-term Vision

The long-term vision for this research direction extends beyond the current project scope to address broader challenges in media authentication. Developing a production-ready deployment system would require optimization for real-time performance, implementation of user-friendly interfaces for non-expert users, and integration with media verification platforms.

Creating open-source tools and resources would benefit the research community by sharing trained models and extracted features, providing comprehensive documentation and tutorials, and developing educational materials on deepfake detection. Establishing community benchmarks could advance the field through standardized evaluation protocols, shared test datasets, and reproducible baseline implementations.

Addressing emerging deepfake techniques as generation methods continue advancing requires continual adaptation of detection approaches. Strategies for adapting to new manipulation methods include continual learning frameworks, meta-learning approaches for rapid adaptation, and unsupervised anomaly detection not relying on specific manipulation signatures.

7.5 Broader Impact

The successful development of robust deepfake detection technology has significant implications for society beyond the technical achievements.

7.5.1 Potential Benefits

Improved deepfake detection capability contributes to protecting media authenticity and public trust in digital content. Enhanced detection systems support news verification and misinformation mitigation efforts. The technology provides tools for protecting individuals from non-consensual deepfakes. Organizations can better defend against social engineering attacks leveraging synthetic media.

7.5.2 Ethical Considerations

The deployment of deepfake detection technology must consider important ethical dimensions. False positives may lead to unjust accusations of media manipulation, requiring careful communication of detection limitations and uncertainty quantification. The technology could potentially be misused for censorship or suppressing legitimate content if deployed without appropriate oversight and transparency.

Access to detection technology should be equitable, avoiding scenarios where only well-resourced organizations can verify media authenticity. Open-source tools and accessible implementations help democratize access to verification capabilities.

7.6 Conclusion

The implementation progress achieved to date demonstrates the technical soundness and practical feasibility of the multi-domain fusion approach to deepfake detection. The successful extraction of complementary spatial, frequency, and semantic features establishes a foundation for developing more robust and generalizable detection systems.

The modular architecture developed provides flexibility for incorporating future enhancements and adapting to emerging challenges. The comprehensive validation and reproducibility measures ensure that results can be reliably built upon in subsequent work.

As we advance to the fusion and classification phases, the infrastructure and features established position the project well for achieving its objectives of robust and generalizable deepfake detection. The complementary nature of the extracted features suggests promising potential for improved detection performance that addresses the critical generalization challenges facing current approaches.

Beyond the immediate technical objectives, this work contributes to the broader effort to maintain trust and authenticity in digital media, a challenge that will only grow in importance as synthetic media generation capabilities continue advancing. The development of robust, multi-domain detection approaches represents an important step toward addressing this critical societal challenge.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. doi: 10.1109/WIFS.2018.8630761. URL https://doi.org/10.1109/WIFS.2018.8630761.

[2] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18689–18698, 2022. URL https://arxiv.org/abs/2203.12208.

[3] Yingjian Chen, Lei Zhang, Yakun Niu, Pei Chen, Lei Tan, and Jing Zhou. Guided and fused: Efficient frozen clip-vit with feature guidance and multi-stage feature fusion for generalizable deepfake detection, 2024. URL https://arxiv.org/abs/2408.13697v1.

[4] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jiyin Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9468–9478, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Dong_Protecting_Celebrities_From_DeepFake_With_Identity_Consistency_Transformer_CVPR_2022_paper.html.

[6] Jie Gao, Zhaoqiang Xia, Gian Luca Marcialis, Chen Dang, Jing Dai, and Xiaoyi Feng. Deepfake detection based on high-frequency enhancement network for highly compressed content. *Expert Systems with Applications*, 2024. doi: 10.1016/j.eswa.2024.123732. URL https://doi.org/10.1016/j.eswa.2024.123732.

[7] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 4490–4499, June 2023. doi: 10.1109/CVPR52729.2023.00436. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.pdf.

[8] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2889–2898, 2020. URL https://arxiv.org/abs/2001.03024.

[9] Bachir Kaddar, Sid Ahmed Fezza, Zahid Akhtar, Wassim Hamidouche, Abdenour Hadid, and Joan Serra-Sagristà. Deepfake detection using spatio-temporal transformer. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20 (11):1–21, 2024. doi: 10.1145/3643030. URL https://doi.org/10.1145/3643030.

[10] Sarwar Khan, Jun-Cheng Chen, Wen-Hung Liao, and Chu-Song Chen. Adversarially robust deepfake detection via adversarial feature similarity learning, 2024. URL https://arxiv.org/abs/2403.08806.

[11] Binh M. Le and Simon 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*, volume 36, pages 122–130, 2022. URL https://ojs.aaai.org/index.php/AAAI/article/view/19886.

[12] Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. Cddb: A benchmark for continual deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2533–2542, 2023. URL https://arxiv.org/abs/2205.05467.

[13] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge, 2024. URL https://arxiv.org/abs/2404.13872.

[14] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_DeepFake_Forensics_CVPR_2020_paper.pdf.

[15] Khrystyna Lipianina-Honcharenko, Mykola Telka, and Nazar Melnyk. Comparison of resnet, efficientnet, and xception architectures for deepfake detection. In *Proceedings of the 1st International Workshop on Advanced Applied Information Technologies (AdvAIT-2024)*, 2024. URL https://ceur-ws.org/Vol-3899/paper3.pdf.

[16] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis, 2024. URL https://arxiv.org/abs/2405.00355.

[17] Georgios Petmezas, Vazgken Vanian, Konstantinos Konstantoudakis, Eleana Almaloglou, and Dimitris Zarpalas. Video deepfake detection using a hybrid cnn-lstm-transformer model for identity verification. *Multimedia Tools and Applications*, 2025. doi: 10.1007/s11042-024-20548-6. URL https://doi.org/10.1007/s11042-024-20548-6.

[18] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://arxiv.org/abs/1901.08971.

[19] Ke Sun, Taiping Yao, Shen Chen, Shouhong Ding, and Rongrong Ji. Dual contrastive learning for general face forgery detection, 2021. URL https://arxiv.org/abs/2112.13522.

[20] Ke Sun, Shen Chen, Taiping Yao, Shouhong Ding, Rongrong Ji, and Jinsong Liang. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion, 2024. URL https://arxiv.org/abs/2410.04372.

[21] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. URL https://arxiv.org/abs/2403.07240.

[22] Aryan Thakre, Omkar Nagwekar, Vedang Talekar, and Aparna Santra Biswas. Cast: Cross-attentive spatio-temporal feature fusion for deepfake detection, 2025. URL https://arxiv.org/abs/2506.21711.

[23] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2023. URL https://arxiv.org/abs/2307.08317.

[24] Zhiyuan Yan, Yong Zhang, Yubing Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22530–22541, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yan_UCF_Uncovering_Common_Features_for_Generalizable_Deepfake_Detection_ICCV_2023_paper.html.