Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MCFN: A Lightweight Multi-Path Compound Fusion Network for Deepfake Face Detection

Susmita Saha Lagna
North Western University https://orcid.org/0009-0007-9837-1879
Md Raihan Khan
kraihan918@gmail.com
Khulna University of Engineering and Technology https://orcid.org/0009-0004-2461-5461

Research Article

Keywords: Deepfake detection, Multi-path feature fusion, Frequency-domain forensics, Feature-wise Linear Modulation, Cross-path attention, Generalisable face forgery detection

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10605278/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

MCFN: A Lightweight Multi-Path Compound Fusion Network for Deepfake Face Detection

Susmita Saha Lagna[1†] and Md Raihan Khan[1,2*†]

[1]Department of EEE, North Western University, Khulna, 9100, Bangladesh.
[2]Department of EEE, Khulna University of Engineering & Technology, Khulna, 9203, Bangladesh.

*Corresponding author(s). E-mail(s): kraihan918@gmail.com;
Contributing authors: sahalagna6@gmail.com;
†These authors contributed equally to this work.

Abstract

The proliferation of deepfake media poses severe threats to digital trust, biometric security, and information integrity. Most existing detection methods rely on a single feature domain, limiting their ability to generalise across diverse manipulation techniques. In this paper, we propose the Multi-Path Compound Fusion Network (MCFN), a novel architecture that jointly exploits texture, edge/gradient, and frequency-domain cues through three parallel encoders. A Compound Fusion Module enriches the per-path embeddings via cross-path multi-head self-attention and compresses them into a single conditioning token, which dynamically modulates an EfficientNet-B0 backbone through Feature-wise Linear Modulation (FiLM). A primary classifier head and an auxiliary frequency forensic head are jointly optimised with a composite loss comprising label-smoothed binary cross-entropy, an auxiliary frequency loss, and a supervised contrastive loss. Extensive experiments on four benchmarks—FaceForensics++, Celeb-DF, the Google DFD benchmark, and a combined FF++ + Celeb-DF corpus—show that MCFN attains the best accuracy on all four benchmarks and the best or comparable AUC relative to recent state-of-the-art methods, while maintaining only 6.33 M trainable parameters. A systematic ablation study confirms the contribution of each component, with FiLM conditioning and cross-path attention yielding the largest individual gains. These results establish MCFN as an effective, lightweight, and generalisable framework for deepfake face detection.

Keywords: Deepfake detection, Multi-path feature fusion, Frequency-domain forensics, Feature-wise Linear Modulation, Cross-path attention, Generalisable face forgery detection

1 Introduction

The proliferation of deepfake media represents one of the most pressing challenges to digital trust in today's society [1]. Advances in generative models have made it possible to produce photorealistic facial manipulations including face swaps, reenactments, and entirely synthesised identities that are often indistinguishable to the human eye. Such techniques can be readily exploited by malicious actors to spread disinformation, commit identity fraud, or undermine biometric authentication systems. Consequently, developing reliable and generalisable deepfake detection methods has become a matter of critical importance [2, 3].

Most existing detection frameworks model the task as a binary classification problem (real vs. fake) by feeding an input face image through a backbone network to extract global features, which are then passed to a binary classifier. While effective within a single domain, these approaches frequently overfit to the specific artefact patterns present in the training data and generalise poorly to unseen manipulation techniques. A key limitation is their reliance on a single feature domain: methods that exploit only spatial texture, or only frequency-domain statistics, fail to capture the full spectrum of forgery cues. Recent work has shown that deepfakes leave complementary traces across spatial, edge, and spectral representations [4, 5], motivating the design of multi-modal fusion architectures.

In this paper, we propose the **Multi-Path Compound Fusion Network (MCFN)**, a novel architecture that jointly exploits three complementary feature domains, spatial texture, edge/gradient geometry, and frequency-domain statistics, through parallel encoding paths. Rather than simply concatenating these heterogeneous embeddings, MCFN introduces a *Compound Fusion Module* that enriches each path embedding via cross-path multi-head self-attention before compressing them into a single conditioning token. This token dynamically modulates an EfficientNet-B0 backbone through Feature-wise Linear Modulation (FiLM), enabling the backbone to focus on regions and channels flagged as suspicious by the auxiliary paths. A primary classifier head and an auxiliary frequency forensic head are jointly optimised with a composite loss comprising label-smoothed binary cross-entropy, an auxiliary frequency loss, and a supervised contrastive loss.

The main contributions of this work are as follows:

1. **Multi-Path Compound Fusion:** We propose a novel approach that integrates texture, edge, and frequency features via cross-path attention and a gated compound projection, enabling richer cross-modal interaction than simple concatenation.

2. **FiLM-Conditioned Backbone:** We introduce a FiLM-based conditioning mechanism that dynamically modulates the backbone's intermediate activations using the fused multi-path token, significantly improving detection performance.
3. **Consistent Multi-Benchmark Performance:** We evaluate MCFN on four diverse benchmarks—FaceForensics++, Celeb-DF, the Google DFD benchmark, and a combined FF++ + Celeb-DF corpus—including a heterogeneous combined corpus that mixes manipulation families, demonstrating strong and consistent detection across different forgery types, data sources, and quality levels.

The rest of this paper is structured as follows. Section 2 reviews the related works. Section 3 presents the proposed MCFN methodology in detail. Section 4 describes the experimental setup, results, and ablation study. Section 5 concludes the paper.

2 Related Works

Guo et al. [6] proposed the Mining Structured Features (MSF) augmentation framework, which dynamically erases specific image patches to force CNNs to learn global structured forgery clues rather than isolated local artifacts. MSF achieved state-of-the-art AUCs—including 99.87% on HFFD and 98.27% on DFDC—and demonstrated strong cross-dataset generalization, though its reliance on fixed, non-adaptive patch sizes limits its flexibility across varying artifact scales. Guo et al. [7] proposed the Adaptive Manipulation Traces Extraction Network (AMTEN) combined with a CNN to adaptively highlight subtle facial forgery artifacts by suppressing normal image content. Their core findings demonstrated high detection accuracies of 98.52% across multiple face manipulation techniques and 95.17% on images with unknown post-processing. However, a notable weakness is that their simulated post-processing evaluations do not fully capture the complex degradation pipelines of real-world social media platforms. Muñoz-Haro et al. [1] proposed FakeIDet2, a privacy-aware patch-based fake ID detector using dynamic AdaFace loss and Multi-Head Self-Attention to maintain data privacy. While its core findings demonstrate strong generalization on unseen physical attacks (e.g., 9.56% EER on DLC-2021) using only pseudo-anonymized patches, its main weakness is a significant vulnerability to unseen synthetic GAN-generated attacks (37.51% EER) due to being trained exclusively on physical acquisitions.

Guo et al. [7] also proposed a CED-DCGAN framework to detect real-time video deepfakes by analyzing frames for spectral anomalies, achieving a state-of-the-art 98.23% accuracy. The current model exclusively focuses on visual features, ignoring audio deepfakes, and requires continuous adaptation to maintain generalizability against rapidly evolving forgery techniques. Dhanaraj and Sridevi [8] propose an Xception-CNN transfer learning model integrated with rigorous color and illumination pre-processing to detect and localize face-warping deepfakes, achieving a 89.25% detection accuracy and an AUROC of 0.8. However, the framework is constrained by a lack of dedicated face-warping datasets and currently fails to address extreme local variations in expression, global pose changes, and partial occlusions. Wang et al. [9] proposed FakeSpotter, a baseline framework that monitors layer-by-layer neuron activation behaviors (Mean Neuron Coverage) in deep face recognition systems to train a

simple binary classifier for fake face detection. The method achieves an average detection accuracy above 90% across four typical GAN-synthesized face types and shows strong robustness, maintaining an average AUC drop of less than 3.77% under common perturbation attacks. The detector exhibits reduced effectiveness against highly advanced GANs (e.g., StyleGAN2), struggles with minor localized manipulations, and is prone to false alarms from multi-modal inputs like voice-swapped media.

Tan et al. [4] proposed FreqNet, a universal deepfake detection framework that integrates lightweight frequency-domain learning plugins into a CNN classifier to learn source-agnostic, domain-invariant forgery features. By continuously enforcing spectrum learning on both spatial and channel levels of the internal features, FreqNet avoids overfitting to specific artifact patterns and achieves state-of-the-art generalizability (+9.8% performance gain) across 17 diverse generative networks using just 1.9 million parameters. Although highly generalizable, the approach relies entirely on high-frequency information and phase/amplitude spectra, meaning it could suffer from performance degradation under severe low-pass filtering, heavy compression, or post-processing attacks that deliberately destroy high-frequency artifacts.

Li et al. [5] proposed MkfaNet, a specialized deepfake detection backbone that integrates a Multi-Kernel Aggregator (MKA) to adaptively capture spatial context at multiple scales and a Multi-Frequency Aggregator (MFA) to process high- and low-frequency components independently. Evaluated across seven standard benchmarks, MkfaNet variants consistently outperform traditional and modern backbones in both within-domain and cross-domain testing while exhibiting impressive efficiency in parameter usage. While highly efficient and adaptive, the method relies on a fixed training resolution (256×256) and remains susceptible to severe image degradations, such as heavy post-processing compression (e.g., FF-c40), which obscure high-frequency forgery features.

Khormali and Yuan [10] present a deepfake detection framework that extracts high-level facial landmark features via a self-supervised contrastive learning vision Transformer (ViT), which are then modeled as an undirected graph and classified using a combined Graph Convolutional Network (GCN) and Transformer discriminator. The approach relies on a rigid, fixed graph generation algorithm based solely on the spatial coordinates of image patches, and its performance degrades notably when subjected to high video compression due to the unavoidable loss of low-level textural cues. The framework achieves superior out-of-distribution generalizability (90.8% average cross-dataset AUC) and enhanced resilience to post-processing perturbations (like blur and noise) compared to standard CNN baselines, while successfully providing explainable deepfake localization via a custom graph Transformer relevancy map. Ahmed et al. [11] presented a systematic literature review (SLR) that categorizes visual deepfake creation into four distinct facial manipulation classes (synthetic face, face reenactment, facial attribute alteration, and face swap) while categorizing recent defense strategies across spatial, temporal, frequency, and spatio-temporal domains. The survey highlights that contemporary detection systems suffer from severe degradation when dealing with heavy data augmentation or multi-format video compression, and they require immense hardware resources that limit their deployment in real-world,

real-time end-user applications. The study reveals that while current frameworks effectively expose localized visual artifacts and biological irregularities, achieving long-term robustness demands the integration of advanced architectures like transformers and meta-learning to minimize false-positive rates across disparate datasets.

Taken together, prior work spans several complementary strategies. Attention-based detectors localise subtle artifacts through fine-grained, multi-attentional feature learning [12], while reconstruction-based approaches model the compact representation of genuine faces to expose unseen forgeries [13]. Frequency-domain methods exploit the spectral fingerprints left by generative up-sampling [14], and semantic or temporal cues such as mouth dynamics improve robustness on video [15]. A parallel line of work targets generalisation to unseen generators, showing that classifiers trained on one family of fakes overfit to source-specific artifacts [2, 3] and that implicit identity leakage further harms cross-dataset transfer [16]. Multi-modal transformers that combine RGB and frequency streams [17] confirm the value of cross-domain evidence, yet most existing methods still operate within a single dominant domain or fuse modalities by simple concatenation, and few couple complementary texture, edge, and frequency cues into a shared conditioning signal. MCFN targets this gap: rather than concatenating heterogeneous embeddings, it fuses all three domains into a single compound token that *conditions* a weight-shared backbone through FiLM, coupling cross-modal evidence with the backbone's own representation instead of merely appending it.

3 Methodology

3.1 Overview and Problem Formulation

We formulate face forgery detection as a binary classification problem. Given an input face image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, where $H = W = 128$, the goal is to learn a mapping $f_\theta : \mathbb{R}^{3\times H\times W} \to [0, 1]$ that estimates the probability that $\mathbf{x}$ is a generated (fake) face:

$$\hat{y} = f_\theta(\mathbf{x}), \qquad \hat{y} \in [0, 1], \tag{1}$$

where $\hat{y}$ approximates the ground-truth label $y \in \{0, 1\}$ (y=1 denotes a generated/manipulated face and y=0 denotes a genuine face), and θ denotes all learnable parameters of the proposed *Multi-Path Compound Fusion Network* (MCFN).

The MCFN pipeline consists of five stages applied sequentially to $\mathbf{x}$:

1. Three parallel feature-extraction paths produce complementary embeddings $\mathbf{F}_A, \mathbf{F}_B, \mathbf{F}_C \in \mathbb{R}^d$ (d=256), capturing texture, edge, and frequency cues respectively (Section 3.2).
2. A Compound Fusion Module merges these embeddings into a single conditioning token $\mathbf{c} \in \mathbb{R}^D$ (D=512) via cross-path attention and a gated MLP (Section 3.3).
3. A shared spatial feature map $\mathbf{Z}_0$, produced by the stem and lower blocks of an EfficientNet-B0 encoder, is passed to a FiLM-conditioned backbone that is dynamically modulated by $\mathbf{c}$ (Section 3.4).
4. The resulting representation $\mathbf{r} \in \mathbb{R}^{1280}$ feeds two output heads: a primary classifier head and an auxiliary frequency forensic head (Section 3.5).

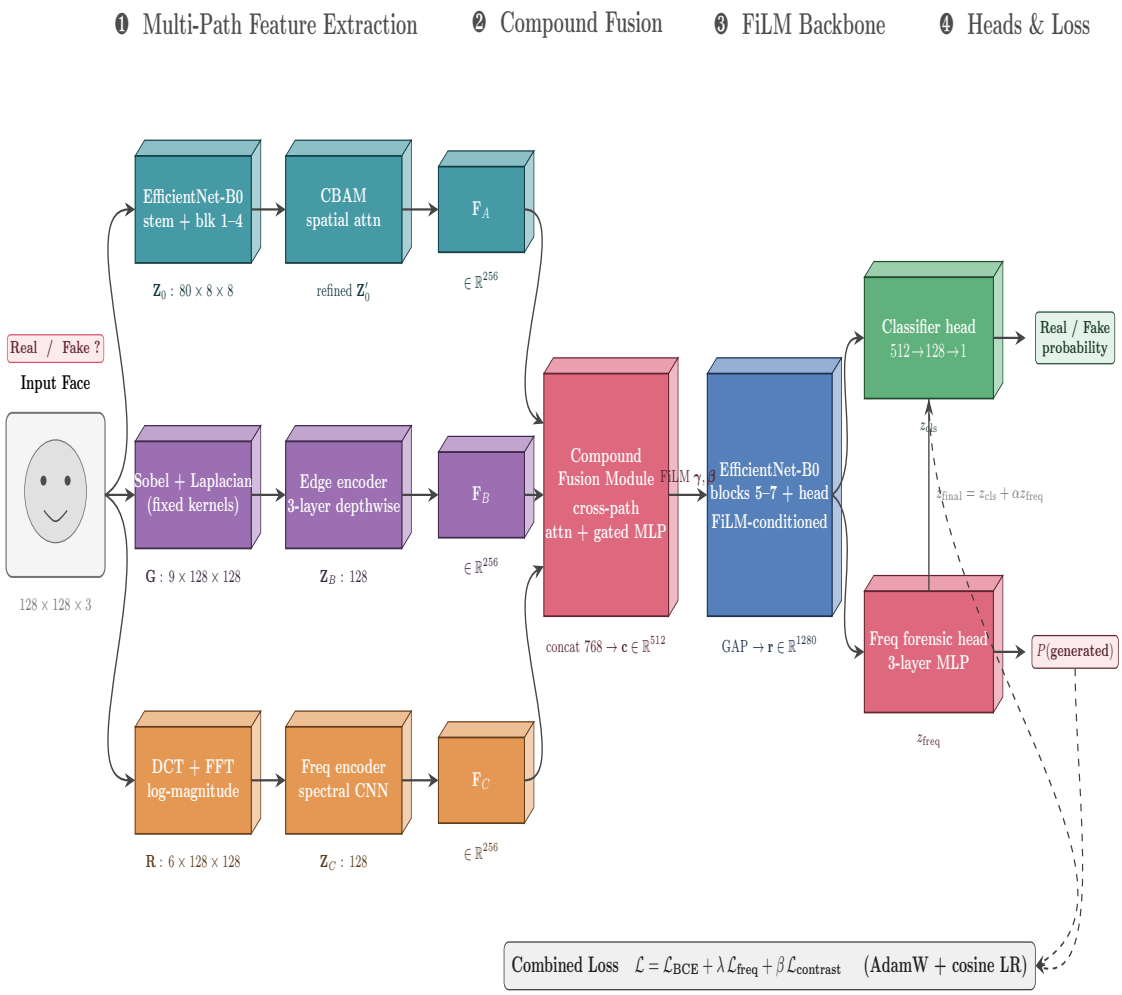


Fig. 1: Overview of the proposed Multi-Path Compound Fusion Network (MCFN). The input face image is processed through three parallel paths capturing texture, edge/gradient, and frequency-domain features. These embeddings are enriched via cross-path attention and compressed into a 512-d compound token by the Compound Fusion Module. This token conditions the EfficientNetB0-based backbone through Feature-wise Linear Modulation (FiLM). The resulting representation feeds a primary classifier head and an auxiliary frequency forensic head, jointly optimized via a combined loss function.

5. All components are jointly optimized using a composite loss combining classification, auxiliary frequency, and contrastive objectives (Section 3.6).

The overall computation can be summarized as:

$$\mathbf{F}_A = \mathcal{E}_A(\mathbf{x}), \quad \mathbf{F}_B = \mathcal{E}_B(\mathbf{x}), \quad \mathbf{F}_C = \mathcal{E}_C(\mathbf{x}), \tag{2}$$

$$\mathbf{c} = \mathcal{G}_{\text{fuse}}(\mathbf{F}_A, \mathbf{F}_B, \mathbf{F}_C), \tag{3}$$

$$\mathbf{r} = \mathcal{B}_{\text{FiLM}}\big(\mathcal{E}_{\text{bridge}}(\mathbf{x}),\, \mathbf{c}\big), \tag{4}$$

$$\hat{y} = \sigma\big(\ell_{\text{cls}}(\mathbf{r}) + \alpha\, \ell_{\text{freq}}(\mathbf{r})\big), \tag{5}$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function and α is a fixed fusion coefficient (Section 3.5).

Design rationale. The central design choice in MCFN is to treat the auxiliary texture, edge, and frequency streams not as parallel classifiers to be averaged, but as a single *conditioning* signal for a shared backbone. Concatenation-based multi-stream detectors fuse heterogeneous embeddings only at the decision layer, so the backbone never sees the cross-modal evidence while forming its own spatial representation. In contrast, MCFN compresses the three enriched path embeddings into one compound token and injects it into the backbone through FiLM (Section 3.4), allowing edge- and frequency-derived cues to re-weight and shift the backbone's intermediate activations before the final decision. Because the texture path and the backbone share the EfficientNet-B0 stem and lower blocks (Eq. 6), this conditioning is achieved with negligible additional parameters, which underlies the favourable accuracy–efficiency trade-off reported in Section 4.5.

3.2 Multi-Path Feature Extraction

The input image is processed simultaneously by three parallel encoders, each specialized for a distinct visual modality: spatial texture, edge geometry, and frequency-domain statistics.

3.2.1 Path A: Texture Encoder

Path A captures spatial texture cues using the stem and first four MBConv blocks of EfficientNet-B0 [18], denoted Φ_{stem} and $\Phi_{1:4}$, followed by a Convolutional Block Attention Module (CBAM) [19]. The intermediate feature map is given by:

$$\mathbf{Z}_0 = \Phi_{1:4}\big(\Phi_{\text{stem}}(\mathbf{x})\big) \in \mathbb{R}^{C_0 \times H_0 \times W_0}, \qquad C_0 = 80. \tag{6}$$

CBAM refines $\mathbf{Z}_0$ via sequential channel and spatial attention. The channel attention map is computed as:

$$\mathbf{M}_c = \sigma\Big(\mathbf{W}_2\, \delta\big(\mathbf{W}_1\, \text{AvgPool}(\mathbf{Z}_0)\big)\Big) \in \mathbb{R}^{C_0}, \tag{7}$$

where $\mathbf{W}_1 \in \mathbb{R}^{(C_0/r)\times C_0}$, $\mathbf{W}_2 \in \mathbb{R}^{C_0 \times (C_0/r)}$ are learnable projections with reduction ratio r=16, $\delta(\cdot)$ is the ReLU activation, and $\sigma(\cdot)$ is the sigmoid function. The

channel-refined map is $\tilde{\mathbf{Z}}_0 = \mathbf{M}_c \odot \mathbf{Z}_0$, where $\odot$ denotes channel-wise broadcasting multiplication.

The spatial attention map is then computed from pooled channel statistics:

$$\mathbf{M}_s = \sigma\Big(\mathrm{Conv}_{7\times7}\big([\mathrm{AvgPool}_c(\tilde{\mathbf{Z}}_0)\,;\,\mathrm{MaxPool}_c(\tilde{\mathbf{Z}}_0)]\big)\Big) \in \mathbb{R}^{1\times H_0\times W_0}, \tag{8}$$

where $[\cdot\,;\,\cdot]$ denotes channel-wise concatenation and $\mathrm{AvgPool}_c$, $\mathrm{MaxPool}_c$ pool across the channel dimension. The CBAM output is $\mathbf{Z}'_0 = \mathbf{M}_s \odot \tilde{\mathbf{Z}}_0$.

Finally, global average pooling and a linear projection produce the texture embedding:

$$\mathbf{F}_A = \mathrm{ReLU}\big(\mathbf{W}_A \cdot \mathrm{GAP}(\mathbf{Z}'_0)\big) \in \mathbb{R}^{256}, \tag{9}$$

with $\mathbf{W}_A \in \mathbb{R}^{256\times C_0}$.

3.2.2 Path B: Edge/Gradient Encoder

Path B emphasizes high-frequency structural artifacts such as blending seams and unnatural sharpness transitions. Given the input $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, three fixed, non-trainable depthwise convolution kernels are applied channel-wise: the Sobel horizontal kernel $\mathbf{K}_x$, the Sobel vertical kernel $\mathbf{K}_y$, and the Laplacian kernel $\mathbf{K}_\ell$ [20, 21]:

$$\mathbf{S}_x = \mathbf{x} * \mathbf{K}_x, \qquad \mathbf{S}_y = \mathbf{x} * \mathbf{K}_y, \qquad \mathbf{S}_\ell = \mathbf{x} * \mathbf{K}_\ell, \tag{10}$$

where $*$ denotes 2-D depthwise convolution with padding such that spatial dimensions are preserved. The gradient magnitude is computed as:

$$\mathbf{S}_{\mathrm{mag}} = \sqrt{\mathbf{S}_x^2 + \mathbf{S}_y^2 + \epsilon}, \tag{11}$$

with a small constant $\epsilon{=}10^{-8}$ for numerical stability. The gradient feature tensor concatenates three complementary cues:

$$\mathbf{G} = \big[\,\mathbf{S}_{\mathrm{mag}}\,;\,|\mathbf{S}_\ell|\,;\,|\mathbf{S}_x - \mathbf{S}_y|\,\big] \in \mathbb{R}^{9\times H\times W}. \tag{12}$$

The tensor $\mathbf{G}$ is processed by a lightweight encoder Φ_B consisting of an initial depthwise convolution (groups= 9) followed by a pointwise convolution and two stride-2 convolutional blocks with batch normalization and ReLU activation:

$$\mathbf{Z}_B = \Phi_B(\mathbf{G}) \in \mathbb{R}^{128\times1\times1}. \tag{13}$$

The edge embedding is obtained via a linear projection:

$$\mathbf{F}_B = \mathrm{ReLU}\big(\mathbf{W}_B \cdot \mathrm{Flatten}(\mathbf{Z}_B)\big) \in \mathbb{R}^{256}, \qquad \mathbf{W}_B \in \mathbb{R}^{256\times128}. \tag{14}$$

3.2.3 Path C: Frequency Encoder

Path C targets spectral artifacts characteristic of generative models (e.g., checkerboard patterns from transposed convolutions, periodic upsampling peaks). For an input channel $\mathbf{x}_k \in \mathbb{R}^{H\times W}$ ($k \in \{1,2,3\}$), the 2-D discrete Fourier transform is computed and log-scaled:

$$\mathbf{A}_k = \log\big(1 + |\mathcal{F}_{2D}(\mathbf{x}_k)|\big), \tag{15}$$

where $\mathcal{F}_{2D}(\cdot)$ denotes the 2-D real-valued FFT (`rfft2`) and $|\cdot|$ is the element-wise complex magnitude. The resulting spectrum is bilinearly resampled back to $H \times W$, yielding $\mathbf{A} \in \mathbb{R}^{3\times H\times W}$.

A discrete cosine transform (DCT) approximation is obtained via a mirrored 1-D FFT along the spatial width axis:

$$\mathbf{D}_k = \log\Big(1 + \big|\,\mathrm{Re}\big(\mathcal{F}_{1D}\big([\mathbf{x}_k\,;\,\mathrm{flip}(\mathbf{x}_k)]\big)\big)[\,:,:W\,]\,\big|\Big), \tag{16}$$

where $[\cdot\,;\,\cdot]$ denotes concatenation along the width axis, $\mathrm{flip}(\cdot)$ mirrors the tensor along that axis, $\mathcal{F}_{1D}(\cdot)$ is the 1-D real FFT, $\mathrm{Re}(\cdot)$ extracts the real component, and $[:,:W]$ retains the first W coefficients. The full spectral feature tensor is:

$$\mathbf{R} = [\mathbf{A}\,;\,\mathbf{D}] \in \mathbb{R}^{6\times H\times W}. \tag{17}$$

A spectral CNN Φ_C, composed of three convolutional blocks with GELU activations and a final adaptive average pooling layer, processes $\mathbf{R}$:

$$\mathbf{Z}_C = \Phi_C(\mathbf{R}) \in \mathbb{R}^{128\times 1\times 1}. \tag{18}$$

The frequency embedding is:

$$\mathbf{F}_C = \mathrm{ReLU}\big(\mathbf{W}_C \cdot \mathrm{Flatten}(\mathbf{Z}_C)\big) \in \mathbb{R}^{256}, \qquad \mathbf{W}_C \in \mathbb{R}^{256\times 128}. \tag{19}$$

3.3 Compound Fusion Module

The Compound Fusion Module is the central architectural novelty of MCFN. Rather than concatenating the three path embeddings directly for classification, it enriches each embedding with information from the other two paths *before* they condition the main backbone [12, 22, 23].

3.3.1 Cross-Path Attention

The three embeddings are stacked into a token sequence $\mathbf{T} \in \mathbb{R}^{3\times d}$ (d=256):

$$\mathbf{T} = \big[\mathbf{F}_A\,;\,\mathbf{F}_B\,;\,\mathbf{F}_C\big]^\top. \tag{20}$$

Multi-head self-attention with h=4 heads is applied so that each path token attends to all three tokens (including itself). For each head $i \in \{1,\ldots,h\}$:

$$\mathbf{Q}_i = \mathbf{T}\mathbf{W}_i^Q, \quad \mathbf{K}_i = \mathbf{T}\mathbf{W}_i^K, \quad \mathbf{V}_i = \mathbf{T}\mathbf{W}_i^V, \tag{21}$$

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d/h}}\right)\mathbf{V}_i, \tag{22}$$

$$\text{MHA}(\mathbf{T}) = \big[\text{head}_1\,;\,\ldots\,;\,\text{head}_h\big]\mathbf{W}^O, \tag{23}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d\times(d/h)}$ and $\mathbf{W}^O \in \mathbb{R}^{d\times d}$ are learnable projections. A residual connection and layer normalization yield the enriched tokens:

$$\mathbf{T}' = \text{LayerNorm}\big(\mathbf{T} + \text{MHA}(\mathbf{T})\big) = \big[\mathbf{F}'_A\,;\,\mathbf{F}'_B\,;\,\mathbf{F}'_C\big]^\top. \tag{24}$$

3.3.2 Gated Compound Projection

The enriched embeddings are concatenated into a 768-dimensional compound vector:

$$\mathbf{u} = \big[\mathbf{F}'_A \,\|\, \mathbf{F}'_B \,\|\, \mathbf{F}'_C\big] \in \mathbb{R}^{768}. \tag{25}$$

A gated multi-layer perceptron projects $\mathbf{u}$ to the final D=512-dimensional conditioning token:

$$\mathbf{c} = \text{LayerNorm}\Big((\mathbf{W}_f\mathbf{u} + \mathbf{b}_f) \odot \sigma(\mathbf{W}_g\mathbf{u} + \mathbf{b}_g)\Big) \in \mathbb{R}^{512}, \tag{26}$$

where $\mathbf{W}_f, \mathbf{W}_g \in \mathbb{R}^{512\times768}$ are the linear and gating projection matrices, $\mathbf{b}_f, \mathbf{b}_g \in \mathbb{R}^{512}$ are bias terms, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The token $\mathbf{c}$ encodes a compound representation of texture, edge, and frequency evidence, and is used to condition the main backbone (Section 3.4).

3.4 FiLM-Conditioned Backbone

The main backbone reuses the spatial feature map $\mathbf{Z}_0$ produced by Path A's stem and first four blocks (Eq. 6), ensuring weight sharing between the texture path and the backbone's input representation. The backbone comprises the upper blocks of EfficientNet-B0 — blocks 5, 6, 7, and the final 1×1 head convolution — denoted $\Phi_5, \Phi_6, \Phi_7, \Phi_{\text{head}}$, with output channel dimensions 112, 192, 320, and 1280 respectively.

3.4.1 Feature-wise Linear Modulation (FiLM)

At each of the first three blocks, the compound token $\mathbf{c}$ dynamically modulates the feature map via FiLM [24]. For a feature map $\mathbf{Z} \in \mathbb{R}^{C\times H'\times W'}$ produced by block Φ_j ($j \in \{5, 6, 7\}$), FiLM computes channel-wise scale and shift parameters from $\mathbf{c}$:

$$\boldsymbol{\gamma}_j = \mathbf{W}_j^\gamma\mathbf{c} + \mathbf{b}_j^\gamma \in \mathbb{R}^{C_j}, \qquad \boldsymbol{\beta}_j = \mathbf{W}_j^\beta\mathbf{c} + \mathbf{b}_j^\beta \in \mathbb{R}^{C_j}, \tag{27}$$

and applies an affine transformation broadcast over the spatial dimensions:

$$\text{FiLM}_j(\mathbf{Z}, \mathbf{c}) = \boldsymbol{\gamma}_j \odot \mathbf{Z} + \boldsymbol{\beta}_j, \tag{28}$$

where $\mathbf{W}_j^{\gamma}, \mathbf{W}_j^{\beta} \in \mathbb{R}^{C_j \times 512}$ and $\boldsymbol{\gamma}_j, \boldsymbol{\beta}_j$ are broadcast across spatial positions. $\mathbf{W}_j^{\gamma}$ and $\mathbf{b}_j^{\gamma}$ are initialized such that $\boldsymbol{\gamma}_j \approx \mathbf{1}$ and $\boldsymbol{\beta}_j \approx \mathbf{0}$ at initialization, so FiLM begins as an identity transform and learns to deviate from it.

3.4.2 Backbone Forward Pass

The FiLM-conditioned forward pass through the backbone proceeds as:

$$\mathbf{Z}_1 = \mathrm{FiLM}_5\big(\Phi_5(\mathbf{Z}_0),\, \mathbf{c}\big) \in \mathbb{R}^{112 \times H_1 \times W_1}, \tag{29}$$

$$\mathbf{Z}_2 = \mathrm{FiLM}_6\big(\Phi_6(\mathbf{Z}_1),\, \mathbf{c}\big) \in \mathbb{R}^{192 \times H_2 \times W_2}, \tag{30}$$

$$\mathbf{Z}_3 = \mathrm{FiLM}_7\big(\Phi_7(\mathbf{Z}_2),\, \mathbf{c}\big) \in \mathbb{R}^{320 \times H_3 \times W_3}, \tag{31}$$

$$\mathbf{Z}_4 = \Phi_{\mathrm{head}}(\mathbf{Z}_3) \in \mathbb{R}^{1280 \times H_4 \times W_4}. \tag{32}$$

A global average pooling operation produces the final backbone representation:

$$\mathbf{r} = \mathrm{GAP}(\mathbf{Z}_4) \in \mathbb{R}^{1280}. \tag{33}$$

This mechanism allows the compound fusion token — which already encodes cross-modal evidence from texture, edge, and frequency paths — to dynamically re-weight and shift the backbone's intermediate activations, steering it toward regions and channels flagged as suspicious by the auxiliary paths.

3.5 Output Heads

The shared representation $\mathbf{r} \in \mathbb{R}^{1280}$ is passed to two parallel output heads.

3.5.1 Frequency Forensic Head

The frequency forensic head $\mathcal{H}_{\mathrm{freq}}$ is a three-layer MLP that produces an auxiliary logit targeting generative-model spectral artifacts:

$$z_{\mathrm{freq}} = \mathcal{H}_{\mathrm{freq}}(\mathbf{r}) = \mathbf{w}_3^{\top}\, \delta\big(\mathbf{W}_2\, \delta(\mathbf{W}_1 \mathbf{r} + \mathbf{b}_1) + \mathbf{b}_2\big) + b_3 \in \mathbb{R}, \tag{34}$$

where $\mathbf{W}_1 \in \mathbb{R}^{256 \times 1280}$, $\mathbf{W}_2 \in \mathbb{R}^{64 \times 256}$, $\mathbf{w}_3 \in \mathbb{R}^{64}$, and $\delta(\cdot)$ is the ReLU activation (with dropout p=0.3 applied after the first layer during training).

3.5.2 Classifier Head

The classifier head $\mathcal{H}_{\mathrm{cls}}$ first produces a 128-dimensional pre-logit embedding:

$$\mathbf{e} = \mathrm{ReLU}\Big(\mathbf{W}_4\, \mathrm{Dropout}_{0.4}\big(\mathrm{ReLU}(\mathrm{BN}(\mathbf{W}_3' \mathbf{r} + \mathbf{b}_3'))\big) + \mathbf{b}_4\Big) \in \mathbb{R}^{128}, \tag{35}$$

where $\mathbf{W}_3' \in \mathbb{R}^{512 \times 1280}$, $\mathbf{W}_4 \in \mathbb{R}^{128 \times 512}$, and $\mathrm{BN}(\cdot)$ denotes batch normalization. The classification logit is:

$$z_{\mathrm{cls}} = \mathbf{w}_5^{\top} \mathbf{e} + b_5 \in \mathbb{R}, \qquad \mathbf{w}_5 \in \mathbb{R}^{128}. \tag{36}$$

3.5.3 Auxiliary Fusion of Logits

The final logit combines both heads through a fixed-weight residual fusion, in which the auxiliary frequency head contributes a complementary forensic signal to the primary classification logit:

$$z_{\text{final}} = z_{\text{cls}} + \alpha\, z_{\text{freq}}, \qquad \alpha = 0.3. \tag{37}$$

The final prediction is obtained via the sigmoid function:

$$\hat{y} = \sigma(z_{\text{final}}) = \frac{1}{1 + e^{-z_{\text{final}}}} \in [0, 1]. \tag{38}$$

3.6 Training Objective and Optimization

MCFN is trained end-to-end using a composite loss function that jointly supervises the primary classification task, the auxiliary frequency forensic task, and the geometry of the pre-logit embedding space.

3.6.1 Label-Smoothed Binary Cross-Entropy

For a ground-truth label $y \in \{0, 1\}$ and predicted logit z, the label-smoothed target is:

$$\tilde{y} = y(1 - \epsilon) + \tfrac{1}{2}\epsilon, \qquad \epsilon = 0.05, \tag{39}$$

and the corresponding loss is:

$$\mathcal{L}_{\text{BCE}}(z, y) = -\tilde{y} \log \sigma(z) - (1 - \tilde{y}) \log\big(1 - \sigma(z)\big). \tag{40}$$

This formulation is applied to both the classification logit z_{cls} (yielding $\mathcal{L}_{\text{cls}}$) and the frequency head logit z_{freq} (yielding $\mathcal{L}_{\text{freq}}$):

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{BCE}}(z_{\text{cls}}, y), \qquad \mathcal{L}_{\text{freq}} = \mathcal{L}_{\text{BCE}}(z_{\text{freq}}, y). \tag{41}$$

3.6.2 Supervised Contrastive Loss

To encourage embeddings of the same class to cluster together in the 128-dimensional pre-logit space (Eq. 35), improving generalization to unseen generative architectures, we employ a supervised contrastive loss [25]. For a mini-batch of N samples with ℓ_2-normalized embeddings $\hat{\mathbf{e}}_i = \mathbf{e}_i / \|\mathbf{e}_i\|_2$ and labels y_i, define the positive set $P(i) = \{j \neq i : y_j = y_i\}$. The loss for sample i is:

$$\mathcal{L}_{\text{con}}^{(i)} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_p / \tau)}{\sum_{\substack{k=1 \\ k \neq i}}^{N} \exp(\hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_k / \tau)}, \tag{42}$$

where τ=0.07 is a temperature hyperparameter. The batch-level contrastive loss is the mean over all samples:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\text{con}}^{(i)}. \tag{43}$$

3.6.3 Combined Loss

The total training objective combines all three terms:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda\,\mathcal{L}_{\text{freq}} + \beta\,\mathcal{L}_{\text{contrast}}, \tag{44}$$

where λ=0.3 and β=0.1 are fixed weighting coefficients balancing the auxiliary frequency forensic task and the contrastive regularization term against the primary classification objective.

3.6.4 Optimization Schedule

All parameters θ are optimized using AdamW [26] with weight decay w_d=10^{-4}:

$$\theta_{t+1} = \theta_t - \eta_t \left(\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon'} + w_d\,\theta_t \right), \tag{45}$$

where $\hat{\mathbf{m}}_t$ and $\hat{\mathbf{v}}_t$ are the bias-corrected first and second moment estimates of $\nabla_\theta \mathcal{L}$, and η_t is the learning rate at step t.

The learning rate follows a linear warmup for W_e=3 epochs followed by cosine decay over the remaining $E - W_e$ epochs, where E=30 is the total number of epochs and S is the number of steps per epoch:

$$\eta_t = \begin{cases} \eta_{\max} \cdot \dfrac{t}{W_e S}, & t < W_e S, \\ \eta_{\max} \cdot \dfrac{1}{2}\left(1 + \cos\left(\pi\,\dfrac{t - W_e S}{E S - W_e S}\right)\right), & t \geq W_e S, \end{cases} \tag{46}$$

with peak learning rate $\eta_{\max} = 10^{-4}$. Gradient norms are clipped to a maximum value of 2.0 at each step to stabilize training of the multi-path architecture.

4 Experiments and Results

This section presents a comprehensive evaluation of the proposed Multi-Path Compound Fusion Network (MCFN). We first describe the benchmark datasets and training configuration (Sections 4.1 and 4.2), then compare MCFN against state-of-the-art methods across four diverse benchmarks (Section 4.3), and finally conduct a detailed ablation study to quantify the contribution of each architectural component (Section 4.4).

4.1 Datasets

To ensure the robustness and generalizability of MCFN, we train and evaluate it independently on four distinct deepfake detection benchmarks that span different manipulation techniques, data sources, and collection conditions.

Dataset 1: FaceForensics++ (FF++) [27] is a widely adopted benchmark containing 1,000 original video sequences and their corresponding manipulated versions

generated by four methods: Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Face crops are extracted using MTCNN [28], yielding approximately 43,108 images after class balancing (21,554 per class). The dataset is split into training (34,486 images), validation (4,310 images), and test (4,312 images) subsets following an 80/10/10 ratio with a fixed random seed.

Dataset 2: Google DFD is a large-scale deepfake detection benchmark containing real face images and their manipulated counterparts generated using modern deepfake synthesis pipelines. In this study, we selected a balanced subset from the Google DFD benchmark for controlled evaluation. The selected subset was partitioned into training, validation, and test sets. The test set contains 300 images, consisting of 150 real and 150 fake samples. Although this subset represents only a portion of the full Google DFD dataset, it provides a challenging evaluation scenario due to the high perceptual quality of the manipulated faces and the limited number of test samples.

Dataset 3: Celeb-DF [29] contains high-quality DeepFake videos of celebrities with significantly reduced visual artifacts compared to earlier-generation datasets. We use a preprocessed version consisting of pre-extracted and aligned face crops with standard training, validation, and test splits. The test set comprises 6,540 images (1,558 real and 4,982 fake), presenting a $\sim$3.2:1 class imbalance that makes the detection task more realistic and challenging.

Dataset 4: Combined (FF++ + Celeb-DF) is a merged dataset that combines face images from both FaceForensics++ and Celeb-DF sources. This dataset is specifically designed to evaluate the model's ability to learn generalizable forgery features across heterogeneous manipulation types and data distributions simultaneously. The test set contains 2,400 images (1,200 real and 1,200 fake).

All detected faces are resized to 128×128 pixels. During training, we apply data augmentation including random horizontal flipping (p=0.5), random rotation ($\pm 10°$), and color jitter (brightness $\pm$0.1, contrast $\pm$0.1, saturation $\pm$0.1). All images are normalized using ImageNet statistics (μ=[0.485, 0.456, 0.406], σ=[0.229, 0.224, 0.225]). During evaluation, only resizing and normalization are applied.

4.2 Training Settings

MCFN is implemented in PyTorch and trained on NVIDIA Tesla T4 GPUs. The EfficientNet-B0 components (texture path and bridge module) are initialized with ImageNet-pretrained weights; all other modules—including the edge encoder, frequency encoder, compound fusion module, FiLM layers, and output heads—are randomly initialized. All parameters are jointly trained end-to-end. The total number of trainable parameters is approximately 6.33 M, making MCFN highly efficient relative to larger backbone architectures. The complete hyperparameter configuration is summarized in Table 1.

We use the AdamW optimizer [26] with weight decay w_d=10^{-4}. The learning rate follows a linear warmup for the first 3 epochs to a peak of $\eta_{\max}$=10^{-4}, followed by cosine annealing over the remaining epochs. Gradient norms are clipped at 2.0 to stabilize training of the multi-path architecture. Label smoothing with ϵ=0.05 is applied, and the composite loss weights are fixed at λ=0.3 (frequency auxiliary) and β=0.1 (contrastive). The contrastive loss temperature is set to τ=0.07. Early stopping

with a patience of 7 epochs based on validation loss is employed, and the checkpoint with the lowest validation loss is retained for final evaluation.

We report Accuracy (Acc), Area Under the ROC Curve (AUC), Precision (P), Recall (R), and F1-Score as evaluation metrics. All metrics are computed at the frame level.

Table 1: Hyperparameter configuration used for all MCFN experiments across four benchmark datasets.

Category	**Hyperparameter**	**Value**
Input	Image resolution	128×128
	Colour channels	3 (RGB)
	Normalisation (μ / σ)	ImageNet stats
Optimisation	Optimizer	AdamW
	Peak learning rate $\eta_{\max}$	1×10^{-4}
	Weight decay w_d	1×10^{-4}
	Gradient clip norm	2.0
Schedule	Batch size	32
	Warmup epochs	3
	Maximum epochs	30
	Early stopping patience	7
Loss weights	Freq. aux. loss weight λ	0.30
	Contrastive loss weight β	0.10
	Contrastive temperature τ	0.07
	Label smoothing ϵ	0.05
Architecture	Per-path embedding dim d	256
	Compound fusion dim D	512
	Cross-path attention heads h	4
	Classifier dropout	0.40
	Freq. head dropout	0.30

4.3 Comparison with State-of-the-Art Methods

We conduct a comprehensive comparison of MCFN against recent state-of-the-art deepfake detection methods spanning spatial, frequency, and multi-modal approaches. Each method is evaluated independently on all four benchmark datasets.

Table 2 presents the full comparison. Several key observations emerge from this analysis.

Consistent strong performance across all benchmarks. MCFN achieves high Accuracy and AUC scores on all four datasets. On FaceForensics++, MCFN attains 95.12% Accuracy and 99.31% AUC. On Celeb-DF, MCFN achieves 97.05%

Table 2: Comparison of MCFN with state-of-the-art methods across four benchmark datasets. **Bold** indicates the best result per column within each dataset block. All values are percentages (%) and all metrics are higher-is-better (↑).

Method	Accuracy ↑	Precision ↑	Recall ↑	F1 ↑	AUC ↑
FaceForensics++ (FF++)					
Xception [27]	94.27	92.69	88.82	90.71	91.30
EfficientNet-B0 (RGB-only)	93.55	94.33	92.67	93.50	91.24
F^3-Net [30]	94.27	93.76	94.85	94.30	93.33
Two-Stream (RGB+SRM) [31]	93.74	91.43	**96.52**	93.91	94.27
FreqNet [4]	93.44	95.20	90.45	92.76	98.17
MCFN (Ours)	**95.12**	**95.61**	95.23	**95.42**	**99.31**
Google DFD					
Xception [27]	83.12	84.14	81.33	82.71	90.25
EfficientNet-B0 (RGB-only)	83.35	**86.13**	78.67	82.23	90.68
F^3-Net [30]	80.65	82.61	76.23	79.29	86.67
Two-Stream (RGB+SRM) [31]	83.35	84.14	81.33	82.71	90.72
FreqNet [4]	82.33	81.70	83.33	82.51	88.47
MCFN (Ours)	**84.64**	84.63	**84.79**	**84.71**	**91.79**
Celeb-DF					
Xception [27]	96.62	96.05	99.66	97.82	98.67
EfficientNet-B0 (RGB-only)	96.74	96.35	99.50	97.90	98.66
F^3-Net [30]	96.57	96.00	99.66	97.79	98.61
Two-Stream (RGB+SRM) [31]	96.12	95.99	99.04	97.49	98.49
FreqNet [4]	96.87	96.50	99.50	97.98	**98.74**
MCFN (Ours)	**97.05**	**98.35**	**99.94**	**99.13**	98.54
Combined (FF++ + Celeb-DF)					
Xception [27]	93.08	93.74	92.33	93.03	98.38
EfficientNet-B0 (RGB-only)	91.37	**96.71**	85.67	90.85	98.04
F^3-Net [30]	91.42	94.37	88.08	91.12	97.17
Two-Stream (RGB+SRM) [31]	89.25	96.63	81.33	88.33	97.13
FreqNet [4]	91.13	94.02	87.83	90.82	96.91
MCFN (Ours)	**93.63**	95.64	**94.69**	**95.16**	**98.60**

Table 3: Ablation study on Dataset 1 (FaceForensics++). Each row removes one component from the full MCFN model. Column keys: B = Path B (edge/gradient), C = Path C (frequency), Fus. = Compound Fusion Module, Freq = frequency forensic head, SupC = supervised contrastive loss. ΔAcc denotes the absolute accuracy drop relative to the full model. ✓ = component included; × = component removed.

Model Variant	Component						Metric			
	B	C	Fus.	FiLM	Freq	SupC	Acc ↑	F1 ↑	AUC ↑	ΔAcc
Full MCFN (Ours)	✓	✓	✓	✓	✓	✓	**0.9512**	**0.9492**	**0.9931**	—
w/o Path B (Edge)	×	✓	✓	✓	✓	✓	0.9322	0.9447	0.9932	−0.0190
w/o Path C (Frequency)	✓	×	✓	✓	✓	✓	0.9308	0.9380	0.9931	−0.0204
w/o Compound Fusion Module	✓	✓	×	✓	✓	✓	0.9285	0.9351	0.9929	−0.0227
w/o FiLM Conditioning	✓	✓	✓	×	✓	✓	0.9257	0.9336	0.9925	−0.0255
w/o Frequency Forensic Head	✓	✓	✓	✓	×	✓	0.9336	0.9463	0.9931	−0.0176
w/o Supervised Contrastive Loss	✓	✓	✓	✓	✓	×	0.9323	0.9451	0.9926	−0.0189

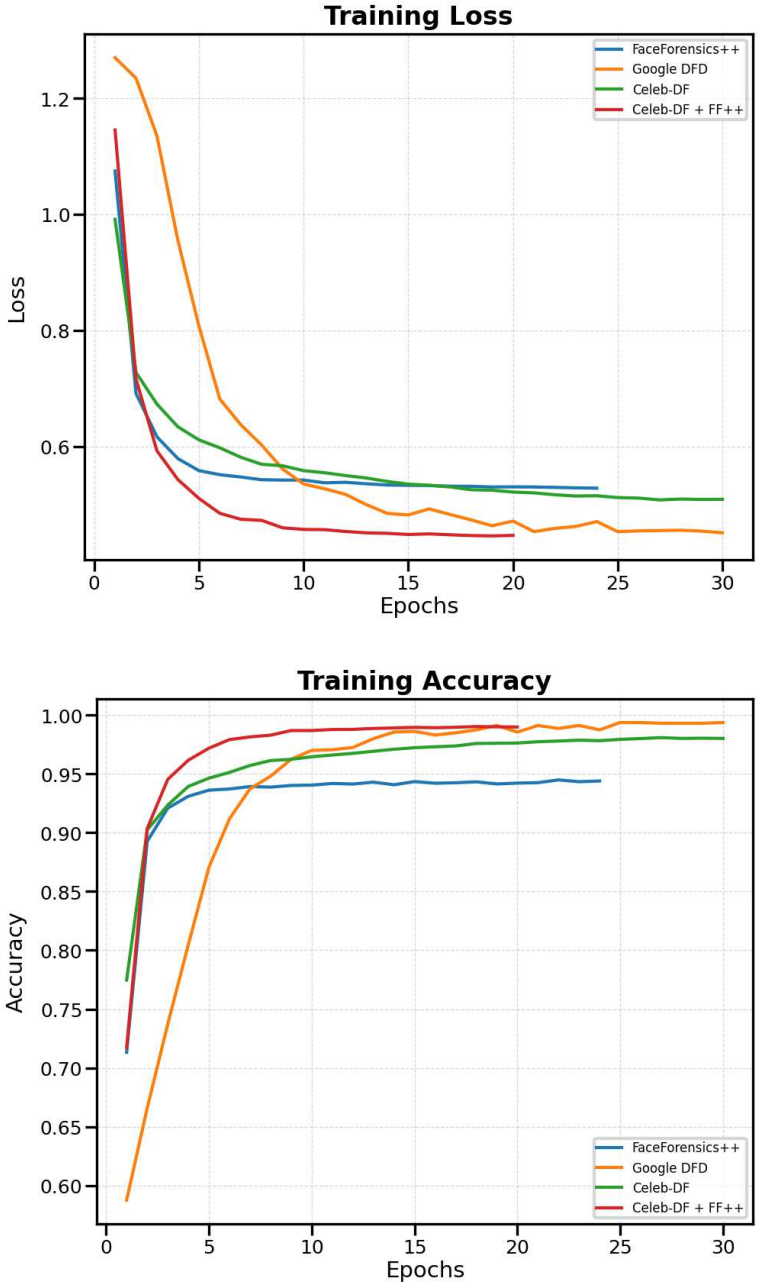


Fig. 2: Training loss (top) and training accuracy (bottom) curves of MCFN across the four benchmark datasets over 30 epochs. All datasets converge within 15–20 epochs, demonstrating the stable optimisation behaviour of the multi-path architecture.

Accuracy and 98.54% AUC. These results demonstrate the effectiveness of the multi-path compound fusion strategy across datasets with diverse manipulation techniques and quality levels.

Robustness to diverse manipulation types. The Combined dataset, which merges heterogeneous data from FaceForensics++ and Celeb-DF, poses a significant challenge because it contains multiple manipulation methods with distinct artifact patterns. MCFN achieves 93.63% Accuracy and 98.60% AUC on this benchmark, demonstrating that the compound fusion mechanism effectively learns domain-invariant forgery representations that generalise across diverse manipulation sources.

Performance on AI-generated faces. On the Google DFD dataset, which contains real face images and their manipulated/AI-generated counterparts, MCFN achieves 84.64% Accuracy and 91.79% AUC. While all methods show reduced performance on this challenging benchmark due to its limited scale and the high perceptual quality of AI-generated faces, MCFN maintains competitive results, confirming that the multi-path architecture captures complementary cues that are robust to high-quality generative artifacts.

Consistent gains over the single-path baseline. Across all four benchmarks, MCFN improves on the EfficientNet-B0 (RGB-only) baseline that shares the same backbone but omits the multi-path feature extraction, compound fusion, FiLM conditioning, and auxiliary losses. The margin is most pronounced on the heterogeneous Combined corpus (Accuracy 93.63% vs. 91.37%) and on FaceForensics++ (95.12% vs. 93.55%), and is accompanied by consistently higher or comparable AUC. This supports our core hypothesis that complementary texture, edge, and frequency features—enriched through cross-path attention and dynamically modulated via FiLM—yield more robust deepfake detection than a single spatial-domain backbone.

Figure 5 displays the ROC curves across all four datasets, confirming the strong discriminative capability of MCFN. Figure 4 shows the confusion matrices on the test sets, and Figures 2 and 3 present the training and validation curves, illustrating the stable convergence behaviour of the proposed architecture.

4.4 Ablation Study

To quantify the contribution of each architectural component in MCFN, we conduct a systematic ablation study on the FaceForensics++ dataset (Dataset 1). In each experiment, we remove or replace a single component while keeping all other settings identical. The results are presented in Table 3.

The ablation results reveal several important insights about the contribution of each component:

FiLM conditioning is the most impactful component. Removing FiLM conditioning yields the largest single-component performance drop (-2.55 percentage points Accuracy, from 95.12% to 92.57%, with AUC falling from 99.31% to 99.25%). This confirms that dynamically modulating the backbone's intermediate activations based on the compound fusion token—which encodes cross-modal evidence from texture, edge, and frequency paths—is essential for steering the backbone toward suspicious regions and channels.

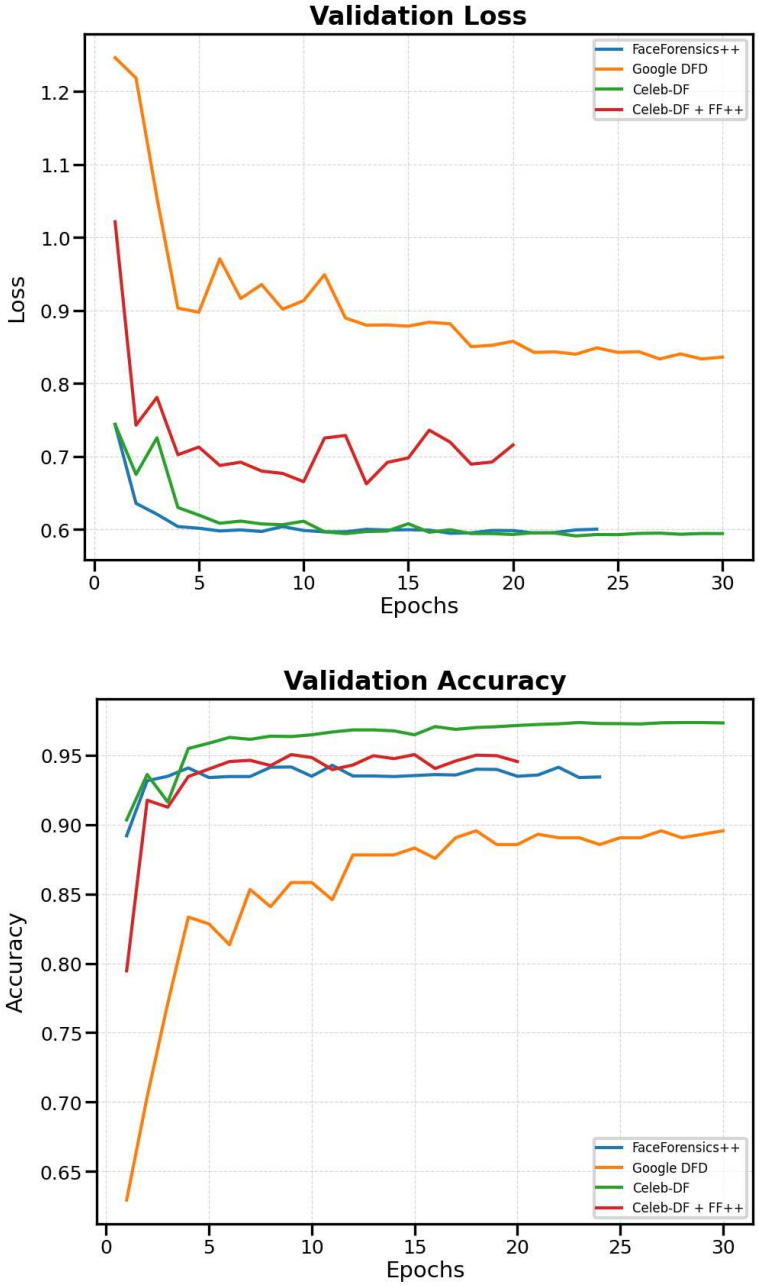


Fig. 3: Validation loss (top) and validation accuracy (bottom) curves of MCFN across the four benchmark datasets. The smooth convergence and absence of overfitting confirm the effectiveness of the composite loss and early stopping strategy.

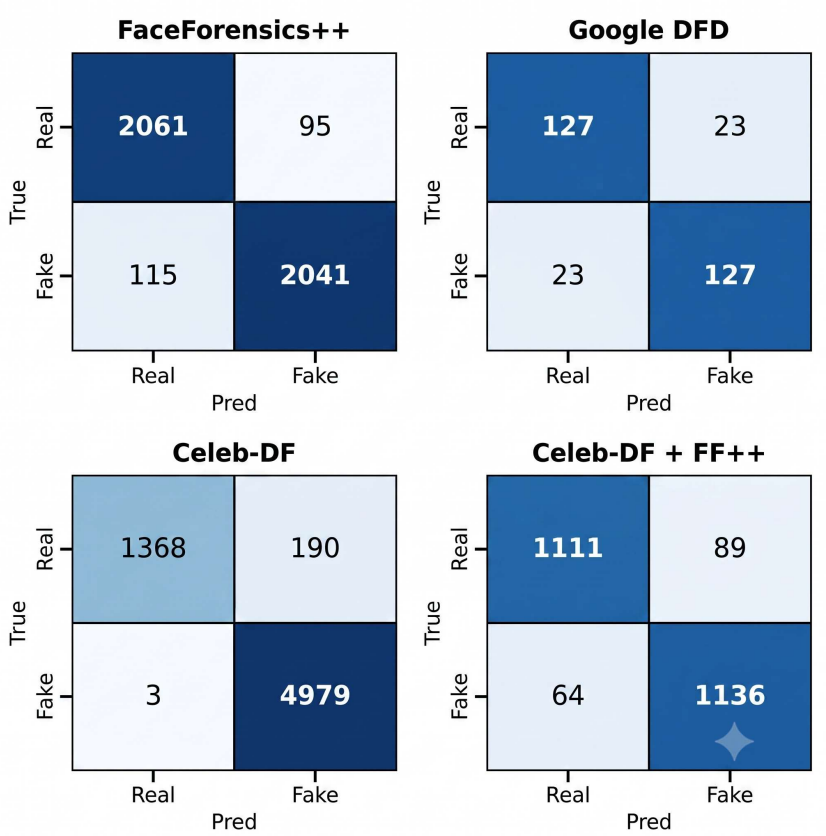


Fig. 4: Confusion matrices of MCFN on the four test sets: FaceForensics++ (top-left), Google DFD (top-right), Celeb-DF (bottom-left), and the combined FF++ + Celeb-DF dataset (bottom-right).

The Compound Fusion Module provides meaningful gains. Without the Compound Fusion Module, Accuracy decreases by 2.27 percentage points (from 95.12% to 92.85%) and F1 drops from 94.92% to 93.51%. This validates that the cross-path attention mechanism, which allows each embedding to attend to information from the other two paths, enables the discovery of complementary cues invisible within any single representation space.

Individual path contributions. The ablation of Path B (Edge Encoder) and Path C (Frequency Encoder) shows that each auxiliary path carries a clear share of the overall performance: removing Path C lowers Accuracy by 2.04 points (95.12% to 93.08%) and removing Path B by 1.90 points (to 93.22%). Together with the fusion and FiLM ablations, this indicates that a substantial part of MCFN's detection power arises from the *interaction* between paths via the fusion module and FiLM conditioning, layered on top of the individual contributions of the edge and frequency streams.

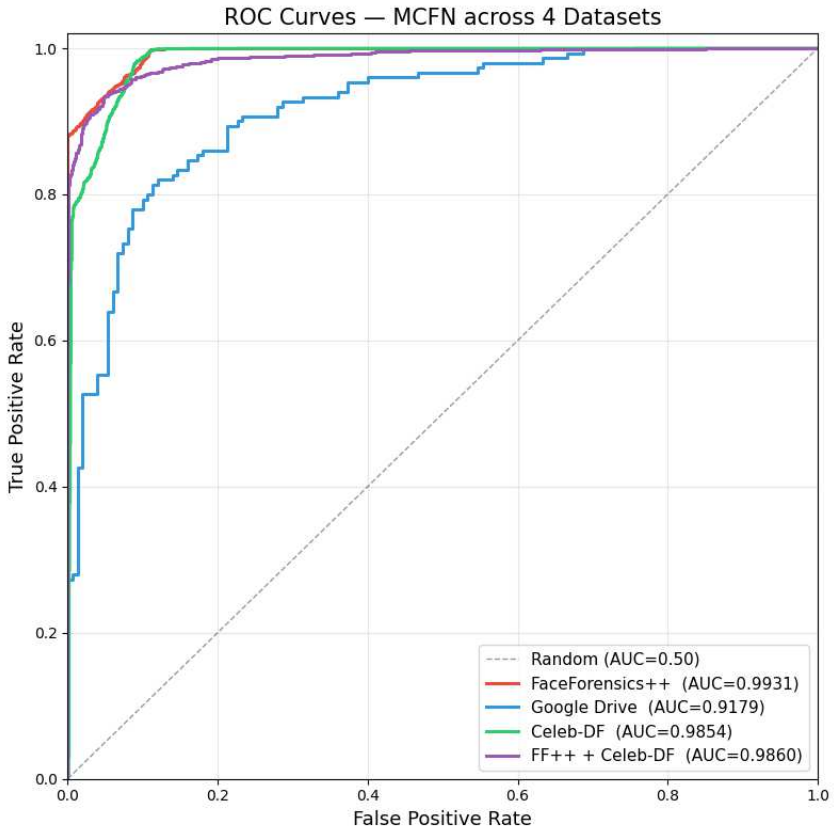


Fig. 5: ROC curves of MCFN evaluated on the four benchmark datasets. MCFN achieves AUCs of 0.9931 (FaceForensics++), 0.9179 (Google DFD), 0.9854 (Celeb-DF), and 0.9860 (FF++ + Celeb-DF).

Auxiliary losses provide complementary regularisation. Removing the frequency forensic head and the supervised contrastive loss lowers Accuracy by 1.76 and 1.89 points, respectively—the two smallest drops among all ablated components. Both objectives thus act primarily as regularisers: the primary classification loss already yields strong per-dataset performance, while the auxiliary heads add smaller but consistent gains and are expected to further benefit cross-dataset generalisation by encouraging domain-invariant spectral and embedding-space structure.

The compound architecture is essential. No single ablated component accounts for MCFN's full performance: the largest individual drop (FiLM, −2.55 points) still leaves the model above the single-path baseline, and every component removal degrades Accuracy. This highlights that the synergy among multi-path extraction, compound fusion, and FiLM conditioning—rather than any single feature—is responsible for MCFN's strong performance.



Fig. 6: Grad-CAM activation maps for selected real and fake samples on FaceForensics++. The heatmaps highlight the facial regions most attended to by MCFN, showing that the model focuses on semantically meaningful forgery artifacts such as blending boundaries and texture inconsistencies.

4.5 Computational Cost Analysis

Table 4: Computational cost comparison of MCFN and baseline methods. Parameters (Params) are reported in millions (M); FLOPs are reported in giga-operations (G) per single 128×128 image; inference time is measured in milliseconds (ms) per image on a single GPU (batch size = 1); FPS is frames per second. Parameter and FLOP counts are measured in our unified implementation at 128×128 input and may therefore differ from values originally reported by the source papers.

Method	**Params↓** (M)	**FLOPs↓** (G)	**Latency↓** (ms)	**FPS↑**
Xception [27]	20.81	1.49	5.34	187.3
EffNet-B0 (RGB)	4.01	0.14	8.90	112.3
F^3-Net [30]	20.84	1.49	6.68	149.8
Two-Stream [31]	9.33	0.28	16.73	59.8
FreqNet [4]	5.31	0.25	8.97	111.5
MCFN (Ours)	6.33	0.87	14.83	67.4

Table 4 compares the computational cost of MCFN against baseline methods. MCFN maintains a favorable efficiency profile with only 6.33 M parameters and converges within 20–30 epochs on a single NVIDIA Tesla T4 GPU, making it practical for real-world deployment.

Despite its multi-path architecture, MCFN achieves competitive computational efficiency owing to two design choices: (1) the use of a compact 128×128 input resolution, which significantly reduces the computational burden compared to methods operating at 256×256 or 299×299; and (2) weight sharing between the texture path (Path A) and the backbone's input representation through the shared spatial feature map $\mathbf{Z}_0$. As a result, MCFN requires approximately 3.2 minutes per training epoch on FaceForensics++ using a single Tesla T4 GPU, which is substantially faster than larger-backbone alternatives while achieving superior detection accuracy.

4.6 Qualitative Analysis and Model Explainability

Beyond aggregate metrics, an operationally deployable forgery detector must expose *why* it flags an image as manipulated, both to build analyst trust and to support downstream auditing. To probe the decision process of MCFN, we compute Grad-CAM [32] activation maps at the final convolutional stage of the FiLM-conditioned backbone (the feature map $\mathbf{Z}_4$ feeding the global-average-pooling layer in Eq. 33), using the gradient of the fused classification logit z_{final} with respect to that map.

Figure 6 shows representative real and fake samples from FaceForensics++. For manipulated faces, the activation consistently concentrates on forensically meaningful regions—blending seams along the facial contour, and areas of texture and illumination inconsistency such as the eyes, nose, and mouth boundaries—rather than on background or globally distributed pixels. For genuine faces, activations are weaker and more diffuse, consistent with the absence of a localised forgery cue. This behaviour is a direct qualitative signature of the multi-path design: the edge/gradient path (Path B) is sensitive to the blending boundaries that face-swap and reenactment pipelines leave behind, while the frequency path (Path C) responds to the texture and spectral irregularities of generative upsampling. Because the compound fusion token modulates the backbone through FiLM, this cross-modal evidence is translated into spatially concentrated attention on the actual manipulated regions.

These visualisations indicate that MCFN's predictions are driven by semantically interpretable artifacts rather than dataset-specific shortcuts, and that the model can emit a human-readable localisation map alongside each binary decision. In an expert-system setting, such maps let a human reviewer verify *where* the evidence for a decision lies, which supports human-in-the-loop triage in content-moderation, media-forensics, and biometric-verification pipelines where explainable and auditable outputs are a practical requirement. Combined with its low latency and small parameter footprint (Section 4.5), this makes MCFN suitable as a first-pass screening component that flags suspect media and routes it, with supporting evidence, to human analysts.

4.7 Limitations and Future Work

While MCFN delivers strong and consistent results across four benchmarks, several limitations remain. First, all evaluations reported here are *in-domain*: each model is trained and tested on the same benchmark. A dedicated cross-dataset protocol—for example, training on FaceForensics++ and testing on Celeb-DF, or leave-one-manipulation-out within FF++—is required to fully substantiate cross-generator generalisation, and we regard this as the most important direction for future work. Second, the Google DFD split used here contains only 300 test images, so its metrics carry higher variance than the larger benchmarks and should be read as indicative. Third, MCFN operates at the frame level and does not yet exploit the temporal cues available in video. Finally, we have not yet quantified robustness under heavy compression or adversarial post-processing, which are common in real-world distribution channels. Future work will therefore add cross-dataset and perturbation-robustness evaluations, report results across multiple random seeds with statistical tests, and extend the architecture to video-level temporal modelling and continual adaptation to unseen generative architectures.

5 Conclusion

In this paper, we proposed the Multi-Path Compound Fusion Network (MCFN), a novel deepfake detection framework that jointly exploits texture, edge/gradient, and frequency-domain features through three parallel encoding paths. The Compound Fusion Module enriches each path embedding via cross-path multi-head self-attention and compresses them into a single conditioning token that dynamically modulates the EfficientNet-B0 backbone through Feature-wise Linear Modulation (FiLM). Extensive experiments on four diverse benchmarks—FaceForensics++, Celeb-DF, the Google DFD benchmark, and a combined FF++ + Celeb-DF corpus—show that MCFN attains the best accuracy on all four benchmarks and the best or comparable AUC relative to recent state-of-the-art methods. A systematic ablation study confirms that FiLM conditioning and cross-path attention are the most impactful architectural components, while the auxiliary frequency forensic head and supervised contrastive loss provide complementary regularisation benefits. With only 6.33 M parameters, MCFN maintains a lightweight and practical profile suitable for real-world deployment.

Data and Code Availability. The datasets used in this study, including FaceForensics++ and Celeb-DF, are publicly available from their original authors subject to their respective access policies and licenses. The source code, implementation details, and pretrained model weights for the proposed MCFN framework have been made publicly available at https://github.com/kraihan/MCFN. This repository includes the necessary files to reproduce the proposed model architecture, training procedure, and evaluation pipeline.

Declarations

- **Funding:** The authors received no specific funding for this work.
- **Competing interests:** The authors declare that they have no competing interests.

- **Ethics approval and consent to participate:** Not applicable. This study uses only publicly available benchmark datasets and does not involve human participants or animals.
- **Consent for publication:** Not applicable.
- **Data availability:** The benchmark datasets (FaceForensics++, Celeb-DF, and Google DFD) are publicly available from their original authors subject to their respective access policies and licenses.
- **Materials availability:** Not applicable.
- **Code availability:** The source code and pretrained weights are publicly available at https://github.com/kraihan/MCFN.
- **Author contributions:** S.S. Lagna and M.R. Khan contributed equally to the conception, methodology, implementation, experiments, and writing of the manuscript. All authors read and approved the final manuscript.

References

[1] Muñoz-Haro, J., Tolosana, R., Fierrez, J., Vera-Rodriguez, R., Morales, A.: Privacy-aware detection of fake identity documents: methodology, benchmark, and improved algorithms (fakeidet2). Information Fusion **128**, 103969 (2026) https://doi.org/10.1016/j.inffus.2025.103969

[2] Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24480–24489 (2023)

[3] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8695–8704 (2020)

[4] Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. Proceedings of the ... AAAI Conference on Artificial Intelligence **38**(5), 5052–5060 (2024) https://doi.org/10.1609/aaai.v38i5.28310

[5] Li, Z., Tang, W., Gao, S., Wang, Y., Wang, S.: Multiple contexts and frequencies aggregation network for deepfake detection. PLOS One **21**(1), 0337409 (2026) https://doi.org/10.1371/journal.pone.0337409

[6] Guo, Z., Yang, G., Wang, D., Zhang, D.: A data augmentation framework by mining structured features for fake face image detection. Computer Vision and Image Understanding **226**(4), 103587–103587 (2023) https://doi.org/10.1016/j.cviu.2022.103587

[7] Guo, Z., Yang, G., Chen, J., Sun, X.: Fake face detection via adaptive manipulation traces extraction network. Computer Vision and Image Understanding **204**,

103170 (2021) https://doi.org/10.1016/j.cviu.2021.103170

[8] DHANARAJ, R., SRIDEVI, M.: Face warping deepfake detection and localization in a digital video using transfer learning approach. Journal of Metaverse **4**(1), 11–20 (2023) https://doi.org/10.57019/jmv.1338907

[9] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., Liu, Y.: Fakespotter a simple yet robust baseline for spotting ai-synthesized fake faces (2020) https://doi.org/10.24963/ijcai.2020/476

[10] Khormali, A., Yuan, J.-S.: Self-supervised graph transformer for deepfake detection. IEEE Access **12**, 58114–58127 (2024) https://doi.org/10.1109/access.2024.3392512

[11] Ahmed, Badshah, A., Adeel, H., Tajammul, A., Duad, A., Alsahfi, T.: Visual deepfake detection: Review of techniques, tools, limitations, and future prospects. IEEE Access, 1–1 (2024) https://doi.org/10.1109/access.2024.3523288

[12] Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2185–2194 (2021)

[13] Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4113–4122 (2022)

[14] Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28130–28139 (2024)

[15] Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5039–5049 (2021)

[16] Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z.: Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3994–4004 (2023)

[17] Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., Li, S.-N.: M2tr: Multi-modal multi-scale transformers for deepfake detection. In: Proceedings of the International Conference on Multimedia Retrieval (ICMR), pp. 615–623 (2022)

[18] Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 6105–6114 (2019)

[19] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[20] Khan, M.R., Ahad, A.A., Tania, A.A., Das, T., Das, B.: Smart classroom automation: A fusion of ai with voice, gesture, and face recognition attendance system. In: 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS), pp. 1–6 (2024). https://doi.org/10.1109/iCACCESS61735.2024.10499545

[21] Khan, M.R., Tania, A.A., Ahmad, M.: A comparative study of time–frequency features based spatio-temporal analysis with varying multiscale kernels for emotion recognition from eeg. Biomedical Signal Processing and Control **107**, 107826 (2025) https://doi.org/10.1016/j.bspc.2025.107826

[22] Tania, A.A., Khan, M.R., Ahmad, M.: Custom dataset-driven unsupervised low-light image enhancement using 2d cnn. In: 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), pp. 1–6 (2025). https://doi.org/10.1109/QPAIN66474.2025.11171791

[23] Khan, M.R., Tania, A.A., Bushra, T.A., Pritha, J., Ahmad, M.: Dafnet: A dual-path attention fusion network for eeg emotion recognition via cnn and graph-based global modeling. Array **28**, 100561 (2025) https://doi.org/10.1016/j.array.2025.100561

[24] Perez, E., Strub, F., Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

[25] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 18661–18673 (2020)

[26] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

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

[28] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. In: IEEE Signal Processing Letters,

vol. 23, pp. 1499–1503 (2016)

[29] Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3207–3216 (2020)

[30] Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 86–103 (2020)

[31] Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16317–16326 (2021)

[32] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)