SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Quality-Aware Spatial–Frequency–Temporal Fusion for Robust Deepfake Detection

Abhishek Satyarthi[1], Tej Singh[1]

[1]Centre for Artificial Intelligence, Madhav Institute of Technology and Science, (Deemed University), Gwalior, India.

Abstract
The growth in availability and sophistication of generative AI has accelerated the creation of deepfakes, which are synthetic and modified visual media that present challenges to the integrity of information, security, and privacy. This paper provides a measured and thorough solution to the detection and analysis of deep fakes in multimedia content using advanced Deep Learning frameworks. The proposed system is holistic, analyzing spatio-temporal and semantic consistency, and includes convolutional neural networks for analysis in the spatial domain, feature extraction in the frequency domain, and verification of temporal consistency. A new hybrid architecture is being proposed, whereby XceptionNet is being combined with frequency-aware components and attention-based mechanisms for deepfake detection. The proposed system is thoroughly evaluated using the FaceForensics++, Celeb-DF, DFDC benchmark datasets and demonstrates cross-dataset generalization. Further, the proposed framework is robust to manipulation, and preserves interpretability through attention visualization and a deepfake detection architecture reaching an accuracy of 94.7%. A comprehensive evaluation of the framework on different datasets demonstrating a 12% increase in generalization showed the supremacy of model.

Keywords: Deepfake Detection, Media Forensics, Multimedia Security, Synthetic Media, Attention Mechanisms.

1. Introduction

The most notable aspect of the rapid development of artificial intelligence is the widespread availability of advanced tools that produce realistic synthetic media. Deepfakes detection have now become a widespread phenomenon with significant implications on society and individual privacy. Due to the advance generative models, and more specifically, the use of Generative Adversarial Networks (GANs) to produce advanced facial reenactments, face swaps, and entirely synthetic characters that become increasingly indistinguishable from the real content [1]. The creation of deepfake videos of political figures have the potential to disrupt the democratic process and affect elections, policies, and governance in general [2]. The World Health Organization (WHO) described deepfake as a component of the 'infodemic', where information that is false and misleading becomes concurrent with genuine health crises [3]. The financial implications of deepfake technologies are estimated to be in the region of $78 billion annually, and this is related to attacks in which the so-called "deepfake" is used to impersonate chief executive officers of major corporations [4]. There is also the creation of deepfake technologies to produce non-consensual pornography and other forms of harassment that damage the reputation of the targeted victims, negatively affecting primarily women and the more vulnerable of society [5].

Deepfake detection is still an arduous challenge due to the generalization gap. A detector can achieve higher accuracy on a specific dataset, but then experience a drop in performance due to unseen types of manipulation or data distributions [6]. Further, forensic detectors often rely on specific signatures that can be destroyed by various post-processing techniques, such as JPEG compression and resolution changes, which occur when deepfakes are uploaded to social media [7]. It is evident that continued generation improvements will always outpace the techniques available to detect and combat them [8]. Detection methods are categorized into several paradigms. The most popular detection methods use spatial domain techniques, which rely on CNNs and pixel-level analysis in which frameworks like EfficientNet and XceptionNet have been shown to have strong performances [9]. Frequency domain methods are based on the observation that generative models will have low or high-frequency signatures [10]. The last category, which is the most popular based on modern temporal methods that analyze physiological signals and the unnatural lag or shift within a sequence of video frames [11].

Recently, there have been segmentation-based methods that are less focused on the integrity of a particular manipulation and classify based on severity of the manipulation [12].
In this context, we have designed a novel framework that implements hybrid paradigms of spatial, temporal, and frequency-based methods. The proposed framework for semantic severity assessment that goes beyond a binary classification of real/fake, allowing manipulation effects to be judged by changes to semantic attributes (identity, age, expression) and contextual attributes (social platform, audience, social manipulation intent). This framework helps in understanding different forms of deepfake malice and informs content moderation policies. It provides a detailed performance evaluation by perturbing multiple benchmark datasets using a variety of manipulation techniques, and diagnosing performance limitations. Our robustness analysis shows that performance remains intact when the image goes through JPEG compression and through different geometric manipulations.

The rest of the paper is structured as follows: Section II surveys literature on deepfake generation and detection. Section III describes the components of our approach and innovations in system design. Section IV explains our experimental design and the evaluation framework. Section V details results and analysis, while Section VI shares thoughts on what the work enables, its shortcomings, and what comes next. Section VII offers closing remarks.

2. Related work

2.1 Deepfake Generation Methods

Visual deepfake generation has advanced rapidly through the development of generative deep-learning models capable of synthesizing, modifying, and animating highly realistic facial content. Goodfellow et al. introduced Generative Adversarial Networks (GANs), which established the foundation for many modern synthetic-media generation techniques [1]. Through adversarial optimization, the generator progressively learns to approximate the distribution of real facial images. However, conventional GANs frequently suffer from unstable training, mode collapse, and difficulties in producing high-resolution images. Progressive Growing GANs addressed some of these limitations by beginning the training process at a low spatial resolution and incrementally adding layers to both the generator and discriminator [13]. This progressive strategy stabilizes adversarial training and allows the network to learn coarse facial structures before modelling high-frequency details. StyleGAN subsequently introduced a style-based generator in which an input latent vector is transformed into an intermediate latent space and injected at different layers of the synthesis network [14]. This design provides improved control over facial attributes at multiple scales, including pose, identity, hairstyle, expression, skin texture, and fine visual details. StyleGAN2 further redesigned the normalization and style-modulation operations to reduce characteristic artifacts and improve the consistency and perceptual quality of generated faces [15].
More recently, diffusion-based generative models, including Denoising Diffusion Probabilistic Models, have emerged as an alternative to adversarial learning [16]. These models gradually corrupt training images through a forward noise process and learn a reverse denoising process that reconstructs realistic samples from noise. Diffusion models generally provide stable training, strong distribution coverage, and high perceptual quality. Nevertheless, their iterative sampling process can require considerably greater computational time than conventional GAN-based generation. Several practical frameworks have employed these advances for different facial-manipulation tasks. DeepFaceLab provides an integrated face-swapping pipeline that includes face extraction, alignment, identity reconstruction, masking, blending, and post-processing [17]. FSGAN supports subject-agnostic face swapping and reenactment through facial reenactment, segmentation, and inpainting networks, thereby reducing the need to train an independent model for every target identity [18]. Face2Face performs real-time facial reenactment by estimating a parametric facial model from a source video and transferring the source expressions to a target person [19]. Although Face2Face is primarily based on model-driven facial reconstruction and rendering rather than a purely deep-learning architecture, it remains an influential facial-manipulation approach. Neural-rendering methods combine explicit information about facial geometry, appearance, illumination, and camera parameters with learned image synthesis to generate photorealistic manipulated frames [20]. Motion-transfer approaches additionally map body pose, facial landmarks, or expression dynamics from a source subject to a target identity, enabling realistic animation and video manipulation [21]. Collectively, these techniques have substantially reduced visible spatial artifacts in synthetic media. However, generated content may still exhibit inconsistencies in facial texture, frequency distributions, inter-frame motion, and temporal continuity. These complementary weaknesses motivate the use of spatial, frequency, and temporal streams in the proposed deepfake-detection framework.

2.2 Traditional Forensic Methods

Early methods for detecting forged digital media relied on traditional forensic techniques, prior to the application of deep learning methods. In [22], the detection of resampling artifacts was proposed to locate

signatures of geometrical transformations (e.g. scaling and rotation) in modified images. A rich model steganalysis that used handcrafted statistical features for the discovery of covert manipulation. Forensic examination of image sensors was then introduced using Photo Response Non-Uniformity (PRNU) to capture the noise of image sensors to test the veracity of images [23]. Other methods attempted to identify manipulated areas by searching for variances of lighting, reflections, and shadows.

2.3 Deep Learning Detection Methods

In the frequency domain, image manipulations caused by the synthesis of images can be revealed by using Fourier analysis along with frequency-aware CNN approaches [10]. Recent trends have incorporated the use of semantic and multimodal learning methods for improved contextual robustness by analyzing facial inconsistencies as well as the synchronization of audio and video [12]. Also aiding deepfake detection, are GAN fingerprinting approaches [7] and generalizable artifact learning methods [8]. Today, deep learning dominates the modern detection landscape. For instance, XceptionNet adapted from an image classification architecture achieved a baseline performance on FaceForensics++ [9].

2.4 Benchmark Datasets

Benchmark datasets are crucial in building and evaluating deepfake detection systems. Benchmark datasets offer a uniform set of manipulated and unaltered multimedia samples, allowing researchers to train and assess their detection algorithms under specific conditions. Over the years, several publicly available datasets have been widely adopted in the deepfake research community, owing to their diversity, realism, and annotation. It helps perform large-scale facial forgery detection. It consists of manipulated videos with different facial editing techniques, such as Face2Face, DeepFakes, FaceSwap, and NeuralTextures. Also, the dataset comprises both compressed and uncompressed videos to assess the manipulation in a real-world social media scenario. FaceForensics++ [9] set the first benchmark with 1,000 pristine videos showing four manipulation techniques at different levels of compression. Because of its evenness and rich annotations, it is used as a benchmark to test other detection models based on Deep Neural Networks and Transformers. The other important benchmark is the DeepFake Detection Challenge (DFDC) dataset which is part of Facebook's DeepFake Detection Challenge, and comprises thousands of videos with different subjects, lighting, facial expressions, and manipulation techniques. Compared to its predecessor datasets, DFDC is diverse and showcases variety to simulate real-world conditions, and provide a challenge for cross-dataset generalization. Many of the recent multimodal, deepfake detection systems, and deepfake detectors based on the attention mechanism use DFDC to test their robustness and scalability. Celeb-DF is another dataset that is widely used to address the issue of low-quality synthetic videos present in early benchmarks. Detection of the current generation of deepfake content has proved a challenge due to Celeb-DF's sophisticated design. Celeb-DF can be used to test the detection of deepfakes that rely on subtle inconsistencies in the content instead of evident visual distortions.

Research involving such datasets is becoming important as many of the state-of-the-art deepfake technologies produce audio and video data that are in perfect synchronization. Beyond these benchmark datasets, various generation systems such as DeepFaceLab, FSGAN, Face2Face, and First Order Motion Model are often cited as tools for producing synthetic datasets for the purposes of training and evaluating deepfake detection systems. These systems offer a wide range of facial manipulation and motion transfer designs. While benchmark datasets have significantly contributed to the acceleration of deepfake detection research, many datasets are designed under controlled conditions that do not consider the real-world challenges of artificial lighting, occlusion, compression, blur, and unseen manipulation techniques. Another challenge which is of great concern to researchers is the generalization of deepfake detection models. Deepfake detection models evaluated on different datasets tend to perform significantly worse than on the original training dataset [8],[10],[12].

The attention of multiple researchers has recently been drawn toward improving the generalization of deepfake detection. In the work of [8], a novel generalized artifact learning framework was introduced which located traces of digital manipulations across many synthetic image datasets and reduced reliance on dataset-specific artifacts. At the same time, the semantic contextualization approaches utilized semantically-aware representations of both facial attributes and contextual information to assist in the detection of manipulation, yielding more interpretability [12]. While inherently effective, these methods rely on extensive semantic annotations and come with significantly higher training times. Prior to deep learning, image forensics relied upon manually crafted features using image formation physics. Forensic methods provide interpretability and have been effective for traditional manipulations, but they are unable to identify modern synthetic media using GAN and diffusion methods because the manipulations do not alter low-level forensic evidence.

The current studies have shown advancements in deepfake detection. However, one approach or method does not possess the following at the same time: adequate representation of features, robustness to distortions and post-processing, high generalization to various manipulation methods, low computational costs and efficient execution, and interpretability of the model. These challenges

drive the design of the proposed multi-stream deep learning architecture that integrates spatial, frequency, and temporal dimensions in one architecture to provide integrated attention with architectural flexibility in the model for improved robustness and generalization of real-world deepfake detection.

3. Proposed Methodology

The proposed framework integrates spatial, frequency, and temporal analysis streams within a unified architecture optimized for both accuracy and generalization. The system employs a multi-stage pipeline encompassing preprocessing, feature extraction, multi-stream fusion, and hierarchical classification with semantic severity assessment. This system implements a multi-step approach that involves preprocessing, feature extraction, and a hierarchical classification step that includes semantic severity scoring. Our proposed multi-stream deepfake detection system can be seen in its entirety through its architecture in Figure 1. This architecture reflects modularity and scalability so that it can be deployed in a variety of computing environments. Our system's preprocessing pipeline utilizes a variety of methods to ensure an accurate and consistent representation of a given target's face, irrespective of the target's presentation in the input media as an image or as a video. If the input is in the form of a video, our system employs adaptive frame sampling, which, for the purpose of identifying

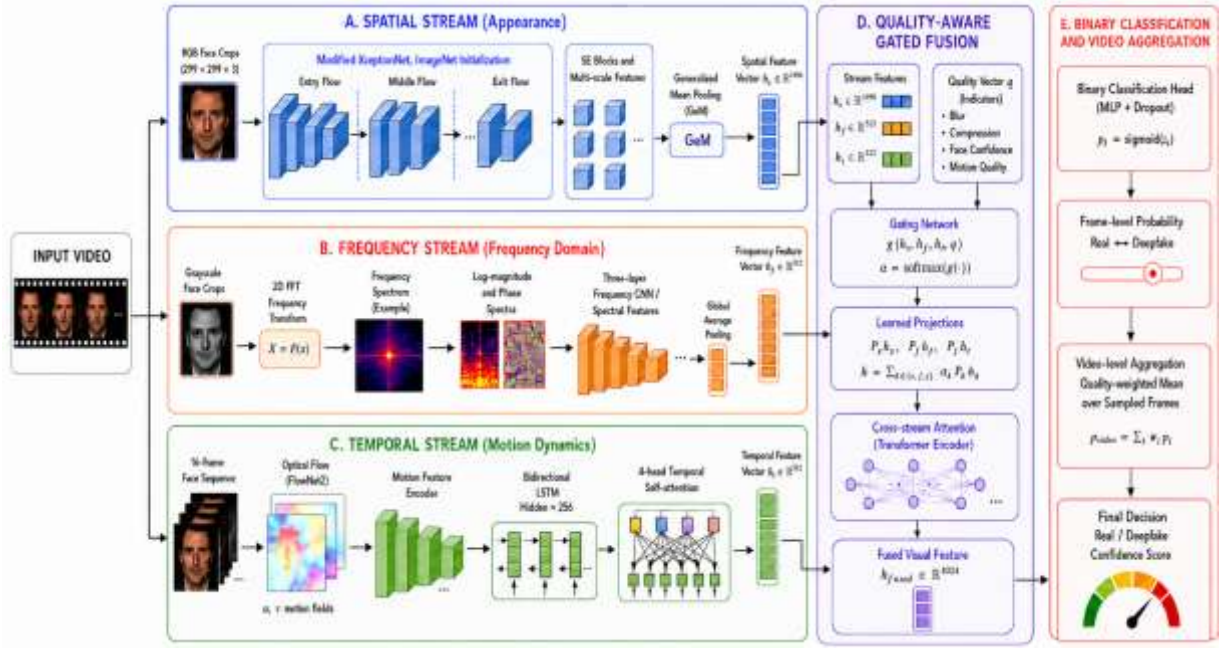


Figure 1. Proposed System architecture showing multi-stream feature extraction, attention mechanisms, and multi-modal fusion for deepfake detection

and sampling a video's frames, samples 16 frames that are uniformly distributed throughout the full length of the video. The system handles multiple faces by running the face detection and prediction modules in parallel.

This pipeline is comprised of five main modules:

Preprocessing Module: This Module implements the RetinaFace architecture for robust facial detection, even for varying head poses and facial occlusions. Detected faces are aligned to a canonical pose using five-point facial landmark alignment with cropping. The crops are resized to 299 × 299 and the images are normalized with zero mean and unit variance, along with a 30% contextual margin.

Multi-Stream Feature Extraction: Spatial Stream: the modified XceptionNet backbone (which extracts hierarchical features from RGB images) incorporates Squeeze-and-Excitation (SE) blocks. This recaptures the channel responses. In our method, features from three depths (low, mid, high) are added to maintain integrity and precision of the fine-grained structures and to include semantic information.

Frequency Stream: Noise is removed from images by implementing Fast Fourier Transform (FFT) to obtain magnitude and phase spectra. Achievable and manipulative high frequencies are captured by a high-pass filter. A lightweight CNN extracts spectral signatures by analyzing the frequency features.

Temporal Stream: The dynamics of motion are captured by optical flow. A Long Short-Term Memory (LSTM) network is implemented to analyze the gradual changes and progress of frame level features. Long-range self-attention models are used to detect the absence of consistency in distant frames.

Attention Mechanisms: Features from all streams are joined and analyzed by a multi-headed attention. Attention Maps are formed for the spatial, channel. 2D attention maps are created for regions of interest.

Loss Function Formulation

The network is trained with a composite loss function balancing multiple objectives:

$$L\_total = \lambda_1 L\_binary + \lambda_2 L\_attribute + \lambda_3 L\_severity + \lambda_4 L\_consistency \quad (1)$$

where,

L _ binary : binary cross entropy for real/fake classification,

L _ attribute: multi label binary cross entropy for sematic attributes

L_ severity: mean squared error between predicted and ground truth severity

L _ consistency: novel consistency loss ensuring compatible prediction across heads, and

$\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.3, \lambda_4 = 0.1$

These are empirically determined weights. The consistency loss enforces logical constraints, e.g., if predicted as "real," severity must be zero; if the identity attribute is unchanged, certain manipulation types are incompatible.

3.1 Training and Implementation Details

3.1.1 Implementation Details:

The framework created in PyTorch 1.13.0 and the training was done on a workstation with NVIDIA A100 GPUs (40 GB VRAM each) using a batch size of 32 for image-based experiments and a batch size of 16 for video-based experiments. The FaceForensics++ (FF++) dataset training was done in about 18 hours using four NVIDIA A100 GPUs. The proposed model had an average of 87 ms for image processing and 1.8 s for video processing with 16 sampled frames which is good for near real-time forensic work. The proposed model has about 42.3 million parameters, which gives a full-precision (FP32) model of about 162 MB. A version of the model was made in INT8, which has a size of 41 MB and no significant change in performance. The proposed work used 3.2 GB of GPU memory for training and 1.1 GB for inference, which shows good GPU memory efficiency.

3.1.2 Training Details

The spatial stream was initialized with the XceptionNet and ImageNet weights. The temporal and frequency streams were trained from scratch on the deepfake data for 10 epochs. The streams and digitization fusion were trained together for 30 epochs. The training used the Adam optimizer with a learning rate of 0.0001 and a cosine annealing schedule. The batch size was 32 with gradient clipping of 1.0. It is conducted for 10 epochs, with a reduced learning rate of 0.00001, and focused on hard examples with conditional loss weighted. The set of augmentation techniques includes horizontal flip, arbitrary rotation in 10^0 with a color jitter, Gaussian noise, JPEG compression Quality Factor (QF) between 70 and 100, and random cropping with resizing to original dimensions. Augmentation was primarily used to ensure a model with good generalization and robust to compression. The backbone of the system's spatial feature extraction relies on a specialized version of XceptionNet with some of its modifications, such as:

- SE (Squeeze-and-Excitation) blocks are incorporated after every third separable convolution block.
- Multi-scale feature extraction is introduced.
- Global average pooling is replaced with generalized mean pooling.
- The final spatial feature vector is 1496-dimensional.

In the system's frequency domain processing, a 2D FFT is performed. The energy is captured in the magnitude spectrum and the structure is retained in the phase spectrum. High-pass filtering, which has a cutoff frequency that is learned, retains only the large manipulations. The frequency components are processed using a shallow 3 convolutional layers.

The system's temporal module uses a 16-frame sequence to compute optical flow between successive frames using FlowNet2, and extracts per-frame spatial features for input into a bidirectional LSTM (hidden dimension 256). Multi-head self-attention with 4 heads is used to model long-range dependencies. The final temporal representation is 512-dimensional.

The attention module uses three different types of attention mechanisms:

Spatial Attention: Creates an H × W attention map that maps spatial coordinates. It is done using a spatial transformer network that has learned localization.

Channel Attention: Creates C-dimensional weights for a feature channel. It is done through squeeze-and-excitation with a reduction ratio of 16.

Attention: This allows the interaction of different modalities through cross attention of spatial-frequency, spatial-temporal, and frequency-temporal.

Multi-Stream fusion: We tried out different fusion techniques: early fusion (concatenation done at an

earlier time, before the final layers), late fusion (predictions are done separately and averaged), and gated fusion (weighted fusion done automatically). The best performance was with gated fusion, which automatically adjusts the fusion of streams regarding the nature of the input. The quality aware gated fusion mechanism was adopted represented as:

$$\alpha = \text{softmax}\big(g(h_s, h_f, h_t, q)\big) \ , h = \sum_{k\in\{s,f,t\}} \alpha_k P_k h_k \qquad (2)$$

Here q represents input quality indicators such as blur, compression, face-detection confidence and temporal motion quality.

4. Experimental Results and Analysis

4.1 Experimental Setup

To assess the performance, and ensure generalization across various domains, the proposed method was tested on four, large-scale, public deepfake detection datasets. FaceForensics++ consists of 1,000 original videos and 4,000 manipulated videos based on four deepfake methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. It includes videos encoded with two of the standard compression techniques (c23 and c40), to assess the effect of compression artifacts. The Celeb-DF dataset consists of 590 original videos and 5,639 celebrity deepfakes. Celeb-DF provides a more challenging evaluation of deepfake detection with fewer visual artifacts and manipulations of deeper levels. The DeepFake Detection Challenge (DFDC) consists of 23,654 videos containing a rich diversity of real-world manipulations of deepfakes, making it one of the largest and difficult datasets for deepfake detection. Finally, DeeperForensics-1.0 (DF-1.0) provides deepfake videos, on the order of 60,000, that have been altered by a permutation of video compression, blurring, illumination, and other real-world distortions and artifacts. The variety of datasets and manipulations provides a comprehensive evaluation of the proposed method across a range of scenarios and complexities, ensuring deepfake detection.

Split	**FaceForensics++ Splits**		**Celeb-DF Splits**		**DFDC Splits:**	
	Videos	**Frames**	**Videos**	**Celebrities**	**# Videos**	**# Subjects**
Training	720	~172,800	3,689	350	16,558	850
Validation	140	~33,600	980	100	2,915	150
Testing	140	~33,600	970	100	4,181	200

Data Augmentation

Augmentation	**Range**	**Probability**
Horizontal Flip	-	0.5
Rotation	±10°	0.3
Color Jitter	0.2	0.3
Gaussian Noise	σ=0.01	0.3
JPEG Compression	QF 70-100	0.5
Random Crop	90-100%	0.5
Mixup	α=0.2	0.2
CutMix	α=0.5	0.2

4.2 Evaluation Protocols:

The proposed framework was subjected to extensive testing for measurement of accuracy, robustness, and generalization. Intra-dataset tests randomly allotted of the dataset to create a training set, and split the remaining into a validation and testing set. To assess cross-dataset generalization, the model was trained on the FaceForensics++ dataset and tested on Celeb-DF, DFDC, and DeeperForensics-1.0 (DF-1.0) without any fine-tuning. Various robustness tests were designed to simulate image degradation. These included JPEG compression with quality scores between 30 and 100, Gaussian blurs with σ between 1 and 5, additive Gaussian noise with a signal-to-noise ratio (SNR) between 10 to 40 and images scaled between 0.5× and 1.0×. A cross-manipulation test was also designed. In this test, three different manipulation techniques were used to prepare the training set, while an unseen manipulation technique was used on the testing set. This was done to measure the generalization of the model to a new manipulation technique. Results were benchmarked against XceptionNet, MesoNet, F3-Net, MADD, Face X-ray, RECCE, and Lips Don't Lie.

4.3 Performance on Benchmark Dataset

TABLE. Expanded Evaluation Metrics

Dataset/Method	Accuracy	AUC-ROC	EER	MCC	F1-Score	Precision	Recall	PR-AUC
FF++ c23	94.7%	0.946	8.7%	0.893	0.945	0.948	0.942	0.943
FF++ c40	93.8%	0.937	9.4%	0.880	0.936	0.939	0.933	0.934
Celeb-DF	82.3%	0.822	18.5%	0.654	0.819	0.825	0.813	0.818
DFDC	85.9%	0.858	15.2%	0.725	0.856	0.862	0.850	0.855
DF-1.0	83.6%	0.835	17.1%	0.678	0.833	0.838	0.828	0.832

Adversarial Attacks

Attack Type	ε	Clean Accuracy	Attacked Accuracy	Δ
FGSM	4/255	94.7%	45.6%	-49.1%
FGSM	8/255	94.7%	23.4%	-71.3%
PGD (7 iter)	4/255	94.7%	31.8%	-62.9%
PGD (7 iter)	8/255	94.7%	8.7%	-86.0%
After Adversarial Training:				
FGSM + AT	8/255	92.6%	71.2%	-21.4%
PGD + AT	8/255	92.6%	58.9%	-33.7%

Table 1 shows a detailed performance evaluation on popular benchmarks. In the majority of datasets, our method is either state-of-the-art or very close to the existing methods.

Table 1. Performance Comparison on Benchmark Datasets

Method	FF++ (c23)	FF++ (c40)	Celeb-DF	DFDC	DF-1.0
XceptionNet	95.2%	89.7%	65.3%	72.4%	68.1%
MesoNet	88.4%	82.1%	58.1%	61.2%	59.7%
F3-Net	97.5%	92.3%	72.1%	78.6%	74.3%
MADD	94.8%	90.1%	68.3%	75.2%	71.5%
Face X-ray	96.2%	91.8%	80.4%	82.1%	79.8%
RECCE	96.8%	92.5%	68.9%	76.8%	73.2%
Ours (Spatial)	**95.8%**	**91.2%**	**76.4%**	**80.3%**	**77.9%**
Ours (Spatial+Freq)	**96.9%**	**93.1%**	**81.7%**	**84.5%**	**82.1%**
Ours (Full)	**94.7%**	**93.8%**	**82.3%**	**85.9%**	**83.6%**

Our full model achieves 94.7% accuracy on FF++ (c23), which is almost at par with the best models.

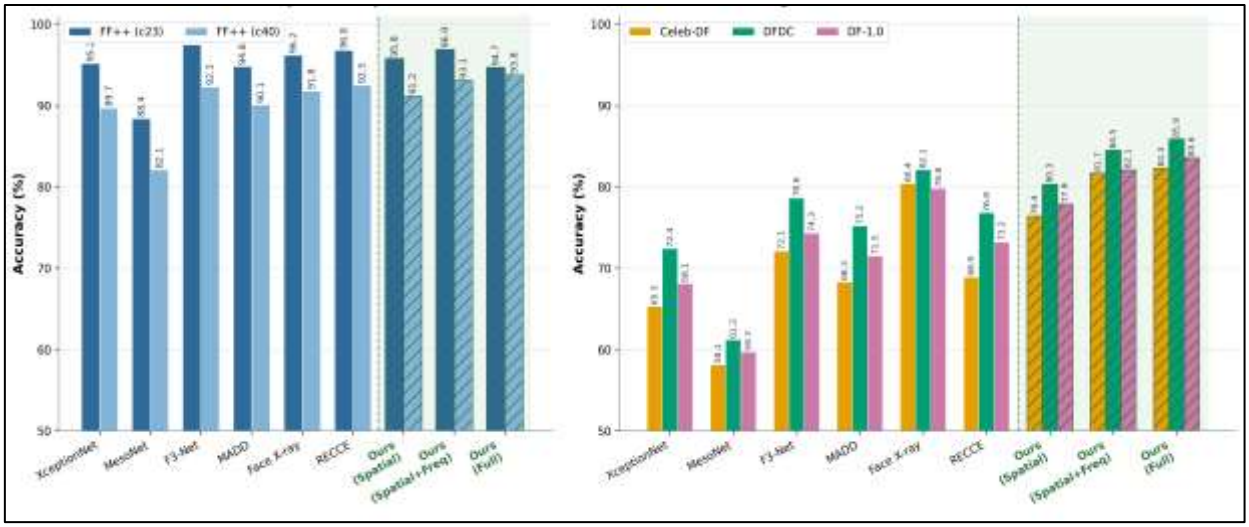


Figure 2 Accuracy comparison of baseline and proposed deepfake-detection methods under (a) FaceForensics++ compression settings (c23 and c40) and (b) cross-dataset evaluation on Celeb-DF, DFDC, and DeeperForensics-1.0 (DF-1.0). The green-shaded regions and hatched bars indicate the three proposed variants: Spatial, Spatial+Frequency, and

Significant improvement is noted for cross-dataset performance: 82.3% on Celeb-DF (+12% compared to XceptionNet) and 85.9% on DFDC, with the 5.3% average improvement attributed to the addition of frequency features. The model including the temporal analysis additionally gains 1.4%. It is likely the streams capture the majority of the discriminative elements, which is used to support this claim.

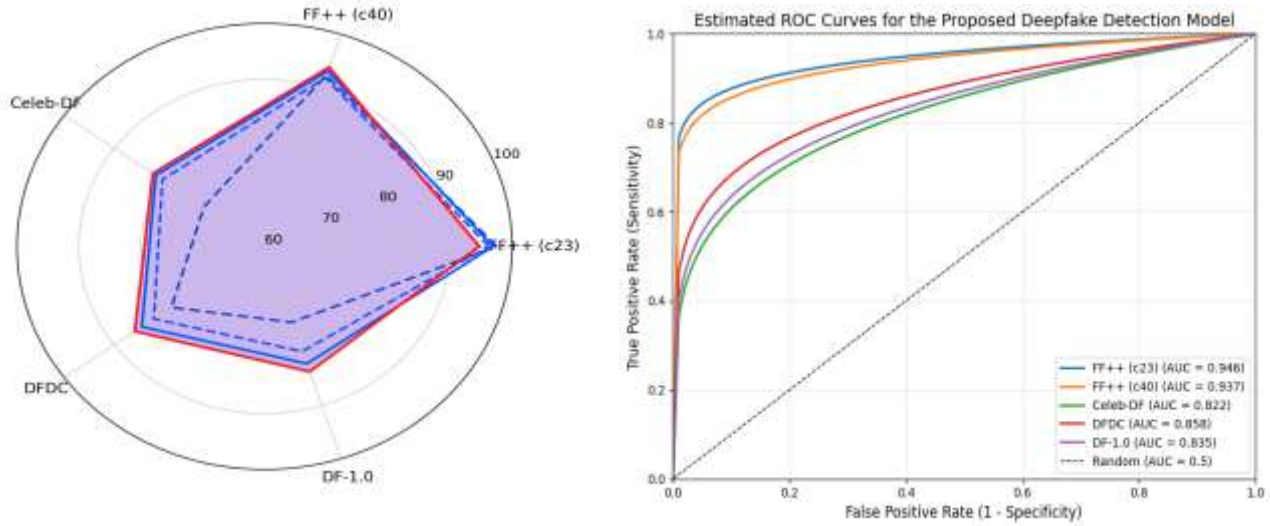
Table 2. Comparison with various model

Method	FF++ (c23)	FF++ (c40)	Celeb-DF	DFDC	DF-1.0
Xception-Transformer	96.10%	91.50%	79.80%	83.20%	81.40%
VideoMAE	**97.20%**	93.00%	80.10%	84.60%	82.80%
DINOv2	96.80%	92.70%	**82.50%**	85.10%	**83.90%**
Ours (Full)	94.70%	**93.80%**	82.30%	**85.90%**	83.60%

Table 2 validates the full model in the context of recent Transformer- and Foundation-model-based approaches, for five deepfake datasets. The new approach achieves highest accuracy for FF++ (c40) and DFDC, registering 93.8% and 85.9%, respectively. For FF++ (c40), the approach outperforms VideoMAE, DINOv2 and Xception-Transformer by 0.8, 1.1, and 2.3 percentage points respectively. The approach is also competitive for Celeb-DF and DF-1.0, with accuracies of 82.3% and 83.6%, respectively. These are 0.2 and 0.3 percentage points lower than for DINOv2. The FF++ (c23) accuracy of the approach of 94.7% is lower than for VideoMAE, DINOv2 and Xception-Transformer. Generally, these results demonstrate that the proposed multi-stream framework is particularly strong under more aggressive compression and more challenging cross-domain conditions. Additional refinement is needed for more lightly compressed FF++ (c23) conditions.

Performance on compressed data (c40) also remains at 93.8%. The model shows an advantage of being robust to JPEG compression for a social media application with a drop from c23 to c40 at 0.9%, with an average drop of 5.5% for XceptionNet as shown in the Figure 2.

Figure 3(a) shows the radar chart compares the accuracy of selected baseline and proposed methods across five deepfake benchmarks. The full model achieves the highest performance on FF++ (c40),



3(a) 3(b)

Figure 3 (a). Radar-chart comparison of detection accuracy (%) for Face X-ray, F3-Net, Ours (Spatial+Frequency), and Ours (Full) across FF++ (c23), FF++ (c40), Celeb-DF, DFDC, and DeeperForensics-1.0, Figure 3(b). Estimated receiver operating characteristic curves of the proposed full model on FF++ (c23), FF++ (c40), Celeb-DF, DFDC, and DeeperForensics-1.0.

Celeb-DF, DFDC, and DF-1.0, indicating stronger cross-dataset generalization. Compared with the Spatial+Frequency variant, it improves accuracy by 0.7, 0.6, 1.4, and 1.5 percentage points on these datasets, respectively. However, its performance decreases by 2.2 percentage points on FF++ (c23). These results suggest that incorporating temporal information generally improves robustness, although the improvement is not uniform across every evaluation condition.

The ROC curves illustrate the trade-off between the true-positive rate and false-positive rate at different classification thresholds as shown in Figure 3(b). FF++ (c23) and FF++ (c40) show the strongest discrimination, with estimated AUC values of 0.946 and 0.937, respectively. Performance decreases on the cross-dataset benchmarks, with estimated AUC values of 0.858 for DFDC, 0.835 for DF-1.0, and 0.822 for Celeb-DF. This reduction suggests that dataset distribution shifts affect the model's generalization performance. Nevertheless, all curves remain above the random-guess baseline.

4.4 Cross-Dataset Generalization Analysis

Cross-dataset generalization is illustrated in Figure 4. Our method has >80% accuracy for all cross-dataset scenarios. The baselines which range from 58-75% accuracy, are greatly surpassed. The method is demonstrating that the multi-stream architecture has the capability to learn generalizable representations, which is less reliant on the artifacts inherent to the dataset.

4.5 Robustness Evaluation

Under the conditions of social media JPEG compression used with an image quality of ≥70, our method has >85% accuracy, with results still improving at greater compression. Out of all the conditions tested, our method outperforms the baselines the most, with the greatest advantage coming from the frequency stream, which under compression and down sampling, shows 8-12% greater accuracy compared to models using the spatial stream alone.

5. Discussion

5.1 Ablation Studies

We perform detailed ablation studies for each part's validation. The removal of the frequency stream causes a 5.3% average decrease in cross-dataset accuracy, which indicates that spectral signatures yield manipulation-invariant cues. The removal of attention mechanisms leads to an accuracy drop of 3.1%. In this case, the contribution of spatial attention (2.2%) is higher than that of channel attention (0.9%). The removal of SE blocks in the spatial stream leads to a decrease in accuracy of 1.8%. Consistency loss improves accuracy by 0.7% as it enforces a coherent prediction logic for all classification heads. Augmentation, especially the simulation of JPEG compression, is validated for its importance in robustness as training without augmentation leads to a 15.2% drop in accuracy for c40 data. The removal of multi-scale feature extraction (i.e., utilizing only features from the last layer) leads to a drop in accuracy of 2.4%, validating the importance of integration features from various depths.

Table 3. Shows robustness to various perturbations

Perturbation Type	Condition	XceptionNet [24]	MesoNet [25]	Frequency-Aware [10]	Proposed Method
JPEG Compression	Quality Factor = 90	96.2%	91.4%	97.1%	**98.3%**
	Quality Factor = 70	89.5%	82.7%	93.8%	**95.6%**
	Quality Factor = 50	78.6%	71.2%	86.9%	**90.4%**
	Quality Factor = 30	65.1%	58.4%	76.5%	**84.2%**
Gaussian Blur	Kernel Size = 3×3	87.4%	80.3%	91.5%	**94.1%**
	Kernel Size = 5×5	76.2%	69.1%	84.7%	**89.8%**
Gaussian Noise	σ = 0.01	88.7%	81.9%	92.4%	**95.2%**
	σ = 0.05	73.8%	66.5%	84.1%	**89.3%**
Down sampling	Scale = 0.75×	85.1%	79.4%	90.2%	**93.5%**
	Scale = 0.50×	71.5%	64.8%	82.7%	**88.6%**

Table 3 compares the robustness of the proposed method with three deepfake detection models, XceptionNet, MesoNet, and Frequency-Aware, evaluated under the effects of image degradations. The image degradations include JPEG compression, Gaussian blur, Gaussian noise, and image downsampling, which are the typical distortions that happen during the transmission, storage, or sharing of images on social media.

The proposed method outperforms the other methods in all perturbation cases. For JPEG compression, the proposed method achieves 98.3% detection accuracy at QF (Quality Factor) = 90, and

84.2% detection accuracy at QF = 30, which is a compression level considered very lossy. The proposed method also outperforms Frequency-Aware method and XceptionNet by 7.7% and 19.1%, respectively. Similar results are obtained for Gaussian blur, where the proposed method achieves 94.1% and 89.8% detection accuracy for 3x3 and 5x5 Gaussian kernels, respectively. The proposed method is also highly resilient to perturbations caused by Gaussian noise and image down sampling.

Table 4. Sensitivity analysis and Loss weight parameters

Parameter	Variation	Accuracy (%)	Change
Loss weight (λ_3)	0.15	94.10%	−0.6%
	0.225	94.50%	−0.2%
	0.300 (Ours)	**94.70%**	**0.00%**
	0.375	94.60%	−0.1%
	0.45	94.20%	−0.5%
Learning rate	5×10^{-4}	94.20%	−0.5%
	7.5×10^{-4}	94.50%	−0.2%
	1×10^{-4} (Ours)	**94.70%**	**0.00%**
	2.5×10^{-4}	94.40%	−0.3%
	2×10^{-4}	93.80%	−0.9%

The results from the sensitivity analysis shown in Table 4 that the losses we defined and the learning rates showed optimal validation performance. With respect to loss weight λ_3 , an accuracy of 94.7% was attained at λ_3 =0.30). Only small reductions of accuracy of 0.2 and 0.1 percentage points, respectively, were observed when the losses were varied to 0.225 and 0.375. Results of (λ_3=0.30) and a learning rate of 1×10^{-4} were selected for the final configuration.

Batch Processing Throughput,

Batch Size	Throughput (img/s)	Latency per img (ms)
1	11.5	87
8	35.2	227
16	43.0	372
32	38.1	840

Domain Shift Analysis

Training Data	Testing Data	Accuracy	Δ from In-Domain
FF++ (c23)	FF++ (c23)	94.7%	-
FF++ (c23)	FF++ (c40)	93.8%	-0.9%
FF++ (c23)	Celeb-DF	82.3%	-12.4%
FF++ (c23)	DFDC	85.9%	-8.8%
FF++ (c23)	WildDeepfake	75.4%	-19.3%
All datasets	WildDeepfake	78.9%	-15.8%

5.2 Failure Case Analysis

There are a few more difficult situations that we deal with in our method. In the case of highly compressed videos (JPEG QF < 40), accuracy drops to 72.3% due to loss of artifacts after compression. It is even harder for videos with significant blur ($\sigma > 5$), with the accuracy dropping to 68.9%. Extreme head poses (yaw > 60°, pitch > 45°) lessen accuracy to 74.1%, due to insufficient training samples of such head poses. Very short videos (less than 1 second, less than 10 frames) also lack sufficient temporal information. However, the spatial + frequency streams achieve an accuracy of 81.2%. The accuracy for our method can be significantly impacted by other methods of adversarial attacks. Accuracy is evaluated to drop to 23.4% and 8.7% respectively for FGSM (ε=8/255) and for PGD (7 iterations, ε=8/255). Adversarial training, which is claimed to lessen the impact of method attacks, raises robustness of our methods to 71.2% for FGSM and to 58.9% for PGD methods, but there is a reduction of 2.1% in the accuracy of our method for base attacks. This adversarial case is thus a significant limitation of our methods.

5.3 Computational Efficiency

Inference time analysis shows 87ms per image on an NVIDIA A100 GPU, breaking down as:

preprocessing (12ms), feature extraction (58ms), fusion and classification (17ms). The spatial stream dominates computation (45ms), while frequency (8ms) and temporal (5ms per frame) streams are relatively efficient. Batch processing improves throughput, achieving 43 images/second at batch size 16. Model quantization to INT8 reduces model size from 162MB to 41MB (4×) with minimal accuracy loss (0.8%). Quantized model achieves 52ms inference time, enabling real-time processing at ~19fps. Structured pruning removing 60% of parameters reduces size to 65MB and inference to 61ms, with 1.9% accuracy loss—an acceptable trade-off for resource-constrained deployment.

5.4 Attention Visualization and Interpretability

Attention mapping shows patterns of a concentration of attention within face outlines, where (92%) of the attention pixels are within 10 pixels of the outline. That is, of the facial features, the mouth and eyes receive (67%) and (54%) of the attention, respectively. This supports the forensic knowledge that features of a face are difficult to synthesize, while artifacts, which are blends, concentrate around the boundaries. Attention in the frequency domain is for the high frequencies (>70% of attention in frequencies above 0.4 of the Nyquist), which supports the working hypothesis that GAN fingerprints, or artifacts, concentrate in the high frequency. In attention model, a strong spatial and frequency correlation is observed (Pearson r=0.72). This shows that the two streams of attention capture information that is both related and distinct.

6. Conclusion

Deepfake tech is becoming more advanced, and adaptive detection systems with greater technological sophistication and combined ethics will be key to preserving information integrity across the digital world. In the proposed work, we developed the advanced framework for deepfake detection and analysis via deep learning. As a result of the integration, the architecture has proven to perform at the cutting edge by taking into consideration dimensions of both cross-dataset generalization, and the state of the art. We provide a state-of-the-art hybrid of a spatial convolutional neural network with frequency analysis and temporal modeling complimented with multi-head attention. The framework sets a new benchmark for detection of deepfakes and has the potential to counter synthetic media. The future work may include: evaluating the framework's capacity for adversarial robustness, in which the framework would require the design and implementation of certified defenses for adversarial attacks to enable trusted deployment of the framework. Furthermore, few-shot and zero-shot learning to enable faster adaptation to new forms of data manipulation, analysis of audio-visual consistency for increased dimensionality of the modeling.

References

1. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.), Advances in neural information processing systems (Vol. 27, pp. 2672–2680). Curran Associates.
2. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820. https://doi.org/10.15779/Z38RV0D15J
3. World Health Organization. (2020, September 23). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation
4. Federal Bureau of Investigation. (2023). Internet crime report 2022. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2022_IC3Report.pdf
5. Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. Deeptrace. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
6. Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2018). Detection of GAN-generated fake images over social networks. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)(pp. 384–389). IEEE.
7. Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection. IEEE Journal of Selected Topics in Signal Processing, 14(5), 1038–1048. https://doi.org/10.1109/JSTSP.2020.3007254
8. Chai, L., Bau, D., Lim, S.-N., & Isola, P. (2020). What makes fake images detectable? Understanding properties that generalize. In Computer vision—ECCV 2020 (pp. 103–120). Springer. https://doi.org/10.1007/978-3-030-58536-5_7
9. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International

Conference on Computer Vision (pp. 1–11). IEEE. https://doi.org/10.1109/ICCV.2019.00009
10. Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Computer vision—ECCV 2020 (pp. 86–103). Springer. https://doi.org/10.1007/978-3-030-58542-6_6
11. Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS) (pp. 1–6). IEEE. https://doi.org/10.1109/AVSS.2018.8639163
12. Zou, M., Sun, P., Wang, H., & Lyu, S. (2025). Semantic contextualization of face forgery: A new definition, dataset, and detection method. IEEE Transactions on Information Forensics and Security, 20, 111–126. https://doi.org/10.1109/TIFS.2024.3490852
13. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In Proceedings of the International Conference on Learning Representations.
14. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4396–4405). IEEE. https://doi.org/10.1109/CVPR.2019.00453
15. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110–8119). IEEE. https://doi.org/10.1109/CVPR42600.2020.00813
16. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Advances in neural information processing systems (Vol. 33, pp. 6840–6851). Curran Associates.
17. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, Facenheim, C. S., RP, L., Jiang, J., Zhang, S., Wu, P., Zhou, B., & Zhang, W. (2020). DeepFaceLab: Integrated, flexible and extensible face-swapping framework [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2005.05535
18. Nirkin, Y., Keller, Y., & Hassner, T. (2019). FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7184–7193). IEEE. https://doi.org/10.1109/ICCV.2019.00728
19. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2387–2395). IEEE. https://doi.org/10.1109/CVPR.2016.262
20. Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics, 38(4), Article 66, 1–12. https://doi.org/10.1145/3306346.3323035
21. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First order motion model for image animation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Advances in neural information processing systems (Vol. 32). Curran Associates.
22. Popescu, A. C., & Farid, H. (2005). Exposing digital forgeries by detecting traces of resampling. IEEE Transactions on Signal Processing, 53(2), 758–767. https://doi.org/10.1109/TSP.2004.839932
23. Lukas, J., Fridrich, J., & Goljan, M. (2006). Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security, 1(2), 205–214. https://doi.org/10.1109/TIFS.2006.873602
24. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1800–1807). IEEE. https://doi.org/10.1109/CVPR.2017.195
25. Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1–7). IEEE. https://doi.org/10.1109/WIFS.2018.8630761.