SCIENTIFIC Publishing Limited

Digital Technologies Research and Applications

https://ojs.ukscip.com/index.php/dtra

Article

Evaluating Deep Learning Architectures for Effective Detection of Manipulated Media across Multiple Datasets

Achraf Ibnouzaher * **and Noureddine Moumkine**

Team Data Science and Artificial Intelligence, Laboratory of Intelligence Machine (LIM), Faculty of Science and Technology, Hassan II University of Casablanca, Mohammedia 28806, Morocco

* Correspondence: achraf.ibn19@gmail.com

Received: 1 February 2026; **Revised:** 4 March 2026; **Accepted:** 25 March 2026; **Published:** 15 July 2026

Abstract: The advancement of artificial intelligence has enabled the creation of highly realistic synthetic media raising serious concerns about misinformation, identity manipulation, and digital security. As these technologies continue to evolve and become more accessible the potential misuse of manipulated media increases significantly making it easier to spread deceptive content across digital platforms. This growing threat highlights the urgent need for reliable and efficient deepfake detection techniques capable of identifying manipulated content and limiting its potential societal and technological impact. Deep Learning (DL) has emerged as a promising solution for automatically identifying manipulated multimedia content. In this study, we harnessed Transfer Learning (TL) techniques to fine-tune and evaluate the performance of seven different DL models including EfficientNetB5, XceptionNet, EfficientNet3D, Vision Transformer (ViT), CrossViT, ViViT, and CoAtNet. These models were trained and evaluated on three widely used deepfake detection datasets: FaceForensics++, the DeepFake Detection Challenge (DFDC), and CelebDF. Our experimental results demonstrate that hybrid approaches combining Convolutional Neural Networks (CNNs) with Vision Transformer-based architectures outperform individual models in detecting manipulated media. Furthermore, the findings indicate that models trained on the FaceForensics++ and DFDC datasets exhibit stronger generalization capabilities when applied to previously unseen deepfake samples.

Keywords: Deepfake Detection; Transformers; Convolutional Neural Networks; Hybrid Models; FaceForensics; DFDC; CelebDF

1. Introduction

Synthetic media commonly referred to as deepfakes are digital media that are generated or altered utilizing artificial intelligence algorithms to manipulate visual or auditory information in a realistic manner [1]. They have received significant attention in contemporary research due to their capability to influence and misinform by generating fabricated and misleading digital media. While a deepfake may have recreational or harmless uses, it also poses significant risks if exploited for malevolent purposes such as generating persuasive deceptive media to influence public perception, provoke violence, or manipulate election results [2–4]. Furthermore, given the advancement of artificial intelligence and the easy access to sophisticated editing tools the generation of synthetic media has become easier even for users without expertise [5]. This facilitates the generation of extremely persuasive deepfakes that accurately simulate real content.

The scholarly community has presented new AI-driven networks for the identification of deepfakes which attempt to mitigate the challenges caused by synthetic media [6–9]. Nevertheless, a notable problem associated

with existing synthetic-media identification networks is their limited generalization capacity [1,7]. This indicates that such identification systems perform effectively when handling deepfakes originating from the same dataset on which they were trained. However, they find it difficult to attain satisfactory results when confronted with deepfakes produced utilizing different techniques than those utilized for training.

Previous studies have proposed numerous carefully developed machine learning networks for deepfake identification along with innovative methodologies for training such as comprehensive feature encodings, cross-modal learning frameworks, and data augmentation and have assessed these networks on standard deepfake benchmarks. However, considering the large number of scientific articles, it has become difficult to determine which frameworks produce ideal findings and which data sets are most efficient in promoting resilient network effectiveness thereby improving generalization to unseen datasets. Considering these factors, we argue that an in-depth investigation that aggregates a wide range of advanced neural network frameworks trained and evaluated across several well-known deepfake benchmarks in a consolidated manner is essential for obtaining a more comprehensive insight into this problem. We contend that this evaluative study possesses the capability to reveal significant ideas for determining optimal frameworks and datasets to improve the performance of synthetic media identification. It can also aid in tackling the issue of network generalization in the field of deepfake identification.

In this research, we perform an extensive comparative analysis of multiple well-known deep learning models to evaluate their performance in identifying synthetic media. Our principal goal is to determine which of these networks has the potential for outstanding results on unseen data sets and can generalize well. The architectures chosen in our research include CNN-based networks, Transformer-based networks and hybrid networks which are trained on three deepfake identification benchmarks and assessed under both inter-dataset and intra-dataset conditions. Additionally, we assess the difficulty of each dataset to examine whether more challenging datasets lead to enhanced generalization on unfamiliar data. In summary, our research seeks to offer insights into diverse aspects such as evaluating the complexity of various benchmarks for algorithm training identifying the dataset that most effectively supports generalization to unseen samples determining the most efficient architecture for identifying deepfakes within the set of architectures evaluated and identifying the network with the greatest capability to generalize to novel datasets.

The remainder of this research is organized as follows: Section “Related Work” provides a survey of existing research on deepfake identification. Section “Methods and Materials” describes the dataset and the proposed architecture. Section “Results” presents the findings and Section “Discussion” analyzes and interprets the results. Section “Conclusion” summarizes our analysis and suggests directions for future research.

2. Related Work

Recently, a considerable number of studies on forgery content identification have been conducted. The majority of research employs Convolutional Neural Network architectures trained on extensive datasets to identify deepfake content. These studies also utilize various techniques such as spatiotemporal feature analysis [10], multimodal fusion [11], biological signal extraction [12], hybrid architectural designs [13,14], and data augmentation techniques [15] aiming to enhance the generalization capabilities of the models. In the following, we review some recently proposed approaches as well as several well-known forgery media detection studies. We decided to review research that exhibits similarities with ours concentrating on common elements like the datasets utilized to assess and train the proposed networks and the choice of identification networks.

Rössler et al. [16] introduced FaceForensics in 2019 a dataset designed for synthetic media identification. It comprises more than 1.5 million altered images and has been made publicly accessible. Utilizing the dataset, the authors performed a thorough evaluation of various data-driven deepfake identification approaches. These approaches comprised modern deep learning frameworks such as XceptionNet [17] and MesoNet [18] as well as conventional machine learning models (SVMs) trained on manually crafted features. Through a comprehensive analysis the authors found that a baseline deep convolutional neural network XceptionNet surpasses other evaluated networks for classifying low-quality compressed images. The authors further showed that humans perform worse than data-driven networks in identifying synthetic media. Nonetheless, the research lacks an inter-dataset evaluation of the frameworks which could provide insight into their ability to generalize to unseen and varied contexts.

Cheng et al. [19] introduced a synthetic media identification framework by emphasizing temporal features in

videos to exploit inconsistencies between consecutive images. For analyzing temporal information they utilized a hybrid convolutional-recurrent framework comprising a Convolutional Neural Network (CNN) for feature extraction and a Bidirectional Recurrent Neural Network (BRNN) to process temporal dependencies in video frames. Specifically, two CNN architectures DenseNet [20] and ResNet [21] are investigated for extracting deep features. The framework also incorporates a purposefully designed preprocessing procedure for facial images prior to network input. The frameworks were evaluated using the well-known FaceForensics manipulated media identification dataset [16] achieving excellent performance in an intra-dataset evaluation setting. Similar to the previous study the research does not include a cross-dataset evaluation of the frameworks.

Ciftci et al. [22] proposed using biological signals (e.g., photoplethysmography (PPG)) instead of traditional image features to train their networks to identify subtle variations in motion and color in RGB videos. The authors use a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN) to generate predictions on the extracted features which are then combined to produce a final classification result. The proposed identification approach for synthetic media achieved encouraging results when evaluated on two different datasets namely FaceForensics [16] and CelebDF [23] utilizing inter-dataset and intra-dataset settings.

Zhu et al. [24] proposed an architecture for synthetic media identification that employs three-dimensional facial structural features. The researchers demonstrated that the combination of direct illumination features and three-dimensional identity texture greatly improved identification while enhancing the network's capacity to generalize. The identification network was trained using both extracted face frames and their corresponding three-dimensional features. Xception was used for deep feature extraction. The paper also presents a comprehensive analysis of multiple feature combination techniques. The designed framework was trained on FaceForensic++ [16] dataset and then tested on other benchmarks: DFDC [25], Google Deepfake Detection Dataset [26] and FaceForensic++. The reported assessment metrics demonstrated encouraging outcomes across all three datasets emphasizing the network's strong generalization performance compared to earlier proposed forged video identification approaches.

Wang et al. [27] presented a Multi-modal Multi-scale Transformer network that analyzes patches of various dimensions to detect local anomalies in an input frame across multiple spatial scales. Multi-modal Multi-scale Transformer also uses RGB information in conjunction with frequency-domain information via a sophisticated cross modality merging module to identify tampering-related artifacts more effectively. Through comprehensive evaluations researchers demonstrate the efficacy of Multi-modal Multi-scale Transformer and show that their network surpasses state-of-the-art forged media identification networks by significant degrees.

Coccomini et al. [28] suggested a video-forged media identification network based on a hybrid-transformer framework. Researchers utilized EfficientNet as an extractor of features. The features are subsequently utilized to train 2 distinct models of ViT networks: Convolutional-CrossViT and Efficient-ViT. The result demonstrated that the network consisting of ViT and EfficientNet-B0 feature extractor attained the highest performance metrics relative to the other networks that they evaluated.

Zhao et al. [29] proposed an explainable Temporal-Spatial Video Transformer (STVT) for manipulated video identification. The presented network integrates a novel decomposed self-subtraction mechanism along with spatio-temporal self-attention to capture temporal discrepancies and tampering-related spatial artifacts. STVT can also depict the distinctive areas for both temporal and spatial dimensions by utilizing the relevance distribution method. Comprehensive experiments on various benchmarks were carried out demonstrating a high performance of STVT in both inter-dataset and intra-dataset manipulated media identification and proving the robustness and efficacy of the proposed network.

Ramadhani et al. [30] proposed a spatiotemporal method for deepfake video detection based on analyzing spatial and temporal information across video frames. The authors first extracted 25 facial landmark regions from each video and then applied a combination of Depthwise Separable Convolution (DSC) and Convolutional Block Attention Module (CBAM) to capture discriminative spatial features. These features were then processed using a Video Vision Transformer (ViViT) to model spatiotemporal inconsistencies indicative of deepfake manipulations. The approach was evaluated on the Celeb-DF dataset and demonstrated effective discrimination between real and manipulated videos.

Through this review of existing studies it is apparent that the scientific community extensively utilizes deep-learning networks alongside other methods to develop efficient and robust manipulated content detectors. How-

ever, upon thorough review of the published research most prior works evaluate models under different experimental conditions such as varying dataset splits, preprocessing techniques, or evaluation metrics making direct comparisons between approaches difficult and limiting the ability to extract insights that could guide the development of stronger and more effective architectures. In addition, many works evaluate models only on familiar or in-distribution data which provides limited insight into their robustness and generalization capabilities. Moreover, prior work utilizes different datasets making it challenging to select a dataset from which models can learn most effectively and generalize well to unseen data. To tackle this, we examine several commonly utilized models in the literature on forgery video identification. We also use well-known benchmark datasets for experimentation and aim to determine which benchmarks provide the optimal generalization abilities for the networks.

3. Materials and Methods

This section presents the complete processing steps carried out in this research for evaluating and training the networks, as shown in **Figure 1**. The pipeline begins with data preprocessing where we use the Mediapipe [31] to detect facial landmarks from the extracted frames. All detected faces are then cropped around the center and resized to 224×224 pixels before being input into the networks for training. We use pretrained models on ImageNet, and we remove the last layer and add a new linear classification head. For evaluation, the networks are assessed in two ways: intra-dataset and inter-dataset. In intra-dataset evaluation, networks are tested on the same dataset they were trained on. For example, a network trained on the FaceForensics dataset is tested on the FaceForensics evaluation set. The key goal of intradataset assessment is to determine which network attains the best accuracy in comparison with other networks on each dataset and to gain insights into which benchmark is easier or more challenging for the networks to learn. In interdataset testing, networks are trained on a benchmark and tested on another benchmark; for example, a network trained on FaceForensics is assessed on CelebDF and DFDC. The purpose of interdataset assessment is to evaluate how efficiently the training dataset enables networks to generalize to unseen datasets and to examine how well the models perform on these datasets.

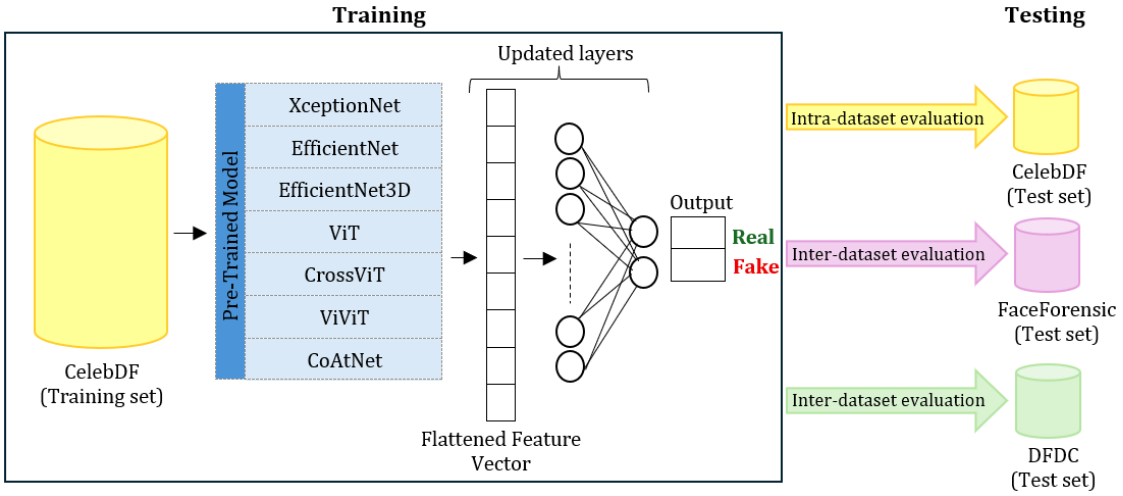

Figure 1. Flow diagram of the steps followed for training and testing the networks.

3.1. Dataset

In this research, we assess and train multiple neural network architectures on three commonly used forged video identification benchmarks: FaceForensics [16], DFDC [25], and CelebDF [23]. All three benchmarks consist of synthetic and real videos, where synthetic ones are produced through different deepfake creation techniques.

- **FaceForensics** [16] is a large-scale dataset of modified facial frames, containing over one million frames. This dataset contains 1,000 genuine videos collected from YouTube. These authentic videos were altered using four face manipulation techniques, which include traditional computer graphics-based techniques (FaceSwap [32], Face2Face [33]) and neural network-based approaches (NeuralTextures [34], Deepfakes [35]), to generate forged

videos. For each technique, 1,000 forged videos were produced. The dataset also contains two compressed formats of videos, with varying quality levels: high-quality (HQ) videos and low-quality (LQ) videos. In our experiments, the networks are evaluated on the low-quality (LQ) subset of the dataset.

- **Deepfake Detection Challenge (DFDC) dataset** [25] consists of approximately 128,000 videos of which roughly 104,000 are synthetic. The synthetic videos were generated using five facial manipulation techniques: StyleGAN [36], FSGAN [37], NTH [38], MM/NN, and Deepfake Autoencoder [39]. A randomly chosen subset of videos was subjected to a basic sharpening post-processing procedure to enhance the perceptual quality of the videos. The DFDC dataset also comprises videos that have experienced audio manipulation. Nevertheless, in this work audio features are not utilized to assess or train our networks.
- **CelebDF** [23] is a large-scale fake video dataset, containing over two million frames. It consists of 5,539 deepfake videos and 590 original videos. The authentic videos were obtained from publicly accessible YouTube content, including interviews of 60 celebrities with different distributions in ethnicity, age, and gender. The deepfake videos were produced using a sophisticated deepfake synthesis model [40,41] which enhances the visual resolution of the generated videos by replacing the facial features of a person with those of another across all 60 subjects.

3.2. Model

We examined a diverse set of deep learning models categorized into transformer-based, CNN-based, and hybrid architectures chosen for their strong performance on the ImageNet benchmark or in deepfake detection tasks. Additionally, in the CNN–ViT combination, we evaluated the use of single and dual CNNs as feature extractors.

3.2.1. CNN-Based Architectures

- **XceptionNet** [17] is a deep convolutional neural network framework derived from Google's Inception architecture [42]. The XceptionNet model replaces standard Inception modules with depthwise separable convolutional layers in which spatial convolutions are applied independently to each channel followed by pointwise (1×1) convolutions to combine information from different channels. This factorization reduces the number of trainable parameters as well as computational cost while still being able to capture spatial and channel-wise correlations effectively. The model comprises three flows: an entry flow that gradually reduces the spatial dimensions and increases the depth of features, a middle flow composed of a number of depth-wise separable convolution layers to extract deep features, and an exit flow to generate high-level features for image classification tasks. This design improves the efficiency of the network and reduces overfitting on unseen datasets. **Figure 2d** illustrates the network architecture. The choice of the XceptionNet algorithm for analysis in this research lies in its strong performance reported in the literature on deepfake detection [16,25,43].
- **EfficientNetB5** [44] is a variant of the EfficientNet family of convolution neural network (CNN) structures. EfficientNetB5 employs a compound scaling approach that proportionally scales the depth, width, and resolution of the network by using a composite coefficient. This scaling enables the network to balance between high precision and computational capability. The architecture is built upon mobile inverted bottleneck (MBConv) blocks along with squeeze-and-excite optimization to further enhance its performance and parameter efficiency. We chose to examine this model in our analysis study because of its successful results in the Google-sponsored Deepfake Detection Challenge (DFDC) competition [25]. Due to its successful results in detecting subtle manipulations in the facial images of the DFDC dataset we would like to examine the capabilities of this model in our study.
- **EfficientNet3D** [45] is a video classification network built upon the same scaling principles as the baseline EfficientNet framework [44] but it is specifically developed to operate on three-dimensional datasets like 3D medical scans and videos. These networks utilize three-dimensional convolutions in place of two-dimensional layers for feature extraction. In addition, EfficientNet3D networks typically utilize a sufficient number of layers which enables them to capture intricate spatial and temporal patterns in 3D datasets. EfficientNet3D networks have been used for numerous computer vision applications such as image segmentation, action recognition, and video classification [46]. An example of 3D and 2D convolutions is illustrated in **Figure 2e**. We opted to utilize the EfficientNet3D network for our research because (1) it has demonstrated strong performance in analyzing dynamic patterns in video data and (2) we wanted to investigate whether temporal information

in a video-based model enhances deepfake detection compared to image-based approaches. We chose the EfficientNet3D network pretrained on 8 frames from each video to carry out our experiments.

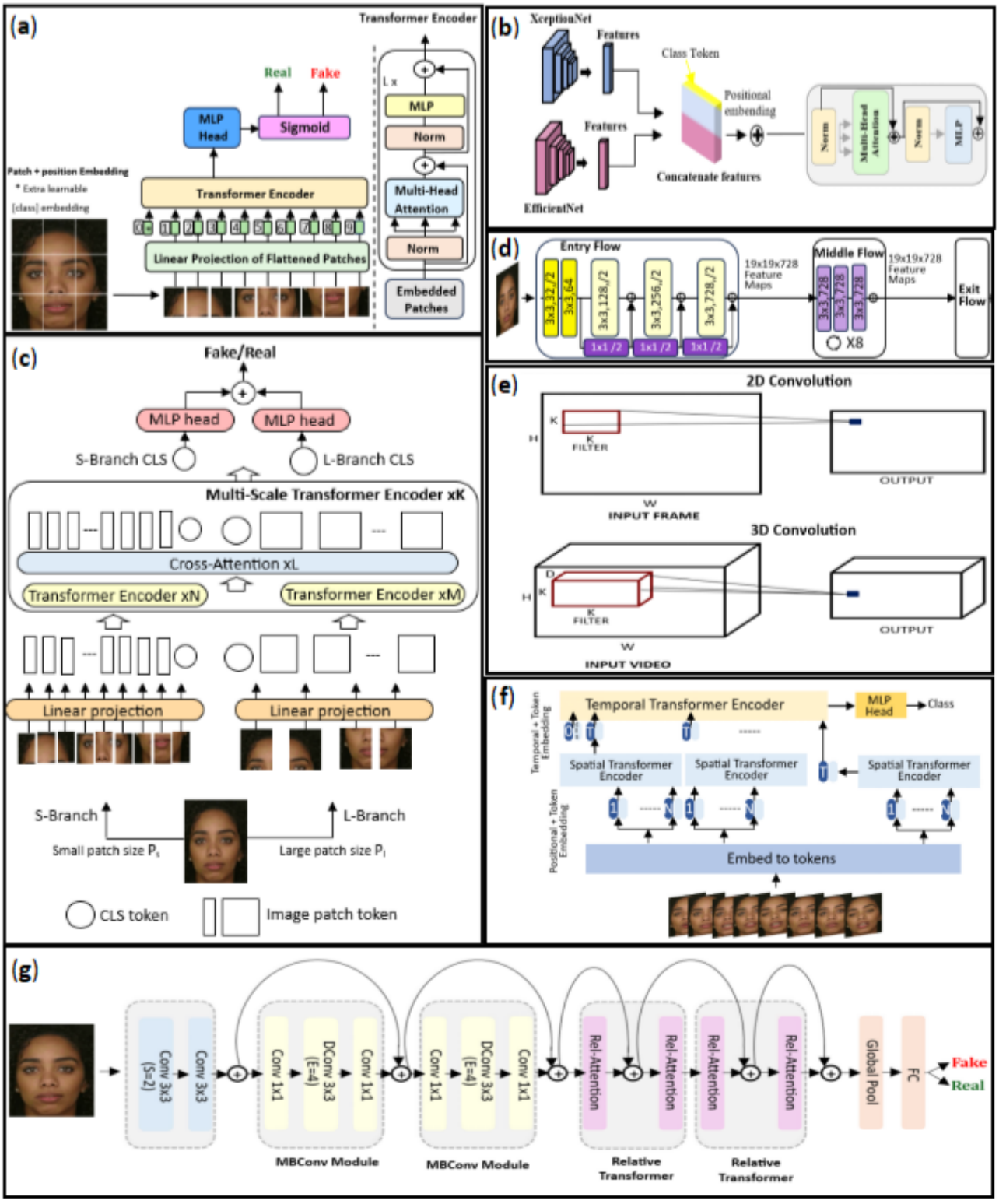

Figure 2. Model Architectures used in this study. (**a**) Vision Transformer architecture, (**b**) 2CNN-ViT hybrid model, (**c**) CrossViT model, (**d**) XceptionNet, (**e**) 3D and 2D convolutions, (**f**) ViViT, (**g**) CoAtNet.

3.2.2. Transformer-Based Architectures

- **Vision Transformer (ViT)** [47] is a deep learning architecture inspired by transformers which are originally used for NLP tasks. In computer vision, ViT was introduced as a foundational transformer-based framework for image recognition applications [48]. Vision Transformer uses self-attention mechanisms to analyze visual information. Its core methodology involves dividing an image into fixed-size patches, flattening them and then projecting these embeddings into a lower-dimensional embedding space. Positional encodings are added to preserve spatial information also a classification token is prepended to the sequence before processing it by the transformer encoder. This allows the network to learn relationships among different regions of the image and capture global contextual information efficiently. Due to its state-of-the-art performance compared to certain traditional convolutional neural networks on large-scale image classification benchmarks we choose to examine the performance of Vision Transformer for our analysis of deepfakes. Specifically, we evaluate and train the ViT-Base model for the identification of deepfakes. **Figure 2a** presents the architecture of a vision transformer.
- **CrossViT** [49] is a variant of the Vision Transformer architecture that introduces a dual-branch design operating at different patch scales to compute image embeddings. The cross-scale attention mechanism enhances feature representation by allowing interactions between different resolutions and enabling complementary learning across them. Such multi-scale designs improve the model's capacity to capture fine-grained local patterns and global semantic context simultaneously giving greater flexibility to model objects with varying sizes. The CrossViT model achieves comparable performance to state-of-the-art image classification networks and in some cases outperforms standard Vision Transformers [50]. Beyond image classification, CrossViT has achieved outstanding results for object detection and image segmentation tasks [50]. **Figure 2c** illustrates the dual-branch architecture and cross-attention interaction in CrossViT. Due to its efficiency in capturing local details that are limited in standard ViTs and outstanding performance on ImageNet we employ CrossViT for deepfake identification and analyze its performance relative to other participating networks.
- **Video Vision Transformer (ViViT)** [51] is a video classification network built upon the transformer structure. ViViT employs self-attention across temporal and spatial dimensions instead of relying solely on convolutional operations or the 2D spatial attention mechanism used in the Vision Transformer for image classification. The Video Vision Transformer framework adjusts the transformer structure traditionally utilized for image identification to capture temporal and spatial representations from a series of image-wise segments. This is performed by expanding attention operation from 2D spatial domain to a three-dimensional spatiotemporal volume. Like the ViT network, the Video Vision Transformer utilizes positional encodings and linear projections to encode the positions of the derived tokens. In the original ViViT paper [51], researchers experimented with several transformer variants. Among various methods, the factorized encoder, which computes spatial and temporal attention independently within each module, was observed to perform best in terms of accuracy and efficiency; therefore, we adopt this factorized spatiotemporal design in this research. Factorized spatiotemporal operation is shown in **Figure 2f**. We choose to assess ViViT for synthetic video identification and compare it with the 3D CNN-based video recognition model EfficientNet3D. We also use the eight-images-per-video version of ViViT network, as with EfficientNet3D network.

3.2.3. Hybrid Architectures

- **CNN+ViT** is a hybrid architecture that combines convolutional neural networks (CNNs) and Vision Transformers (ViTs) for image recognition tasks. The CNN serves as a feature extractor, generating hierarchical representations from the input image. These feature maps are then reshaped into a sequence of tokens and input into the ViT module, where self-attention mechanisms capture long-range dependencies and correlations between different image regions. Unlike traditional classification approaches that rely on handcrafted features or fully connected classifiers, the transformer-based classifier assigns dynamic importance to each feature token through attention, enabling more discriminative and robust decision-making. In this work, we evaluate three feature fusion strategies: Xception + ViT, EfficientNet + ViT, and 2CNN + ViT (where features from both XceptionNet and EfficientNetB5 are fused and passed to the ViT). **Figure 2b** illustrates the architecture of the

2CNN-ViT hybrid model.

- **CoAtNet model** [52] is a hybrid convolution–attention architecture that unifies Convolutional Neural Networks (CNNs) and self-attention modules within a single hierarchical backbone in order to exploit both local feature extraction and global dependency modeling. In contrast to CNN-ViT models, where convolution and self-attention components are separated into feature-extraction and encoding stages, CoAtNet progressively integrates convolutional layers in the early stages and attention mechanisms in deeper stages, forming a continuous architectural transition rather than a two-stage pipeline. The architecture consists of two key components: MBConv modules (mobile inverted bottleneck convolutions) to extract local spatial features, and Transformer blocks with relative self-attention to capture long-range dependencies and global contextual information. An outline of the CoAtNet architecture is illustrated in **Figure 2g**. In this work, we include CoAtNet to investigate whether integrating convolution and attention within a single backbone can enhance the ability of the model to detect deepfakes compared to the separate CNN and attention modules used in CNN-ViT.

3.3. Implementation Details

For each dataset we selected 720 videos (360 real and 360 fake) for training, 140 videos for validation, and 140 videos for testing. To prepare the datasets for training and evaluation we applied a systematic pre-processing pipeline to ensure consistency across all video clips. We employed MediaPipe [31] to detect and extract facial regions from each video frame owing to its capability for real-time, high-precision landmark detection. All detected faces were then cropped around the center and resized to 224 × 224 pixels. To reduce computational load while preserving temporal information we sampled 10 evenly spaced frames per video balancing efficiency with content representativeness. Additionally, image augmentation techniques such as flipping, rotation, scaling, shearing, and translation were applied to mitigate potential overfitting. Finally, all extracted frames were organized into fake and real classes facilitating efficient evaluation and training of the detection networks. This preprocessing pipeline ensures a standardized and well-structured dataset thereby enhancing the reliability of the deepfake detection process. Training was performed using the PyTorch framework with the Adam optimizer a learning rate of 0.0001 and 15 epochs. The batch size was set to 4 for the two video-based networks and 16 for the remaining networks.

3.4. Evaluation Metrics

Several evaluation metrics were utilized to assess the performance of the deep learning frameworks, including accuracy (ACC), specificity (SPE), recall (REC), F1-score, precision (PRE), and AUC [53]. The mathematical definitions of these metrics are provided as follows:

$$\text{accuracy (ACC)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{1}$$

Where TP represents the number of correctly classified real frames, TN represents the number of correctly classified fake frames, FP represents the number of fake frames incorrectly classified as real, and FN represents the number of real frames incorrectly classified as fake. Precision (PRE) is the percentage of frames predicted as real that are actually real and is defined as:

$$\text{precision (PRE)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Specificity is the percentage of correctly classified fake frames out of all actual fake frames and it is calculated as:

$$\text{Specificity (SPE)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Recall represents the proportion of actual real frames correctly classified by the model. It is mathematically defined as:

$$\text{Recall (REC)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

The recall and precision harmonic average is known as the F1 score, mathematically expressed as,

$$\text{F}_1 = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}} \tag{5}$$

4. Results

We carried out extensive experiments and assessments on several deep learning networks: three CNN-based networks, three transformer-based networks, and two hybrid models. These evaluations were performed across three different datasets. The analysis involves assessing all networks across inter-dataset scenarios (trained on one benchmark and tested on another benchmark) and intra-dataset scenarios (tested and trained on the same benchmark).

Tables 1 and **2** summarize the outcomes of networks trained on the CelebDF dataset under intra-dataset and inter-dataset testing, respectively. As shown in **Table 1**, all models achieve high accuracy (ACC), recall (REC), specificity (SPE), precision (PRE), F1 score, and AUC when tested on CelebDF itself, with even the lowest-performing model achieving ACC of 83.76% and similarly strong values for other metrics. This suggests that the CelebDF data is comparatively easy and therefore the networks can precisely differentiate between fake and real instances. In contrast, when CelebDF-trained models are evaluated on unseen datasets FaceForensics and DFDC, all models perform poorly with accuracy not surpassing 70% on the FaceForensics and DFDC datasets, as summarized in **Table 2**. This indicates that training networks on easier data does not yield good results when evaluated on an unseen dataset. We can conclude that the CelebDF dataset does not pose a significant challenge for the networks to learn and is relatively easy to differentiate between real and fake instances. Moreover, this data does not improve the networks' capacity to learn strong discriminative features between authentic and manipulated faces, thereby weakening the models' generalization capability across unseen datasets.

Table 1. Comparison of classification performance for deep learning networks trained and tested on CelebDF [23] dataset.

Network	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	0.8376	0.8457	0.8113	0.8113	0.8114	0.8367
XceptionNet	0.8488	0.8584	0.8369	0.8383	0.8397	0.8497
EfficientNet-3D	0.9183	0.9272	0.9070	0.9098	0.9108	0.9229
ViT	0.8750	0.8720	0.8601	0.8659	0.8719	0.8724
CrossViT	0.8871	0.8890	0.8787	0.8848	0.8910	0.8853
ViViT	0.9250	0.9124	0.9143	0.9162	0.9181	0.9281
EfficientnetB5 + ViT	0.9384	0.9320	0.9230	0.9221	0.9212	0.9220
XceptionNet + ViT	0.9406	0.9330	0.9240	0.9280	0.9323	0.9330
2CNN+ViT	0.9570	0.9450	0.9440	0.9420	0.9401	0.9590
CoAtNet	0.9264	0.9130	0.9040	0.9110	0.9180	0.9157

Table 2. Comparison of classification performance of deep learning networks trained on CelebDF [22] and evaluated on FaceForensics and DFDC datasets.

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	FF	0.5198	0.5394	0.5002	0.5023	0.5046	0.5201
	DFDC	0.4700	0.4752	0.4648	0.4728	0.4810	0.4717
XceptionNet	FF	0.5358	0.5263	0.5353	0.5277	0.5202	0.5468
	DFDC	0.4630	0.4755	0.4527	0.4706	0.4895	0.4850
EfficientNet-3D	FF	0.6329	0.6310	0.6248	0.6224	0.6200	0.6301
	DFDC	0.5850	0.5950	0.5750	0.5725	0.5700	0.5897
ViT	FF	0.5912	0.5822	0.5978	0.5990	0.6001	0.6095
	DFDC	0.5295	0.5240	0.5150	0.5224	0.5302	0.5269
CrossViT	FF	0.6130	0.6159	0.6001	0.6060	0.6120	0.6145
	DFDC	0.5380	0.5509	0.5151	0.5212	0.5280	0.5395
ViViT	FF	0.6443	0.6438	0.6348	0.6324	0.6300	0.6455
	DFDC	0.6025	0.6152	0.5998	0.5961	0.5925	0.6076
EfficientnetB5 + ViT	FF	0.6596	0.6563	0.6609	0.6554	0.6500	0.6590
	DFDC	0.6403	0.6556	0.6550	0.6476	0.6403	0.6397
XceptionNet + ViT	FF	0.6709	0.6786	0.6698	0.6633	0.6570	0.6697
	DFDC	0.6307	0.6313	0.6201	0.6274	0.6350	0.6296

Table 2. *Cont.*

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
2CNN+ViT	FF	0.6901	0.6804	0.6898	0.6824	0.6750	0.6890
	DFDC	0.6687	0.6723	0.6552	0.6537	0.6523	0.6590
CoAtNet	FF	0.6517	0.6332	0.6302	0.6428	0.6556	0.6539
	DFDC	0.6226	0.5993	0.5951	0.6064	0.6180	0.5983

Tables 3 and **4** present the performance of models trained on the FaceForensics dataset under intradataset and interdataset testing, respectively. In the intradataset evaluation **Table 3**, only a few models exceed 90% accuracy, while the remaining models exhibit noticeably lower performance across all evaluation metrics. These relatively low values indicate that several architectures struggle to achieve strong discriminative capability when distinguishing real from manipulated samples, suggesting that FaceForensics presents a relatively complex and challenging dataset for manipulated media detection. In the inter-dataset testing, deep learning networks trained on FaceForensics exhibit satisfactory performance even when tested on previously unseen domains, as illustrated in **Table 4**, demonstrating a stronger ability to generalise compared to models trained on CelebDF, which show comparatively poor generalisation. For instance, the 2CNN-ViT model trained on FaceForensics and tested on CelebDF achieves an accuracy of 81.01% whereas when trained on CelebDF and testing on FaceForensics attains only 69.01%. Overall, these findings support the claim that training on more challenging datasets leads to improved generalisation capability in deepfake detection models.

Table 3. Comparison of classification performance for deep learning networks trained and tested on FaceForensics [16] dataset.

Network	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	0.7795	0.7840	0.7550	0.7535	0.7520	0.7702
XceptionNet	0.7803	0.7950	0.7730	0.7754	0.7778	0.7985
EfficientNet-3D	0.8715	0.8430	0.8300	0.8674	0.8750	0.8684
ViT	0.8207	0.8220	0.8000	0.8010	0.8020	0.8219
CrossViT	0.8571	0.8650	0.8350	0.8335	0.8321	0.8576
ViViT	0.8818	0.8630	0.8651	0.8700	0.9050	0.9060
EfficientnetB5 + ViT	0.9089	0.9020	0.9050	0.9112	0.9174	0.9074
XceptionNet + ViT	0.9116	0.9050	0.8800	0.8974	0.9160	0.9096
2CNN+ViT	0.9263	0.9230	0.9138	0.9219	0.9300	0.9254
CoAtNet	0.9007	0.9080	0.8600	0.8773	0.8950	0.8825

Table 4. Comparison of classification performance of deep learning networks trained on FaceForensics [16] and tested on CelebDF and DFDC dataset.

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	CelebDF	0.6436	0.6434	0.6356	0.6373	0.6391	0.6395
	DFDC	0.6079	0.6017	0.5939	0.5932	0.5926	0.5978
XceptionNet	CelebDF	0.6504	0.6595	0.6404	0.6469	0.6536	0.6500
	DFDC	0.6007	0.6170	0.5880	0.5934	0.5989	0.6025
EfficientNet-3D	CelebDF	0.7414	0.7675	0.7535	0.7607	0.7680	0.7605
	DFDC	0.7292	0.7380	0.7240	0.7252	0.7265	0.7310
ViT	CelebDF	0.6808	0.6605	0.6628	0.6754	0.6887	0.6616
	DFDC	0.6398	0.6395	0.6282	0.6287	0.6292	0.6339
CrossViT	CelebDF	0.7257	0.7205	0.6930	0.6948	0.6970	0.7068
	DFDC	0.6675	0.6595	0.6514	0.6570	0.6630	0.6555
ViViT	CelebDF	0.7698	0.7795	0.7605	0.7577	0.7550	0.7700
	DFDC	0.7489	0.7475	0.7324	0.7377	0.7430	0.7400
EfficientnetB5 + ViT	CelebDF	0.7848	0.7815	0.7780	0.7800	0.7820	0.7798
	DFDC	0.7625	0.7700	0.7650	0.7614	0.7580	0.7675
XceptionNet + ViT	CelebDF	0.7895	0.7800	0.7650	0.7718	0.7789	0.7725
	DFDC	0.7695	0.7670	0.7520	0.7612	0.7705	0.7595

Table 4. *Cont.*

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
2CNN+ViT	CelebDF	0.8101	0.8180	0.8020	0.8035	0.8050	0.8100
	DFDC	0.7987	0.7935	0.7860	0.7925	0.7990	0.7898
CoAtNet	CelebDF	0.7808	0.7765	0.7790	0.7719	0.7650	0.7778
	DFDC	0.7603	0.7450	0.7560	0.7489	0.7420	0.7505

Tables 5 and **6** summarize the performance of deep learning models trained on the DFDC dataset under intra-dataset and interdataset evaluation settings, respectively. In intra-dataset evaluation, all models achieved noticeably lower results compared to their performance on the other two datasets. For example, the hybrid architecture 2CNN+ViT achieved an accuracy of 90.40% on DFDC, whereas it reached 95.70% and 92.63% on CelebDF and FaceForensics, respectively. Similarly, the video-based classification model ViViT obtained an accuracy of 84.75% on DFDC, compared to 92.50% on CelebDF and 88.18% on FaceForensics. These results indicate that DFDC is the most difficult benchmark among the three. In the inter-dataset evaluation, networks trained on DFDC still achieved satisfactory outcomes on out-of-distribution datasets compared to models trained on CelebDF. For example, XceptionNet achieved an accuracy of 46.30% when trained on CelebDF and tested on DFDC, whereas training on DFDC and testing on CelebDF resulted in a higher accuracy of 66.04%. This demonstrates that training on more challenging datasets leads to better generalization across datasets.

Table 5. Comparison of classification performance for deep learning networks trained and tested on the DFDC [25] dataset.

Network	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	0.7682	0.7767	0.7635	0.7654	0.7674	0.7794
XceptionNet	0.7510	0.7630	0.7390	0.7370	0.7350	0.7568
EfficientNet-3D	0.8305	0.8350	0.8200	0.8274	0.8350	0.8390
ViT	0.8035	0.7920	0.7897	0.7961	0.8025	0.8144
CrossViT	0.8105	0.8085	0.8120	0.8162	0.8204	0.8377
ViViT	0.8475	0.8415	0.8403	0.8429	0.8455	0.8469
EfficientnetB5 + ViT	0.8551	0.8520	0.8450	0.8500	0.8550	0.8630
XceptionNet + ViT	0.8784	0.8802	0.8750	0.8775	0.8756	0.8782
2CNN+ViT	0.9040	0.9180	0.8979	0.9009	0.9042	0.9063
CoAtNet	0.8543	0.8580	0.8440	0.8475	0.8510	0.8514

Table 6. Comparison of classification performance of deep learning networks trained on DFDC [25] and tested on CelebDF and FaceForensics dataset.

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
EfficientnetB5	CelebDF	0.6521	0.6586	0.6456	0.6471	0.6486	0.6564
	FF	0.6198	0.6394	0.6002	0.6073	0.6146	0.6301
XceptionNet	CelebDF	0.6604	0.6704	0.6504	0.6570	0.6636	0.6747
	FF	0.6207	0.6334	0.6080	0.6234	0.6289	0.6301
EfficientNet-3D	CelebDF	0.7586	0.7652	0.7520	0.7530	0.7540	0.7647
	FF	0.7377	0.7454	0.7300	0.7317	0.7330	0.7443
ViT	CelebDF	0.6963	0.7046	0.6880	0.6900	0.6920	0.6907
	FF	0.6695	0.6840	0.6550	0.6570	0.6590	0.6817
CrossViT	CelebDF	0.7391	0.7552	0.7230	0.7228	0.7226	0.7327
	FF	0.6812	0.6884	0.6740	0.6855	0.6870	0.7068
ViViT	CelebDF	0.7790	0.7931	0.7649	0.7764	0.7880	0.7883
	FF	0.7532	0.7614	0.7450	0.7575	0.7690	0.7681
EfficientnetB5 + ViT	CelebDF	0.7965	0.8073	0.7857	0.7918	0.7980	0.8043
	FF	0.7837	0.7924	0.7750	0.7765	0.7780	0.7811
XceptionNet + ViT	CelebDF	0.8209	0.8348	0.8070	0.8110	0.8150	0.8202
	FF	0.7798	0.7876	0.7720	0.7735	0.7750	0.7825

Table 6. *Cont.*

Network	Dataset	ACC	SPE	REC	F1 Score	PRE	AUC
2CNN+ViT	CelebDF	0.8548	0.8687	0.8409	0.8479	0.8550	0.8650
	FF	0.8187	0.8234	0.8140	0.8210	0.8280	0.8219
CoAtNet	CelebDF	0.7911	0.7892	0.7930	0.7950	0.7970	0.8021
	FF	0.7806	0.7916	0.7696	0.7728	0.7647	0.7811

Figure 3 supports these observations by illustrating the relative accuracy drops of the models on unseen datasets. As shown in **Figure 3**, models trained on the Celeb-DF dataset show a larger decrease in performance with relative accuracy drops exceeding 40%. In contrast, the same models trained on FaceForensics and DFDC experience smaller reductions. Models trained on FaceForensics have a moderate accuracy drop not exceeding 23% while models trained on DFDC have the smallest accuracy drop not exceeding 19%. Overall, FaceForensics and DFDC are both suitable datasets for training deepfake detection models with DFDC being the better choice to enable models to generalize on unseen datasets.

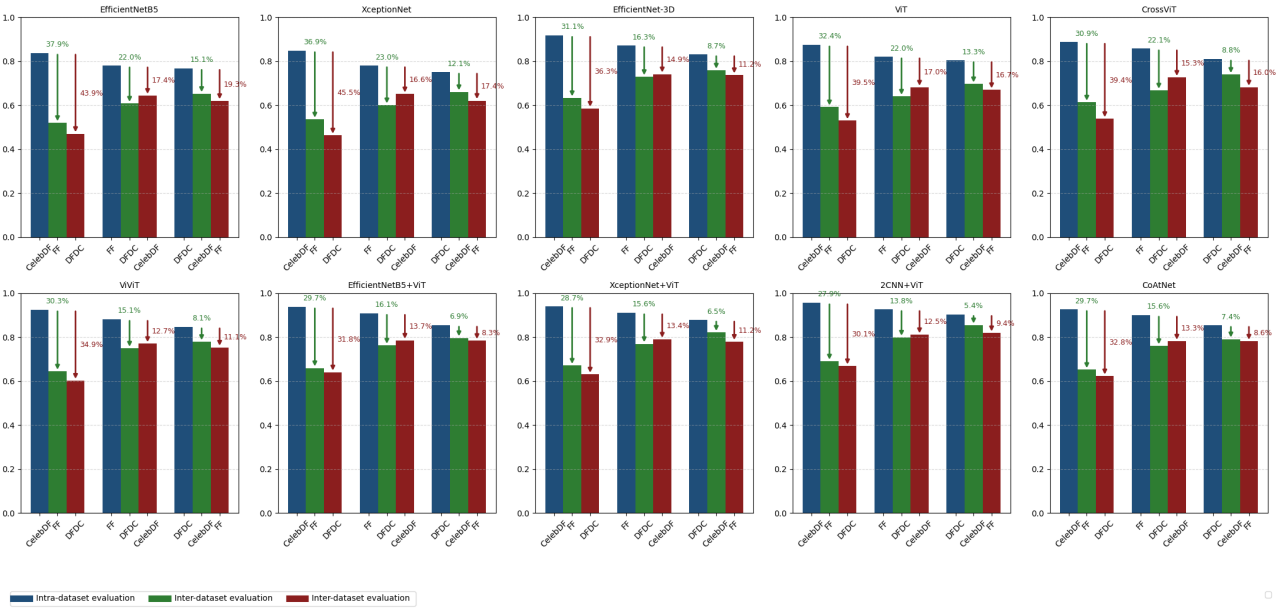

Figure 3. Dataset Impact on Model Generalization Measured by Accuracy Drop.

5. Discussion

Figure 4a provides a detailed summary of the average performance achieved by each network in the intradataset evaluation on the three datasets. Conventional 2D CNN architectures (EfficientNetB5 and XceptionNet) exhibit the lowest performance with no metric exceeding 80%. Transformer-based models (ViT and CrossViT) show noticeable improvements achieving relative performance gains of approximately 5–7% compared to CNN-only models highlighting the advantage of global attention mechanisms in modeling long-range dependencies. Overall, transformer-based architectures outperform CNN-only models across all evaluation metrics. Hybrid architectures achieve substantial performance gains over both individual CNNs and transformers improving all evaluation metrics by approximately 12–15% indicating the effectiveness of combining local feature extraction from CNNs with the global contextual modeling capability of transformers. Furthermore, the utilization of two different CNN backbones for feature extractors results in the highest performance for all the metrics which indicates that the feature fusion strategies effectively exploit complementary information from multiple representations and help the transformer to learn a diverse set of features. The CoAtNet architecture also demonstrates strong performance but remains lower than the multi-branch hybrid model (2CNN+ViT) which indicates that though intrinsic fusion is effective the utilization of multiple diverse CNN backbones before the attention mechanism results in better feature representation and predictive performance.

For the two video-based models both achieve competitive results with ViViT reaching the highest performance among them; however, their performance remains below that of the hybrid models. Overall, these results highlight that architectures leveraging both local and global information either through explicit hybrid designs or integrated convolution–attention frameworks consistently achieve higher performance. In the inter-dataset evaluation the results indicate that the networks show noticeable decreases in performance compared to the intra-dataset assessment as shown in **Figure 4b–d**. This discrepancy is expected as identification networks typically exhibit a performance drop when tested on out-of-distribution datasets. However, hybrid models consistently emerge as the top-performing architectures outperforming individual CNNs (EfficientNetB5, Xception), individual transformers (ViT, CrossViT) and video-based models. These findings demonstrate that hybrid architectures not only achieve the highest intra-dataset performance but also generalize most effectively to unseen datasets.

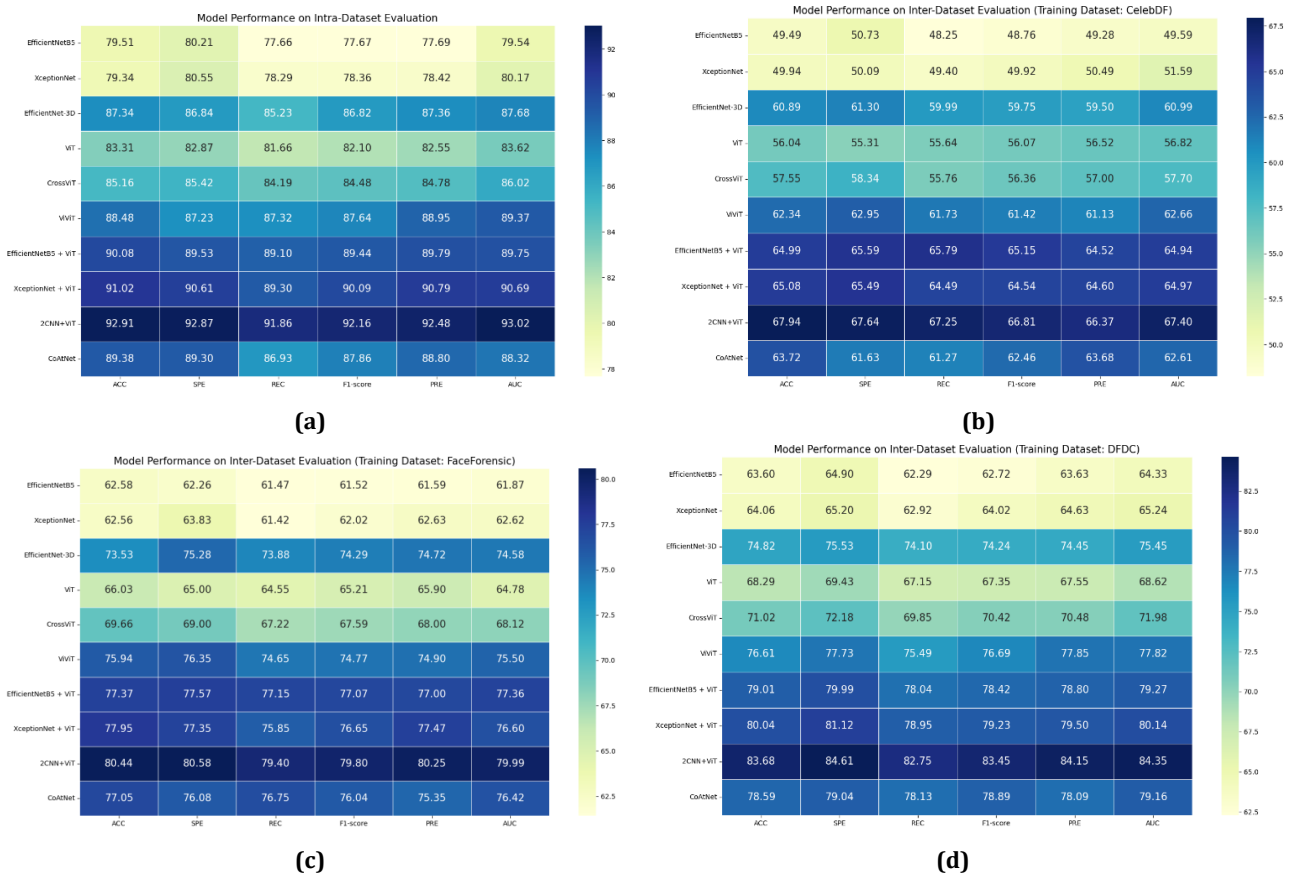

Figure 4. Average performance comparison of deep learning networks evaluated in inter-dataset and intra-dataset setting.

In addition, the comparison with previous deepfake detection methods in **Table 7** further reinforces the superiority of the hybrid model. The 2CNN+ViT approach shows robust and consistent performance across multiple datasets, achieving 92.63% on FaceForensics, 95.70% on CelebDF, and 90.40% on DFDC. This model consistently outperforms several earlier methods, emphasizing its strength in leveraging both local and global feature representations to enhance detection accuracy and its capacity to generalize effectively across various deepfake detection tasks.

Table 7. Accuracy comparison of deepfake detection methods on the FaceForensics, CelebDF, and DFDC datasets.

Approach	FF	CelebDF	DFDC
Rössler et al. [16]	66.69%	-	-
Echizen et al. [18]	61.22%	-	-

Table 7. *Cont.*

Approach	FF	CelebDF	DFDC
Ciftci et al. [22]	94.65%	91.50%	-
Coccomini et al. [28]	80%	-	-
Ramadhani et al. [30]	90.27%	87.17%	88.03%
Trisha et al. [54]	-	-	82.52%
Saikia et al. [55]	91.21%	79.49%	66.26%
Nguyen et al. [56]	91.77%	-	-
Hu et al. [57]	86.61%	80.74%	-
2CNN + ViT	92.63%	95.70%	90.40%

6. Conclusion

In this study, we investigated the effectiveness of several deep learning architectures for detecting manipulated media generated through deepfake techniques. Seven models EfficientNetB5, XceptionNet, EfficientNet-3D, Vision Transformer (ViT), CrossViT, ViViT, and CoAtNet were fine-tuned and evaluated using three widely used datasets: FaceForensics++, DeepFake Detection Challenge (DFDC), and CelebDF. Our experimental results demonstrated that the fusion of Convolutional Neural Networks (CNNs) and Transformer-based architectures exhibits a strong capability to distinguish between deepfake and authentic content. Furthermore, the results revealed that models trained on the FaceForensics++ and DFDC datasets exhibit improved generalization capabilities when applied to unseen samples.

Author Contributions

Conceptualization, A.I. and N.M.; methodology, A.I. and N.M.; software, A.I.; validation, A.I. and N.M.; formal analysis, A.I.; investigation, A.I.; resources, A.I.; data curation, A.I.; writing—original draft preparation, A.I.; writing—review and editing, N.M.; visualization, A.I. and N.M.; supervision, N.M.; project administration, A.I. and N.M. Both authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable. This study does not involve human subjects.

Informed Consent Statement

Not applicable.

Data Availability Statement

The datasets used in this study including FaceForensics, DFDC, and Celeb-DF are publicly available.

Conflicts of Interest

The authors declare no conflict of interest.

AI Use Statement

The authors declare that no artificial intelligence (AI) tools were used in the preparation of this manuscript.

References

1. Nguyen, T.T.; Nguyen, Q.V.H.; Nguyen, D.T.; et al. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
2. Flattery, T.; Miller, C.B. Deepfakes and dishonesty. *Philos. Technol.* **2024**, *37*, 120. [CrossRef]

3. Salman, S.; Shamsi, J.A.; Qureshi, R. Deep fake generation and detection: Issues, challenges, and solutions. *IT Prof.* **2023**, *25*, 52–59. [CrossRef]
4. Agarwal, S.; Farid, H.; Gu, Y.; et al. Protecting world leaders against deep fakes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 16–20 June 2019.
5. Klepper, D. Creating realistic deepfakes is getting easier than ever. Fighting back may take even more AI. Available online: https://www.ap.org/news-highlights/spotlights/2025/creating-realistic-deepfakes-is-getting-easier-than-ever-fighting-back-may-take-even-more-ai/ (accessed on 12 December 2025).
6. Heidari, A.; Jafari Navimipour, N.; Dag, H.; et al. Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Min. Knowl. Discov.* **2024**, *14*, e1520. [CrossRef]
7. Malik, A.; Kuribayashi, M.; Abdullahi, S.M.; et al. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* **2022**, *10*, 18757–18775. [CrossRef]
8. Gupta, G.; Raja, K.; Gupta, M.; et al. A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics* **2024**, *13*, 95. [CrossRef]
9. Rajput, T.; Arora, B. A Systematic Review of Deepfake Detection Using Learning Techniques and Vision Transformer. In *Lecture Notes in Networks and Systems*; Springer: Singapore, 2024. [CrossRef]
10. Lu, T.; Bao, Y.; Li, L. Deepfake Video Detection Based on Improved CapsNet and Temporal–Spatial Features. *Comput. Mater. Contin.* **2023**, *75*, 715–740. [CrossRef]
11. Yang, W.; Zhou, X.; Chen, Z.; et al. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 2015–2029. [CrossRef]
12. Elhassan, A.; Al-Fawa'reh, M.; Jafar, M.T.; et al. DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning. *SoftwareX* **2022**, *19*, 101115. [CrossRef]
13. Ibnouzaher, A.; Moumkine, N. Enhanced Deepfake Detection Using a Multi-model Approach. In *Digital Technologies and Applications*; Springer: Cham, Switzerland, 2024; pp. 317–325. [CrossRef]
14. Ibnouzaher, A.; Moumkine, N. Vgg-ViT: A Framework for Deepfakes Images Detection. In *Networked Systems*; Springer: Cham, Switzerland, 2026; pp. 241–252. [CrossRef]
15. Stanciu, D.C.; Ionescu, B. Autoencoder-based Data Augmentation for Deepfake Detection. In Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation, Thessaloniki, Greece, 12–15 June 2023. [CrossRef]
16. Rössler, A.; Cozzolino, D.; Verdoliva, L.; et al. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
17. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
18. Echizen, I.; Afchar, D.; Nozick, V.; et al. MesoNet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018. [CrossRef]
19. Cheng, J.; Sabir, E.; Jaiswal, A.; et al. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *arXiv preprint* **2019**, *arXiv:1905.00582*.
20. Wu, J.; Hu, W.; Wen, Y.; et al. Skin lesion classification using densely connected convolutional networks with attention residual learning. *Sensors* **2020**, *20*, 7080. [CrossRef]
21. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
22. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**. [CrossRef]
23. Li, Y.; Yang, X.; Sun, P.; et al. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
24. Zhu, X.; Wang, H.; Fei, H.; et al. Face forgery detection by 3D decomposition. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2020. [CrossRef]
25. Dolhansky, B.; Bitton, J.; Pflaum, B.; et al. The DeepFake Detection Challenge (DFDC) dataset. *arXiv preprint* **2020**, *arXiv:2006.07397*.

26. Gully, A.; Dufour, N. Contributing data to deepfake detection research. Available online: https://research.google/blog/contributing-data-to-deepfake-detection-research/ (accessed on 12 December 2025).
27. Wang, J.; Wu, Z.; Ouyang, W.; et al. M2TR: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022. [CrossRef]
28. Coccomini, D.A.; Messina, N.; Gennaro, C.; et al. Combining EfficientNet and Vision Transformers for video deepfake detection. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2022. [CrossRef]
29. Zhao, C.; Wang, C.; Hu, G.; et al. ISTVT: Interpretable spatial-temporal video transformer for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1335–1348. [CrossRef]
30. Ramadhani, K.N.; Munir, R.; Utama, N.P. Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention. *IEEE Access* **2024**, *12*, 8932–8939. [CrossRef]
31. Face detection guide for Python. Available online: https://ai.google.dev/edge/mediapipe/solutions/vision/face_detector/python (accessed on 4 January 2023).
32. MarekKowalski/FaceSwap. Available online: https://github.com/MarekKowalski/FaceSwap/ (accessed on 16 February 2025).
33. Thies, J.; Zollhöfer, M.; Stamminger, M.; et al. Demo of Face2Face: Real-time face capture and reenactment of RGB videos. In Proceedings of the ACM SIGGRAPH 2016 Emerging Technologies, Anaheim, CA, USA, 24–28 July 2016. [CrossRef]
34. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* **2019**, *38*, 66. [CrossRef]
35. Deepfakes/faceswap. Available online: https://github.com/deepfakes/faceswap (accessed on 16 February 2025).
36. Melnik, A.; Miasayedzenkau, M.; Makaravets, D.; et al. Face generation and editing with StyleGAN: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3557–3576. [CrossRef]
37. Nirkin, Y.; Keller, Y.; Hassner, T. FSGAN: Subject-agnostic face swapping and reenactment. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019. [CrossRef]
38. Doukas, M.C.; Ververas, E.; Sharmanska, V.; et al. Free-HeadGAN: Neural talking head synthesis with explicit gaze control. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 9743–9756. [CrossRef]
39. Das, M.K.; Kumar, M.; Kapil, I.K.; et al. Deepfake creation using GANs and autoencoder and deepfake detection. In Proceedings of the 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), Vellore, India, 5–6 May 2023. [CrossRef]
40. Reinhard, E.; Ashikhmin, M.; Gooch, B.; et al. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 36–43. [CrossRef]
41. Ma, L.; Jia, X.; Sun, Q.; et al. Pose Guided Person Image Generation. *arXiv preprint* **2017**, *arXiv:1705.09368*.
42. Szegedy, C.; Liu, W.; Jia, Y.; et al. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015. [CrossRef]
43. Fei, H.; Zhu, X.; Zhang, B.; et al. Face forgery detection by 3D decomposition and composition search. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8342–8357. [CrossRef]
44. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, Long Beach, CA, USA, 9–15 June 2019.
45. Noor, A.; Benjdira, B.; Ammar, A.; et al. DriftNet: Aggressive driving behavior classification using 3D EfficientNet architecture. *arXiv preprint* **2020**, *arXiv:2004.11970*.
46. Zheng, B.; Gao, A.; Huang, X.; et al. A modified 3D EfficientNet for the classification of Alzheimer's disease using structural magnetic resonance images. *IET Image Process.* **2023**, *17*, 77–87. [CrossRef]
47. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
48. Khan, S.; Naseer, M.; Hayat, M.; et al. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 200. [CrossRef]
49. Chen, C.F.; Fan, Q.; Panda, R. CrossViT: Cross-attention multi-scale vision transformer for image classification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
50. Wang, Y.; Deng, Y.; Zheng, Y.; et al. Vision transformers for image classification: A comparative survey. *Technologies* **2025**, *13*, 32. [CrossRef]

51. Arnab, A.; Dehghani, M.; Heigold, M.; et al. ViViT: A video vision transformer. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
52. Dai, Z.; Liu, H.; Le, Q.V.; et al. CoAtNet: Marrying convolution and attention for all data sizes. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online, 6–14 December 2021.
53. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*. [CrossRef]
54. Trisha, M.; Bhattacharya, U.; Chandra, R.; et al. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020. [CrossRef]
55. Saikia, P.; Dholaria, D.; Yadav, P.; et al. A hybrid CNN-LSTM model for video deepfake detection by leveraging optical flow features. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022. [CrossRef]
56. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019. [CrossRef]
57. Hu, J.; Liao, X.; Wang, W.; et al. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1089–1102. [CrossRef]