Article:

Alattas, E., Clark, J., Alsulami, B. et al. (2026) Beyond frames: 3D-CoAtNet for generalizable deepfake video detection. IEEE Access, 14. pp. 29692-29705. ISSN: 2169-3536

https://doi.org/10.1109/access.2026.3666623

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 January 2026, accepted 18 February 2026, date of publication 20 February 2026, date of current version 26 February 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3666623

RESEARCH ARTICLE

Beyond Frames: 3D-CoAtNet for Generalizable Deepfake Video Detection

EMAN ALATTAS[1,2], JOHN CLARK[2], (Member, IEEE), BASSMA ALSULAMI[1], AND SALMA KAMMOUN JARRAYA[3]

[1]Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]School of Computer Science, The University of Sheffield, S1 4DP Sheffield, U.K.
[3]Education and Advancement, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

Corresponding author: Eman Alattas (ealattas@kau.edu.sa)

This work was supported by the General Authority for Defence Development (GADD) in Saudi Arabia under Grant GADD_2024_01_0204.

ABSTRACT Deepfakes pose a growing risk to digital integrity and public trust, driving the need for robust video-level forgery-detection methods. Many existing approaches analyse individual frames independently and overlook temporal dependencies, thereby weakening the generalisation to unseen manipulation techniques. This paper introduces 3D-CoAtNet, a spatiotemporal architecture for deepfake video detection that processes multiple frames simultaneously, thereby reducing reliance on single-frame artefacts. The model inflates CoAtNet's 2D convolutional, residual, pooling, and self-attention layers into their 3D counterparts to learn spatial and temporal representations from multiple frames. We evaluated two input modalities: RGB 15-frame clips sampled from each video, and 15-frame optical-flow sequences that capture motion cues. Extensive experiments on FaceForensics++ (FF++), DFDC, and Celeb-DF under intra- and cross-dataset settings show that 3D-CoAtNet is competitive in intra-dataset evaluations (best in the DeepFakes dataset) and transfers well to Celeb-DF. Moreover, although frame-based CoAtNet16A achieves strong within-dataset accuracy, 3D-CoAtNet improves cross-dataset generalisation. These findings highlight the importance of the proposed 3D-CoAtNet model for deepfake forensics.

INDEX TERMS Convolutional neural networks (CNNs), CoAtNet, deepfake detection, digital forensics, generative adversarial networks (GANs), vision transformers (ViTs).

I. INTRODUCTION

A deepfake refers to synthetic content created using deep learning that can be indistinguishable from real materials by human viewers [1]. The name combines "deep" from deep learning and "fake." It includes hyper-realistic images, speech, and videos created using methods like Generative Adversarial Networks (GANs) [2]. Deepfakes can seriously undermine public trust and even enable fraud, social manipulation, or identity theft [3].

Images contain two distinct types of features that convey different kinds of information: local and global features. Local features relate to small clusters of pixels, while global features cover the entire image [4]. Convolutional Neural Networks (CNNs) are good at learning local image details, but their limited receptive fields restrict their ability to capture the spatial relationships between pixels. Essentially, CNN models focus primarily on the activated section of the face, ignoring other areas. As a result, CNNs struggle to identify and use the connections between different parts of images; for instance, they cannot notice an unnatural link between the mouth and eyes. Additionally, CNNs have issues with overfitting and cannot generalise well to unseen fake videos during training or to various types of deepfake generation methods [5].

Various deep learning architectures with CNNs have been used for deepfake detection, such as MesoNet [6], XceptionNet [7], Capsule Networks [8], and EfficientNet [9].

In 2021, Google introduced the Vision Transformer (ViT) model [10]. This type of neural network, called a

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang.

transformer, is designed to understand the context and meaning of sequential data. It uses attention or self-attention mechanisms to find connections between elements, even if they are far apart. Before transformers were developed, training neural networks needed large, labelled datasets, which are known to be resource-intensive to create. However, transformers avoid this need by mathematically finding patterns among elements. Also, using transformer theory allows for parallel processing, enabling these models to run quickly [11]. In addition, transformers can identify long-term relationships between video frames and can be scaled to manage very complex models on large datasets [12].

Vision Transformers (ViTs) provide two main advantages over Convolutional Neural Networks (CNNs). First, they use input-adaptive weighting, which means their attention weights are flexible and can change depending on the input. This is different from the fixed, input-independent convolution kernels. Second, ViTs have a global receptive field. This allows them to see the whole image at once, a feature that CNNs generally do not have, as mentioned earlier [13].

However, transformers have some drawbacks. One problem is that image attention networks often struggle with translational invariance. This means their effectiveness can change when objects in an image are moved or shifted. Consequently, for Vision Transformers (ViTs) to surpass Convolutional Neural Networks (CNNs), they need to be trained on extensive datasets containing hundreds of millions of images [13]. When ViT-based models are trained with inadequate data, their performance is inferior to CNNs, and they do not generalise effectively. Additionally, ViT-based models emphasise global features and do not perform as well as CNNs when it comes to local features [14].

Recent studies have investigated hybrid models that combine various types of Transformers with CNNs to combine the advantages of both models [15], [16], [17], [18]. Hybrid models that combine CNN and ViT architectures appear to be the most effective for learning both local texture and global semantic inconsistencies.

CoAtNet is a hybrid model that combines Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs). The original CoAtNet formulation was developed and benchmarked primarily for 2D image recognition, not for spatiotemporal video modelling [13]. This creates a mismatch in video forensic tasks where temporal dynamics are often critical.

In this study, we generalise CoAtNet from a 2D image classifier to a 3D spatiotemporal architecture (3D-CoAtNet) for video classification. Specifically, we extracted multiple frames per video and fed them jointly to a 3D variant. We then evaluated the proposed 3D-CoAtNet for deepfake video detection under both intra- and cross-dataset conditions. We train on FaceForensics++ (FF++) and test on DFDC and Celeb-DF, three widely used and progressively more challenging benchmarks for deepfake forensics.

TABLE 1. Comparison between CNN, ViT and CoAtNet based on [13].

Properties	CNN	ViT	CoAtNet
Translation Equivariance	✓		✓
Local Features	✓		✓
Input-adaptive Weighting		✓	✓
Global Features		✓	✓

The key contributions of this research can be outlined as follows:

- Generalising the 2D image classifier into a 3D architecture (3D-CoAtNet).
- Determination of effective high-level architectural parameters for 3D-CoAtNet, including the use of high-profile pretrained models and the number of head channels. Evaluation uses 15 random frames per video. We evaluate the generalisation ability of the 3D-CoAtNet model in deepfake videos.
- Development and evaluation of deepfake detection models (incorporating the effective choices determined above) over multi-frame inputs. These models are evaluated using both 15 random frames and 15 temporal frames (which capture differences in successive frames) in both intra-dataset and cross-dataset contexts (addressing the ''generalisation problem'').

The remainder of this paper is organised as follows: Section II reviews related research on the CoAtNet model and its uses for deepfake detection. Section III describes the proposed model (3D-CoAtNet). Section IV describes the proposed methodology. Section V reports the results of 3D-CoAtNet for deepfake detection. Section VI presents a comparison of performance assessments in both intra-dataset and cross-dataset scenarios. Finally, Section VII outlines the findings, discusses the study's limitations, and suggests avenues for future research.

II. BACKGROUND AND RELATED WORK

A. CoAtNet MODEL

CoAtNet [13], a contraction of Convolution and Self-Attention Network and commonly pronounced ''coat net'', was introduced in late 2021. It is a hybrid architecture that integrates Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs). This design explicitly aims to leverage the complementary advantages of both paradigms by combining convolutional inductive biases with transformer-based global attention, as shown in TABLE 1.

The CoAtNet model enhances the model's ability to generalise, its capacity, and its efficiency [13]. Generalisation in a model refers to its capability to maintain a level of performance relative to unseen data, which is similar to that of

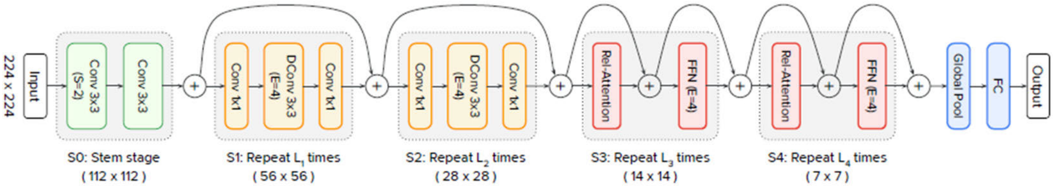


FIGURE 1. **CoAtNet architecture [13].**

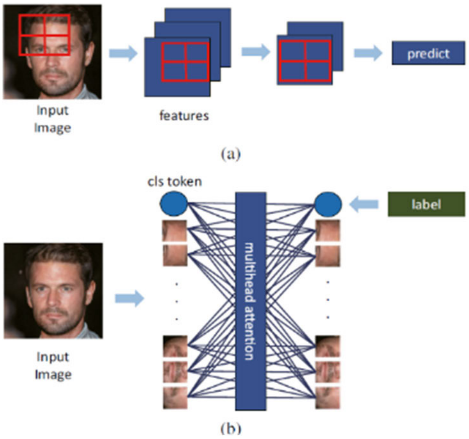


FIGURE 2. **Difference between (a) CNN and (b) Vision Transformer (ViT)in Face Image Processing [16].**

the training data. On the other hand, model capacity pertains to the model's ability to handle extensive training datasets. A model with greater capacity can achieve better final performance outcomes after sufficient training.

FIGURE 1 outlines the CoAtNet architecture [13], organised into five stages (S0–S4) with the pattern S0, then C–C–T–T, where C denotes the convolution, and T denotes the transformer. Stage S0 is a two-layer convolutional stem that is used for early dimensionality reduction. Stages S1 and S2 are convolutional, built from Mobile Inverted Bottleneck Convolution (MBConv) blocks that use depthwise convolution and Squeeze–Excitation (SE) to shrink spatial resolution before handing features to global attention stages. Stages S3 and S4 are transformer stages comprising relative self-attention, followed by a feedforward network (FFN). The network concludes with global pooling and a final fully connected layer.

B. RATIONALE FOR USING THE CoAtNet MODEL

Different deepfake generation strategies change different facial proportions and locations. For example, some techniques only change small areas, such as the mouth region, which reveals itself in colour mismatches in lips, while others change larger areas, such as face borders, in face-swapping techniques. In addition to the importance of local features, global features provide crucial information about the extent and intensity of manipulations. Therefore, these data can be added to the detection process to improve the performance of the deepfake detection algorithm [19]. Therefore, both local and global features are essential for deepfake detection.

There is a difference in the image processing and feature extraction between CNNs and Vision Transformers. In a CNN, the features are reduced progressively using the CNN kernel, which moves over the input image. Finally, the features are converted into a single feature used to classify the image, as shown in FIGURE *2*.

For the deepfake detection problem, CNN uses the facial image's partial features to search for a complete face to detect anomalous features and predict fake images. On the other hand, in Vision Transformers, there is a Cls token that is connected to all patch features to identify closely related components. The patches that are strongly related to class tokens appear as active areas, which is essential for detecting fake images. The CoAtNet model was selected to address these limitations because it combines both CNN and ViT architectures.

To the best of our knowledge, only two recent studies [20] [21] have leveraged the CoAtNet-type architecture for the deepfake detection problem.

The authors in [20] proposed a bagging ensemble approach that leverages CoAtNet for deepfake detection. The model employs CutMix data augmentation and integrates predictions from multiple CoAtNet learners by using majority voting, sum, or product rules. The authors evaluated CoAtNet only in an intra-dataset setting (training and testing in the same dataset) on FF++ [7] and Celeb-DF [22] without examining cross-dataset generalisation or broader generalisation performance. Their approach was frame-based with no explicit temporal modelling.

In [21], the authors adapted CoAtNet to deepfake detection in a frame-based manner: predictions are made for each of 15 randomly sampled frames, and also for each of 15 consecutive optical-flow frames. Optical-flow frames, also known as 'temporal frames', are visual representations of the estimated motion field between two consecutive video frames, where each pixel encodes the displacement vector (direction and magnitude) of apparent pixel movement, allowing models to explicitly capture temporal dynamics and improve recognition of actions or events beyond static appearance) [23]. During training, each frame or temporal frame is considered individually. In testing (after training is complete), 15 frames

TABLE 2. Mapping 2D CoAtNet components to their counterparts in our 3D-CoAtNet architecture.

Components in CoAtNet Architecture	Purpose	Corresponding Components in 3D-CoAtNet Architecture	Part of
LayerNormChannels	Layer normalisation is a technique used to normalise the activations of a layer of neurons in a neural network. Unlike batch normalisation, which normalises the activations of a batch of inputs, layer normalisation normalises the activations of each individual example in the batch. Overall, layer normalisation is a helpful technique for improving the training and generalisation performance of deep neural networks (DNNs). [25]	LayerNormChannels3D	TransformerBlock3D, Head3D
Residual	In the forward method, the residual block takes an input tensor x and passes it through the shortcut path and the residual path. The output of the shortcut path is added to the output of the residual path multiplied by the gamma parameter (trainable scaling parameter). [26]	Residual3D	MBConv3D, TransformerBlock3D
Stem	Early convolutional layers that downsample and extract low-level features before deeper conv/attention stages; provides strong inductive bias and stable training for subsequent Transformer blocks. [13]	Stem3D	CoAtNet3D
ConvBlock	This represents a convolutional block in a neural network. The ConvBlock class inherits from nn. Sequential, which means it is a sequence of PyTorch modules that will be executed in order.[27]	ConvBlock3D	MBConv3D, Stem3D
get_shortcut	This represents the shortcut connection in a residual block of a convolutional neural network. The shortcut connection is a way to bypass the convolutional layers in a residual block and directly connect the input to the output. This is done to facilitate the flow of gradients through the network during training, and to help the network learn identity mappings. [26]	get_shortcut3d	MBConv3D, Transformer TransformerBlock3D
SqueezeExciteBlock(SE)	Early convolutional layers that downsample and extract low-level features before deeper conv/attention stages; provides strong inductive bias and stable training for subsequent Transformer blocks.[28]	SqueezeExciteBlock3D	MBConv3D
MBConv	Uses an inverted residual bottleneck with depthwise convolutions and SE to achieve efficient feature extraction with low FLOPs while preserving accuracy. [29]	MBConv3D	CoAtNet3D
positional encodings	Positional encoding in CoAtNet provides the attention layers with explicit location information, capturing both spatial and temporal positions [30].	3D positional encodings	SelfAttention3D
SelfAttention2d	Models long-range dependencies via multi-head self-attention to capture global context that complements local convolutions. [30].	SelfAttention3D	Transformer3D
FeedForward	Position-wise (per token/patch) MLP that expands then projects features (e.g., GELU + linear layers), applied identically at each position [31]	FeedForward3D	Transformer3D
TransformerBlock	A stack of (LayerNorm → Multi-Head Self-Attention → residual) and (LayerNorm → FeedForward → residual) sublayers [30]	TransformerBlock3D	CoAtNet3D
Head	Classification head: global pooling over spatial (and temporal) dimensions followed by a linear classifier (often with dropout), produces final logits. [13]	Head3D	CoAtNet3D
BlockStack	Organises multiple MBConv or Transformer blocks into stages with controlled depth and resolution changes, enabling scalable capacity and hierarchical features.[13]	BlockStack3D	CoAtNet3D
CoAtNet	A hybrid architecture that combines convolution (inductive bias/efficiency) with attention (global context/capacity), stacking conv then attention stages yields strong generalisation and scalability. [13]	CoAtNet3D	Model

or 15 consecutive temporal frames for a specific video are subject to individual prediction, and the results are combined by a voting technique to reach the video level prediction. The authors examined the generalisation performance of their deepfake detection approach by training on the FF++ dataset [7] and testing on unseen data in the Celeb-DF [22] and DFDC [24] datasets.

III. PROPOSED MODEL: 3D-CoAtNet MODEL

Unlike conventional detection approaches that rely heavily on single-frame artefacts and post-hoc score averaging, 3D-CoAtNet adopts 3D spatial and temporal attention modelling, in which features are jointly extracted across both spatial and temporal dimensions. Features are derived by inflating 2D-CoAtNet into 3D, replacing 2D convolutions,

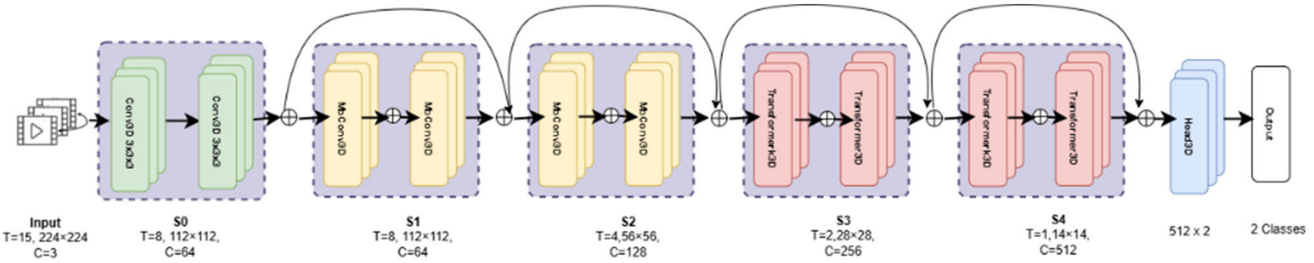
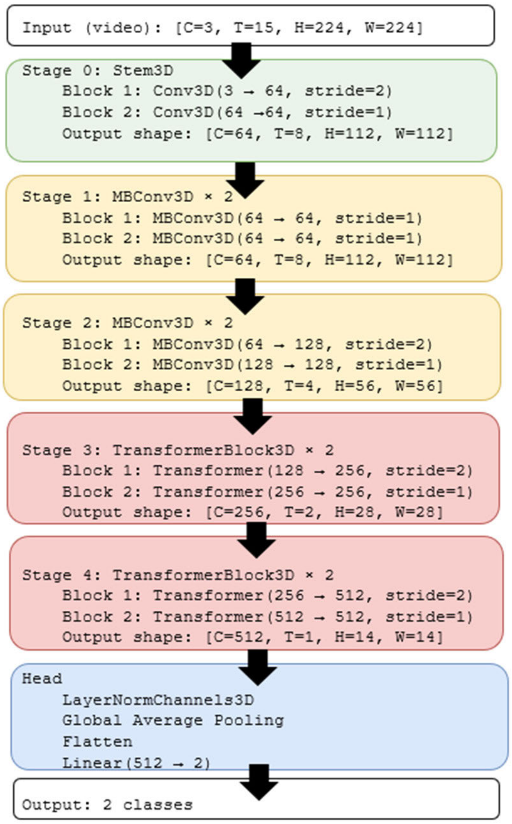


FIGURE 3. 3D-CoAtNet architecture.



```
Input (video): [C=3, T=15, H=224, W=224]

Stage 0: Stem3D
    Block 1: Conv3D(3 → 64, stride=2)
    Block 2: Conv3D(64 →64, stride=1)
    Output shape: [C=64, T=8, H=112, W=112]

Stage 1: MBConv3D × 2
    Block 1: MBConv3D(64 → 64, stride=1)
    Block 2: MBConv3D(64 → 64, stride=1)
    Output shape: [C=64, T=8, H=112, W=112]

Stage 2: MBConv3D × 2
    Block 1: MBConv3D(64 → 128, stride=2)
    Block 2: MBConv3D(128 → 128, stride=1)
    Output shape: [C=128, T=4, H=56, W=56]

Stage 3: TransformerBlock3D × 2
    Block 1: Transformer(128 → 256, stride=2)
    Block 2: Transformer(256 → 256, stride=1)
    Output shape: [C=256, T=2, H=28, W=28]

Stage 4: TransformerBlock3D × 2
    Block 1: Transformer(256 → 512, stride=2)
    Block 2: Transformer(512 → 512, stride=1)
    Output shape: [C=512, T=1, H=14, W=14]

Head
    LayerNormChannels3D
    Global Average Pooling
    Flatten
    Linear(512 → 2)

Output: 2 classes
```

FIGURE 4. 3D-CoAtNet detailed architecture.

pooling, residual blocks, and self-attention layers with their 3D counterparts, ensuring that the network learns dynamic temporal dependencies directly.

We build a 5-stage 3D-CoAtNet that follows CoAtNet's principle of stacking convolutional stages followed by transformer-style attention stages, but with all operators jointly promoted to 3D to model space and time. Specifically, the network is:

- Stem3D: two 3D convolution components.
- Convolutional stages (MBConv3D): two depthwise-separable 3D blocks with Squeeze-and-Excitation and residual connections.
- Attention stages (TransformerBlock3D): two stages Self-Attention3D, 3D positional encodings, and FeedForward3D.
- Head3D: LayerNorm over channels, global 3D average pooling, dropout, and linear classifier for two classes.

TABLE 2 clarifies the detailed components of 2D-CoAtNet, the purpose of each component, and the corresponding components in 3D-CoAtNet. FIGURE 4 shows the architecture of the 3D-CoAtNet model, while FIGURE *3* shows more details of the model.

IV. METHODOLOGY

A. PROPOSED MODEL

Our approach has two distinct stages. In the first, "Experimental Settings", we investigated a range of parameters to determine effective settings. These preliminary experiments served as the foundation for identifying configurations that yielded acceptable performance in terms of the Area Under the Curve (AUC) of the 3D-CoAtNet model.

In the first stage, we examined two parameters: (1) loading pretrained weights and (2) modifying the number of heads in the 3D-CoAtNet model. For the first parameter, we compared three models. The first model had no pretrained weights, the second was initialised with VGG16 weights [32], and the third was initialised with X3D model weights [33]. For the second parameter, we tested configurations with 8, 16, and 32 heads. The multi-head attention mechanism allows each head to engage with different segments of the input data. This helps the model to capture a variety of complex relationships within the input [34]. At this stage, 15 frames were extracted from each video of the DF dataset and used as input instances for training a detection model.

The second stage, "Model Application", uses the best performing settings identified in stage one to train the model on both 15 random frames and 15 consecutive optical flow frames. Consistent with prior video-based deepfake detectors, which typically use 8-32 frames per video, several works adopt 15 or 16 frames as an efficient balance between capturing enough video content and computational cost. For example, GenConViT [35] uses 15 frames per video and [36] uses 16 frames per video. During training, all 15 frames from each video are processed together as a single input to the model. The same applies when using 15 consecutive optical flow frames. During testing, the frames belonging to each video are also combined to generate a final prediction for that video.

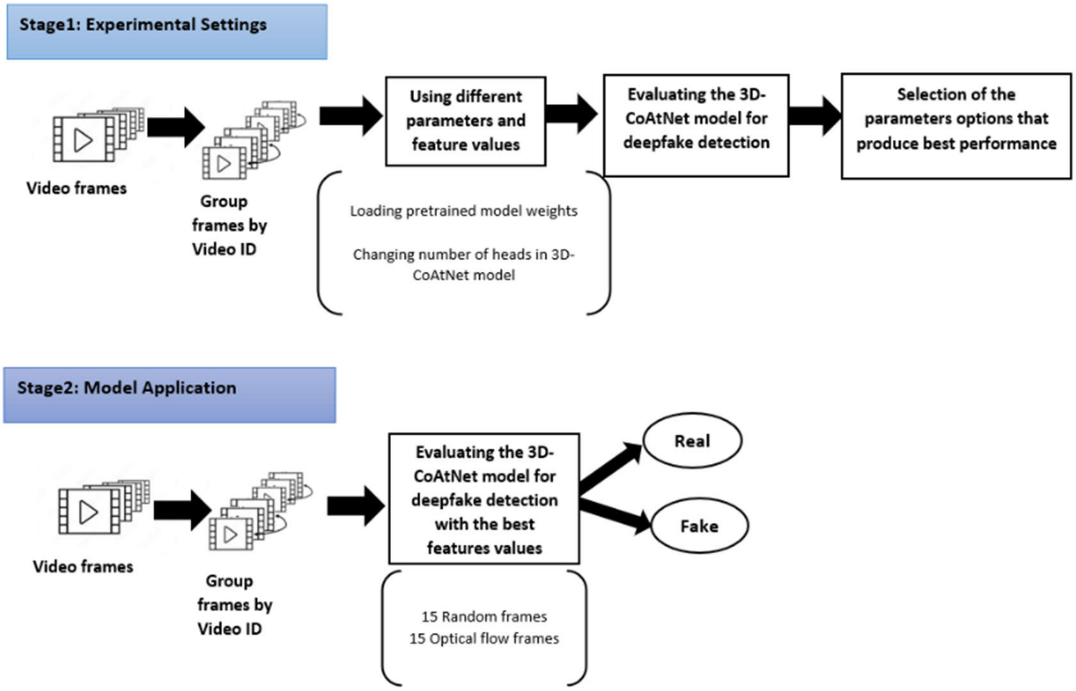

Stage1: Experimental Settings
Video frames
Group frames by Video ID
Using different parameters and feature values
Evaluating the 3D-CoAtNet model for deepfake detection
Selection of the parameters options that produce best performance
Loading pretrained model weights
Changing number of heads in 3D-CoAtNet model
Stage2: Model Application
Video frames
Group frames by Video ID
Evaluating the 3D-CoAtNet model for deepfake detection with the best features values
Real
Fake
15 Random frames
15 Optical flow frames

FIGURE 5. Proposed model using 3D-CoAtNet architecture.

FIGURE 5 illustrates the details of the proposed approach for exploring the 3D-CoAtNet model with combined frames as input, grouped by video ID. In the second stage, the model was trained on four categories of FF++: DF, F2F, FS, and NT. Weighted fusion was then used to evaluate the model on FF++.

Weighted fusion is preferred over a single unified detector because different manipulation-specific models capture complementary artefacts, and combining them improves generalisation to unseen distributions [37]. Weighted fusion assigns each model a weight proportional to its AUC, giving more influence to stronger models. The method converts model probabilities into logits, computes a weighted sum of these logits, and then applies a sigmoid to obtain the final fused probability. This approach ensures that high-performing models contribute more to the final combined model of the four pretrained models for each FF++ category, while weaker models have a reduced influence.

B. DATASETS

This study uses the following publicly available deepfake datasets: FaceForensics++ [7] (raw version), DFDC [24] and Celeb-DF [22] to ensure robust evaluation across diverse manipulation techniques and real-world scenarios. These datasets vary in their manipulated-face generation methods, which aids in assessing model generalisation in both intra-dataset and cross-dataset settings. Using them also allows for direct comparisons with previous studies.

The FF++ and Celeb-DF datasets were divided into training, validation, and testing sets using a 70%, 15%, and 15% split, respectively. This division follows the data-splitting strategy adopted in several prior deepfake-detection studies [38], [39]. The DFDC dataset is provided with predefined training, validation, and testing sets.

C. PREPROCESSING

The frames are extracted from videos, then in each frame, the face is extracted using dlib [40] and resized to 224 × 224 pixels. Face alignment was not used. For the 15 RGB frames, 15 random frames were selected, while for 15 optical flow frames, 16 consecutive frames were extracted, then 15 corresponding consecutive optical flow frames were extracted using Farnebäck [41] algorithm and saved the frame at size 224 × 224 pixels.

D. IMPLEMENTATION DETAILS

The following parameters have been employed in the experiments. The initial learning rate was set to 1×10^{-4}, and the batch size was 8. The AdamW optimiser was employed to train the model for 50 epochs. Training was conducted on an NVIDIA A100 Tensor Core GPU provided by the Aziz Supercomputer at the Center of Excellence in High-Performance Computing [42].

TABLE 3. Average AUC scores of the 3D-CoAtNet model with different pretrained initialisations.

Model	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUC Cross-dataset	AVG AUC All DS
Without weight loading	0.897	0.647	0.539	0.696	0.588	0.617	0.695	0.603	0.664
VGG16	0.861	0.635	0.518	0.660	0.574	0.644	0.669	0.609	0.649
X3D	0.925	0.701	0.503	0.701	0.569	0.672	**0.708**	**0.621**	**0.679**

TABLE 4. Average AUC scores of the 3D-CoAtNet model with different numbers of head channels.

Model	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUC Cross-dataset	AVG AUC All DS
8	0.962	0.706	0.460	0.724	0.588	0.704	**0.713**	**0.646**	**0.691**
16	0.943	0.642	0.453	0.669	0.593	0.692	0.677	0.642	0.665
32	0.950	0.660	0.434	0.696	0.580	0.654	0.685	0.617	0.662

E. COMPUTATIONAL COMPLEXITY

We evaluate the computational cost of the proposed 3D-CoAtNet using an input video of 15 frames at 224 × 224 resolution, with tensor shape B × C × T × H × W = 8 × 3 × 15 × 224 × 224. We report model size in terms of trainable parameters and checkpoint size, theoretical compute in number of floating-point operations (FLOPs). The proposed 3D-CoAtNet contains 11.45 million trainable parameters, corresponding to a checkpoint size of 43.78 MB.

FLOPs are computed using the fvcore library [43] by tracing a single forward pass of the network given a fixed-size input tensor, and reporting the total number of floating-point operations required for inference per sample(video). For a forward pass with batch size B = 8, the total number of floating-point operations is 230.312 GFLOPs, which corresponds to approximately 28.79 GFLOPs per video sample.

V. RESULTS

This section presents the results of the two-stage 3D-CoAtNet approach, including experimental settings and model applications.

A. RESULTS OF STAGE 1

In stage one, two parameters were investigated: the loading of the pretrained weights and the number of heads.

The results in TABLE 3 show that the X3D initialisation achieved the highest average AUC across all settings, with 0.708 for FF++ datasets, 0.621 for cross-dataset evaluation, and 0.679 overall. Training without weight loading produced slightly lower averages (0.695, 0.603, and 0.664), whereas the VGG16 initialisation yielded the weakest results (0.669, 0.609, and 0.645). These findings indicate that video-based pretraining (X3D) provides better results than both random initialisation and image-based pretraining.

The results in TABLE 4 indicate that increasing the number of attention heads from 8 to 16 and 32 does not yield performance gains; instead, a consistent degradation appears across most datasets and aggregated AUC metrics. The 8-head configuration achieves the strongest results, with the highest AVG AUC on FF++ (0.713), cross-dataset evaluation (0.646), and the overall average (0.691), suggesting that additional heads introduce unnecessary model complexity without improving feature discrimination. In contrast, 16 and 32 heads show reduced performance, particularly in F2F, FS, and Celeb-DF. The result aligns with prior transformer analyses that more heads do not necessarily improve the performance [44].

B. RESULTS OF STAGE 2

In stage 2, the 3D-CoATNet model with X3D pretrained weights and 8 heads was applied to 15 frames and 15 optical flow frames. As shown in TABLE 5, the 3D-CoAtNet model achieved an average AUC of 0.770 on the FF++ dataset, outperforming individual dataset training and indicating reliable within-dataset detection. For the cross-dataset evaluation, the model reached an average AUC of 0.656. Combining all the datasets, the model produced an overall average AUC of 0.732.

TABLE 6 presents the results of the 3D-CoAtNet model using the 15 optical flow frames. On the FF++ dataset, it attains an average AUC of 0.812, which is substantially

TABLE 5. Performance (AUC) of the 3D-CoAtNet model using 15 random frames.

Trained on/ Tested on	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUV Cross-dataset	AVG AUC All DS
DF	0.944	0.701	0.536	0.704	0.638	0.68	0.721	0.659	0.701
F2F	0.745	0.77	0.62	0.688	0.502	0.543	0.706	0.523	0.645
FS	0.597	0.66	0.717	0.64	0.481	0.471	0.654	0.476	0.594
NT	0.788	0.732	0.539	0.728	0.594	0.641	0.697	0.618	0.670
Weighted fusion	0.921	0.781	0.642	0.735	0.636	0.676	0.770	0.656	0.732

TABLE 6. Performance (AUC) of the 3D-CoAtNet model using 15 optical flow frames.

Trained on/ Tested on	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUV Cross-dataset	AVG AUC All DS
DF	0.871	0.559	0.68	0.628	0.537	0.601	0.685	0.569	0.646
F2F	0.708	0.707	0.648	0.737	0.538	0.595	0.700	0.567	0.656
FS	0.782	0.596	0.895	0.607	0.545	0.645	0.720	0.595	0.678
NT	0.708	0.654	0.637	0.765	0.509	0.566	0.691	0.538	0.640
Weighted fusion	0.894	0.68	0.896	0.778	0.554	0.681	0.812	0.618	0.747

TABLE 7. Performance (AUC) of the 3D-CoAtNet model using combined 15 random frames (RF) and 15 optical flow frames (OF).

Model	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUV Cross-dataset	AVG AUC All DS
3D-CoAtNet, 15 RF	0.921	0.781	0.642	0.735	0.636	0.676	0.770	0.656	0.732
3D-CoAtNet, 15 OF	0.894	0.680	0.896	0.778	0.554	0.681	0.812	0.618	0.747
3D-CoAtNet, Fused 15 RF+15 OF	0.934	0.782	0.885	0.810	0.623	0.722	0.853	0.672	0.793

higher than the 0.685–0.720 range obtained with single-dataset training. For the cross-dataset evaluation, it reached an average AUC of 0.618, outperforming the 0.538–0.595 range observed with individual training setups.

Then, a combined metric of 15 random frames and 15 optical flow frames was calculated using weighted fusion to enhance the evaluation for both the intra-dataset and the cross-dataset. As shown in TABLE 7 and FIGURE 6, the combined evaluation of 15 random frames and 15 optical flow frames achieved the highest intra-dataset performance with an average AUC of 0.853, which is 8.3 percentage points higher than the 0.770 obtained using only random frames, and 4.1 percentage points higher than the 0.812 achieved with only optical flow. Similarly, in the cross-dataset evaluation, the combined model provides an improvement of 1.6 percentage points (0.672 vs. 0.656) compared to random frames, and 5.4 percentage points compared to optical flow (0.618). When considering all datasets, the combined approach maintained the strongest overall average AUC of 0.793, representing an improvement of 6.1 percentage points over random frames (0.732) and 4.6 percentage points over optical flow (0.747). These results confirm that the fusion results of the 3D-CoAtNet model trained on random frames and optical flow frames provide a more generalisable representation, thereby enhancing the performance for both within-dataset and unseen datasets.

VI. PERFORMANCE EVALUATION COMPARISON

For comparison purposes, the proposed 3D-CoAtNet model is first compared with CoAtNet16A [21] and subsequently with state-of-the-art studies.

A. COMPARISON WITH CoAtNet16A

1) COMPARING 15 RANDOM FRAMES BETWEEN CoAtNet16A AND 3D-CoAtNet

As shown in TABLE 8 and FIGURE 7, CoATNet16A achieved a near-perfect average AUC of 0.997 on the FF++ datasets, which is 22.7 percentage points higher than the 0.770 obtained by the proposed 3D-CoAtNet. However, in the cross-dataset evaluation, 3D-CoAtNet provided an improvement of 12 percentage points (0.656 vs. 0.536), demonstrating stronger generalisation to unseen data. When considering all datasets, CoATNet16A maintained a higher overall average AUC of 0.843, which was 11.1 percentage points greater than the 0.732 achieved by 3D-CoAtNet.

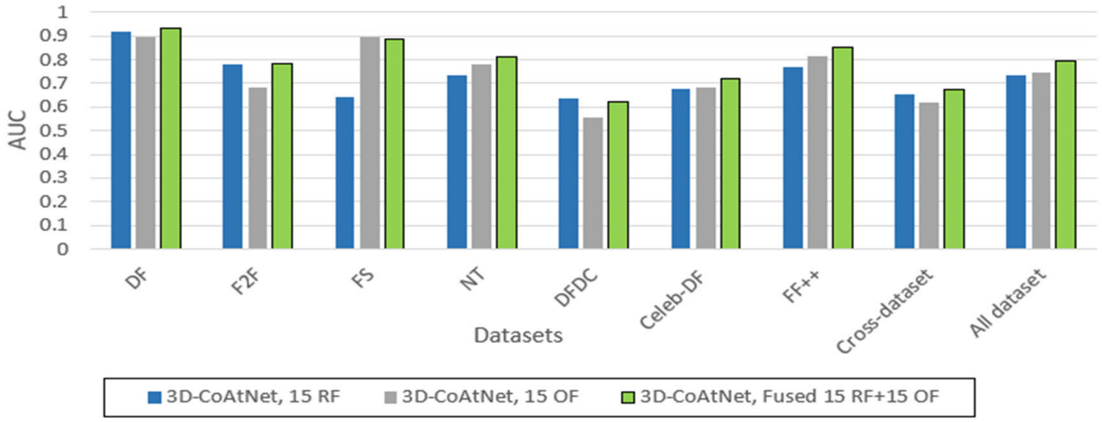


FIGURE 6. 3D-CoAtNet with 15 Random Frames (RF) and 15 Optical Flow (OF) Fusion.

TABLE 8. 15 Random frames (RF) comparison of the CoAtNet16A and 3D-CoAtNet models.

Model	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUV Cross-dataset	AVG AUC All DS
CoATNet16A- 15 RF	0.9988	0.996	0.9955	0.9957	0.5547	0.5167	**0.997**	0.536	**0.843**
3D-CoAtNet- 15RF	0.9210	0.7810	0.6420	0.7350	0.6360	0.6760	0.770	**0.656**	0.732

These results highlight the trade-off between the two models: CoATNet16A is superior in within-dataset detection, whereas 3D-CoAtNet sacrifices some in-domain performance to achieve better cross-dataset robustness.

2) COMPARING 15 OPTICAL FLOW FRAMES BETWEEN CoAtNet16A AND 3D-CoAtNet

As shown in TABLE *9* and FIGURE *8*, the CoATNet16A model achieved a stronger overall performance than 3D-CoATNet when using 15 optical flow frames. On the FF++ datasets, CoATNet16A recorded an average AUC of 0.962, which is 15 percentage points higher than the 0.812 obtained by 3D-CoAtNet. For the DFDC dataset, the results are almost the same in both models, whereas in the Celeb-DF dataset, CoAtNet16A provides an improvement of 6.8 percentage points (0.749 vs. 0.681). These results indicate that, unlike in the random-frame setting, CoAtNet16A surpasses 3D-CoAtNet in both intra-dataset detection and cross-dataset robustness when optical flow frames are utilised.

B. COMPARISON WITH RELATED STUDIES

In this section, we present a comparison between the proposed 3D-CoAtNet model, which fuses 15 random RGB frames with 15 optical-flow frames per video, and 12 state-of-the-art deepfake detection methods across six datasets (DF, F2F, FS, NT, DFDC, and Celeb-DF). The evaluation compares the performance metric (AUC) and reports the absolute difference between the performance of the two models in percentage points (pp).

In our experiments, we divided the FF++ and Celeb-DF datasets into 70% training, 15% validation, and 15% testing sets. In contrast, the authors of [45] randomly sampled the datasets. For DFDC, [45] evaluated the DFDC-Preview subset using approximately 500 randomly selected videos, whereas we used approximately 5,000 videos, that is, the full test set of the DFDC dataset. A similar difference holds for Celeb-DF: [45] tested on approximately 500 videos, whereas we used approximately twice as many videos. In addition, [45] used a larger input resolution of 256 × 256 pixels. Consequently, these differences in splitting strategy, sample size, and input resolution indicate that any cross-study comparison is only an approximate rather than a precise like-for-like evaluation.

As shown in TABLE *10*, intra-dataset evaluation of the four FF++ categories (DF, F2F, FS, and NT) showed a mixed pattern. 3D-CoAtNet attains a DF of 0.934, edging the strongest baseline (F3Net, 0.9335) by +0.05 percentage points (pp). On F2F, 3D-CoAtNet trails the best baseline (Dynamic-Difference Learning) by −13.88 pp (0.7823 vs. 0.9211). In contrast, for FS, the gap to the best result (Dynamic-Difference Learning) is only −3.06 pp (0.8845 vs. 0.9151). Finally, on NT, 3D-CoAtNet and Dynamic-Difference Learning were nearly tied to achieve the best performance (0.8100 vs. 0.8119).

TABLE 9. 15 Optical flow (OF) frames comparison of the CoAtNet16A and 3D-CoAtNet models.

Model	DF	F2F	FS	NT	DFDC	Celeb-DF	AVG AUC FF++	AVG AUV Cross-dataset	AVG AUC All DS
CoATNet16A- 15 OF	0.985	0.954	0.979	0.929	0.545	0.749	**0.962**	**0.647**	**0.857**
3D-CoAtNet- 15 OF	0.894	0.680	0.896	0.778	0.554	0.681	0.812	0.618	0.747

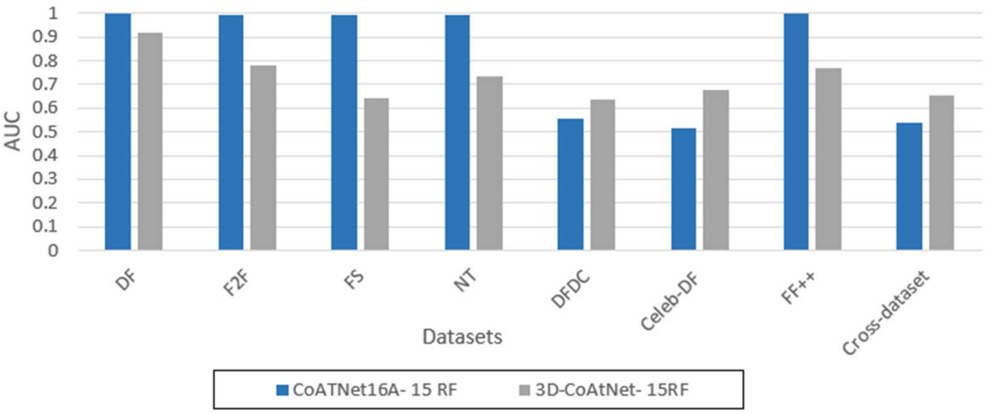


FIGURE 7. 15 Random Frames (RF) for CoAtNet16A and 3D-CoAtNet.

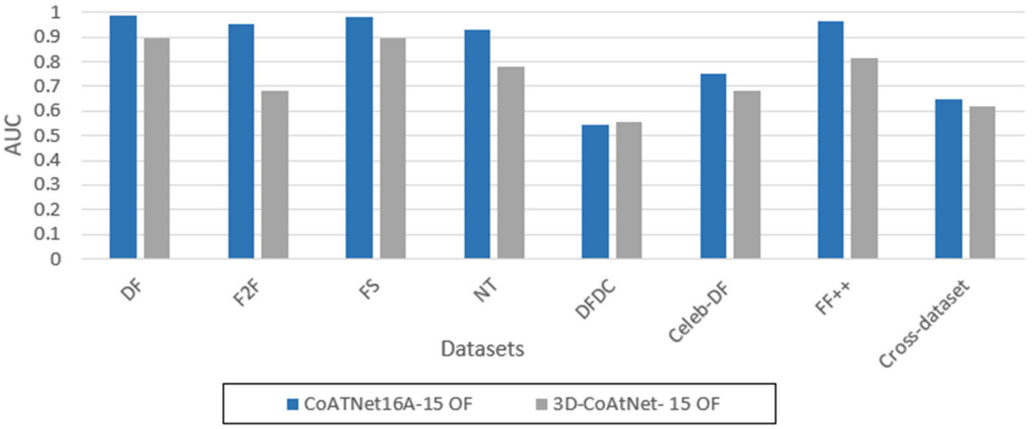


FIGURE 8. 15 Optical Flow Frames (OF) for CoAtNet16A and 3D-CoAtNet.

For cross-dataset transfer to DFDC, 3D-CoAtNet achieved 0.6229, which is 10.79 percentage points (pp) below the top baseline, Dynamic-Difference Learning (0.7308). However, our model was evaluated on the DFDC dataset, whereas Dynamic-Difference Learning [45] reported the results on the DFDC Preview set.

Conversely, 3D-CoAtNet achieved the strongest performance on the Celeb-DF dataset. It achieves 0.7216, outperforming the strongest baseline (Dynamic-Difference Learning, 0.7136) by +0.80 pp. This highlights the 3D-CoATNet's ability to capture temporal-spatial cues that are transferred more effectively to Celeb-DF.

Overall, 3D-CoAtNet is a competitive within-dataset evaluation (best on DF with +0.05 pp) and strong in cross-dataset generalisation to Celeb-DF (+0.80 pp); however, it underperforms the best baseline on F2F (–13.88 pp), FS (–3.06 pp), and NT (–0.19 pp). Taken together, these results suggest that, while 3D-CoAtNet provides measurable gains

TABLE 10. Performance comparisons between 3D-CoAtNet and related studies. The training dataset is FF++. The related studies' results are reported in [45].

Models	DF	F2F	FS	NT	DFDC	Celeb-DF
ShallowNet [46]	0.855	0.7487	0.7523	0.6002	0.5048	0.531
Mesonet [6]	0.8784	0.7988	0.7739	0.638	0.624	0.5837
Xception [47]	0.9042	0.8689	0.836	0.7309	0.5826	0.6171
F3Net [48]	**0.9335**	0.9192	0.9112	0.807	0.6596	0.6204
Multi-att [49]	0.9246	0.9109	0.8986	0.8094	0.6747	0.595
GFF [50]	0.8879	0.873	0.8595	0.7735	0.6599	0.6246
CNN+LSTM [51]	0.7669	0.6549	0.5951	0.6394	0.556	0.5242
DBIRNN [52]	0.9152	0.8665	0.8753	0.6849	0.6263	0.6241
CLRNet [53]	0.9231	0.8545	0.8574	0.6816	0.5994	0.6371
ConvLSTM [54]	0.9323	0.857	0.773	0.7405	0.5251	0.6667
DIL [55]	0.7676	0.676	0.6226	0.5996	0.5401	0.5631
Dynamic-difference learning [45]	0.9298	**0.9211**	**0.9151**	**0.8119**	**0.7308**	0.7136
3D-CoAtNet (rank from 13)	**0.934 (1)**	0.7823(10)	0.8845(4)	0.8100(2)	0.6229(7)	**0.7216(1)**

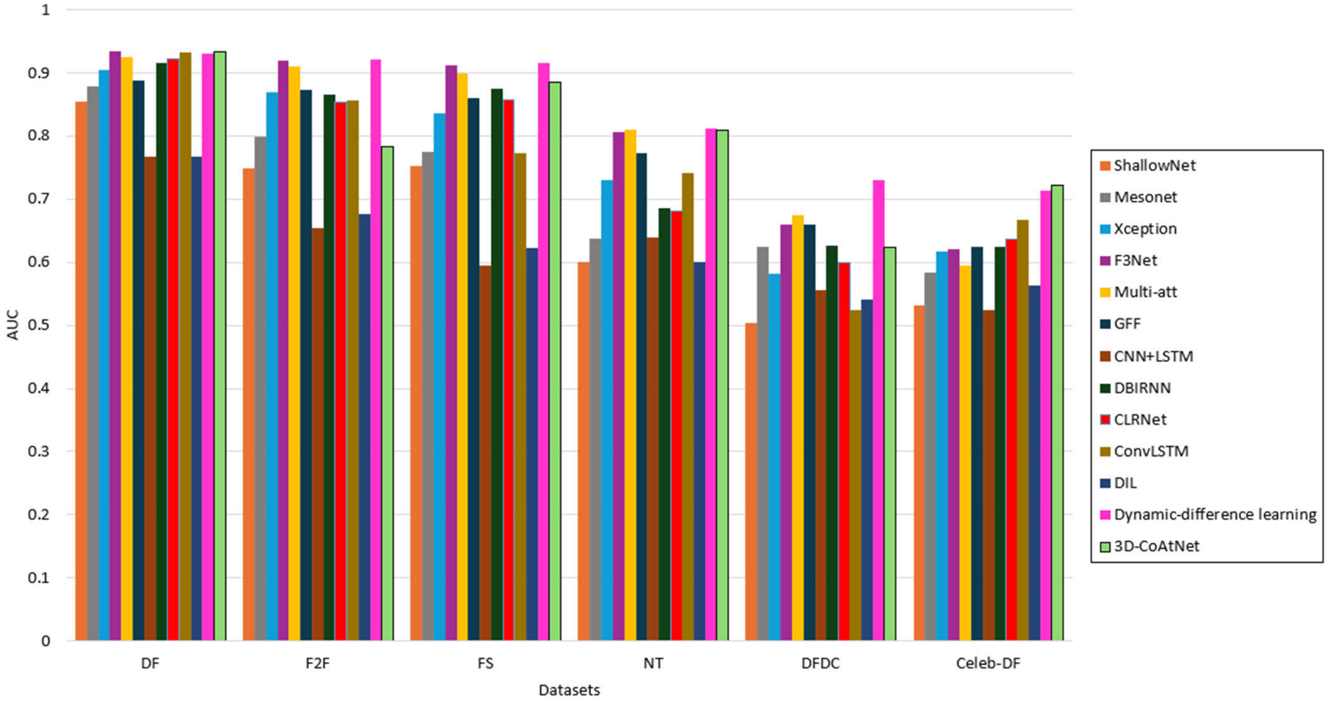


FIGURE 9. AUC scores of the 3D-CoATNet model compared to state-of-the-art deepfake detectors.

on certain manipulations and on challenging cross-dataset transfer to Celeb-DF, further refinement may be required to close the remaining gaps. FIGURE 9 presents a comparison of the performance of the proposed 3D-CoAtNet model with existing baseline methods across various datasets.

VII. DISCUSSION AND LIMITATIONS

The present study demonstrates that extending CoAtNet into a 3D architecture (3D-CoAtNet) enables more effective modelling of spatiotemporal cues in deepfake videos, producing competitive within-dataset performance and improved cross-dataset generalisation compared with 2D CoAtNet variants. The results show that X3D-based video pretraining and 8-head configuration yield effective settings for 3D-CoAtNet, outperforming both random initialisation and image-based pretraining and avoiding the degradation observed with 16 or 32 heads. Moreover, combining 15 RGB frames with 15 optical-flow frames via weighted fusion produced the highest overall AUC (0.793), confirming that multi-frame temporal modelling substantially enhances generalisability across unseen datasets.

Only two prior studies have leveraged CoAtNet architectures for deepfake detection [20], [21], and both operate exclusively on 2D representations. However, the present work therefore offers the first generalisation of CoAtNet into a 3D architecture designed to learn spatiotemporal features. Relative to [21], which aggregated predictions from 15 RGB or optical-flow frames via voting, our study shows that treating multi-frame sequences as unified video volumes - not independent frames - yields more stable performance across datasets. Whereas [21] reported strong intra-dataset detection but weaker cross-dataset robustness, the present work shows that 3D-CoAtNet narrows this gap, particularly through its fused random frames and optical flow frames pipeline. This divergence likely stems from the architectural difference: 3D-CoAtNet learns joint temporal–spatial patterns, whereas [21] relies on voting of single-frame decisions.

This study introduces the first 3D extension of the CoAtNet architecture, demonstrating that CoAtNet's hybrid convolution–attention structure can be successfully generalised to video-based deepfake detection. Practically, these findings highlight the importance of stacking multiple frames into a 3D model for deepfake detection systems. The deepfake detection systems can be deployed in real-world environments, where unseen manipulations and distribution shifts are common. Fusing random frames and optical flow frames as input to 3D-CoAtNet provides a more stable tool for identifying manipulations across heterogeneous datasets.

Despite its strengths, several limitations must be acknowledged. First, although 3D-CoAtNet improves cross-dataset generalisation relative to 2D CoAtNet variants, its absolute performance on some manipulations (e.g., F2F, FS, DFDC) remains below the best published baselines. Second, the study evaluates only 15-frame segments, motivated by computational tractability. Longer temporal windows or dynamic frame-sampling strategies should be explored.

VIII. CONCLUSION

This study introduces 3D-CoAtNet, a spatiotemporal extension of the CoAtNet backbone designed to improve the generalisation of deepfake video detection. Building upon frame-based CoAtNet [21], 3D-CoAtNet eliminates reliance on single-frame artefacts. The core contribution of this work lies in inflating 2D-CoAtNet into a 3D architecture, replacing all convolutional, residual, pooling, and self-attention operators with their 3D counterparts, thereby enabling the direct learning of temporal dynamics essential for video forensics.

This paper provided a comprehensive description of the 3D-CoAtNet framework, covering its architectural design and experimental configurations, including pretrained weight sources and attention-head configuration. The results identified X3D-based video pretraining and an 8-head design as the most effective foundations for robust detection. Building on these findings, the work further developed multi-frame detection models using 15 random frames, 15 optical-flow frames, and a fused representation that jointly captures spatial appearance and motion cues.

Across six datasets, the fused 3D-CoAtNet model achieved competitive intra-dataset performance and demonstrated improved generalisation to unseen datasets, particularly Celeb-DF. These results highlight the importance of incorporating temporal information directly within the model architecture, rather than relying solely on frame-based decision fusion.

Future research should focus on redesigning the internal architecture of 3D-CoAtNet and undertaking comprehensive ablation studies to better understand how individual architectural components influence intra- and cross-dataset generalisation and computational efficiency. Another possible direction is the exploration of multimodal integration, such as incorporating audio alongside visual information, to enhance detection robustness and mitigate reliance on visual artefacts. In addition, extending the application of 3D-CoAtNet to broader video classification problems may provide valuable insights into its transferability beyond deepfake detection.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review," *WIREs Data Mining Knowl. Discovery*, vol. 14, no. 2, Mar. 2024, Art. no. e1520, doi: 10.1002/widm.1520.

[2] M. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble-based learning technique for deepfake detection," in *Proc. 7th IEEE Int. Conf. Cyber Secur. Cloud Comput. (CSCloud)/ 6th IEEE Int. Conf. Edge Comput. Scalable Cloud (EdgeCom)*, Aug. 2020, pp. 70–75, doi: 10.1109/CSCloud-EdgeCom49738.2020.00021.

[3] S. Singh and A. Dhumane, "Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges," *MethodsX*, vol. 15, Dec. 2025, Art. no. 103632, doi: 10.1016/j.mex.2025.103632.

[4] L. Kabbai, M. Abdellaoui, and A. Douik, "Image classification by combining local and global features," *Vis. Comput.*, vol. 35, no. 5, pp. 679–693, May 2019, doi: 10.1007/s00371-018-1503-0.

[5] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 1038–1048, Aug. 2020, doi: 10.1109/JSTSP.2020.3007250.

[6] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS)*, Sep. 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630761.

[7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11, doi: 10.1109/ICCV.2019.00009.

[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2307–2311.

[9] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5012–5019, doi: 10.1109/ICPR48806.2021.9412711.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[11] R. Merritt. (2022). *What Is a Transformer Model*. Accessed: Jul. 14, 2022. [Online]. Available: https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

[12] S. A. Khan and H. Dai, "Video transformer for deepfake detection with incremental learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1821–1828, doi: 10.1145/3474085.3475332.

[13] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 3965–3977.

[14] B. Kaddar, S. A. Fezza, W. Hamidouche, Z. Akhtar, and A. Hadid, "HCiT: Deepfake video detection using a hybrid model of CNN features and vision transformer," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5, doi: 10.1109/VCIP53242.2021.9675402.

[15] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining EfficientNet and vision transformers for video deepfake detection," in *Proc. Int. Conf. Image Anal. Process.*, 2022, pp. 219–229, doi: 10.1007/978-3-031-06433-3_19.

[16] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "DeepFake detection algorithm based on improved vision transformer," *Int. J. Speech Technol.*, vol. 53, no. 7, pp. 7512–7527, Apr. 2023, doi: 10.1007/s10489-022-03867-9.

[17] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[18] S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, "ViXNet: Vision transformer with xception network for deepfakes based video and image forgery detection," *Expert Syst. Appl.*, vol. 210, Dec. 2022, Art. no. 118423, doi: 10.1016/j.eswa.2022.118423.

[19] A. Khormali and J.-S. Yuan, "DFDT: An end-to-end DeepFake detection framework using vision transformer," *Appl. Sci.*, vol. 12, no. 6, p. 2953, Mar. 2022, doi: 10.3390/app12062953.

[20] K. Omar, R. H. Sakr, and M. F. Alrahmawy, "An ensemble of CNNs with self-attention mechanism for DeepFake video detection," *Neural Comput. Appl.*, vol. 36, no. 6, pp. 2749–2765, Feb. 2024, doi: 10.1007/s00521-023-09196-3.

[21] E. Alattas, J. Clark, A. Al-Aama, and S. K. Jarraya, "Evaluating features and variations in deepfake videos using the CoAtNet model," *J. Imag.*, vol. 11, no. 6, pp. 1–28, Jun. 2025, doi: 10.3390/jimaging11060194.

[22] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for DeepFake forensics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213, doi: 10.1109/CVPR42600.2020.00327.

[23] A. Chintha, A. Rao, S. Sohrawardi, K. Bhatt, M. Wright, and R. Ptucha, "Leveraging edges and optical flow on faces for deepfake detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10, doi: 10.1109/IJCB48548.2020.9304936.

[24] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake detection challenge (DFDC) dataset," 2020, *arXiv:2006.07397*.

[25] J. Lei Ba, J. Ryan Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell. AAAI*, vol. 31, 2017, pp. 4278–4284, doi: 10.1609/aaai.v31i1.11231.

[27] C. Wang, "A review on 3D convolutional neural network," in *Proc. IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2023, pp. 1204–1208, doi: 10.1109/ICPECA56706.2023.10075760.

[28] Y. Peng, X. Li, and Y. Wang, "Quantum squeeze-and-excitation networks," in *Proc. IEEE Int. Conf. Quantum Comput. Eng. (QCE)*, Sep. 2024, pp. 39–43, doi: 10.1109/QCE60285.2024.10249.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008, doi: 10.1109/2943.974352.

[31] S. Herath. (Apr. 19, 2024). *The Feedforward Network (FFN) in the Transformer Model*. Accessed: Aug. 13, 2025. [Online]. Available: https://medium.com/image-processing-with-python/the-feedforward-network-ffn-in-the-transformer-model-6bb6e0ff18db

[32] (2025). *Torchvision Main Documentation*. Accessed: Feb. 13, 2025. [Online]. Available: https://pytorch.org/vision/main/models/generated/torchvision.models.vgg16.html

[33] *X3D Weights*. Accessed: Mar. 15, 2025. [Online]. Available: https://dl.fbaipublicfiles.com/pytorchvideo/model_zoo/kinetics/X3D_S.pyth

[34] Z. Bing, L. Li, and J. Liang, "Optimizing knowledge distillation in transformers: Enabling multi-head attention without alignment barriers," 2025, *arXiv:2502.07436*.

[35] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, "GenConViT?: Deepfake video detection using generative convolutional vision transformer," *Appl. Sci.*, vol. 15, no. 12, p. 6622, 2025.

[36] R. Sun, Z. Zhao, L. Shen, Z. Zeng, Y. Li, B. Veeravalli, and Y. Xulei, "An efficient deep video model for deepfake detection," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2023, pp. 351–355, doi: 10.1109/ICIP49359.2023.10222682.

[37] S. Concas, S. M. La Cava, G. Orrù, C. Cuccu, J. Gao, X. Feng, G. L. Marcialis, and F. Roli, "Analysis of score-level fusion rules for deepfake detection," *Appl. Sci.*, vol. 12, no. 15, p. 7365, Jul. 2022, doi: 10.3390/app12157365.

[38] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, "An enhanced deep learning-based DeepFake video detection and classification system," *Electronics*, vol. 12, no. 1, p. 87, Dec. 2022, doi: 10.3390/electronics12010087.

[39] D. C. Subhashree and M. Nikitha, "Detecting deepfake faces with CNN and LSTM," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 9, pp. 448–454, Sep. 2025, doi: 10.22214/ijraset.2025.74079.

[40] D. Zhang, J. Li, and Z. Shan, "Implementation of dlib deep learning face recognition technology," in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, Nov. 2020, pp. 88–91, doi: 10.1109/ICRIS52159.2020.00030.

[41] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Image Analysis: 13th Scandin. Conf.*, 2003, pp. 363–370.

[42] (2023). *Center of Excellence in High Performance Computing Center*. Accessed: Oct. 29, 2023. [Online]. Available: https://hpc.kau.edu.sa/

[43] (2026). *Fvcore Documentation-Detectron2 0.6 Documentation*. Accessed: Jan. 22, 2026. [Online]. Available: https://detectron2.readthedocs.io/en/latest/modules/fvcore.html

[44] Q. Zhao, X. Zhang, F. Wang, P. Fan, and E. Mbeka, "The effect of the head number for multi-head self-attention in remaining useful life prediction of rolling bearing and interpretability," *Neurocomputing*, vol. 616, Feb. 2025, Art. no. 128946, doi: 10.1016/j.neucom.2024.128946.

[45] Q. Yin, W. Lu, B. Li, and J. Huang, "Dynamic difference learning with spatio-temporal correlation for deepfake video detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 4046–4058, 2023, doi: 10.1109/TIFS.2023.3290752.

[46] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "GAN is a friend or foe?: A framework to detect various fake face images," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 1296–1303, doi: 10.1145/3297280.3297410.

[47] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[48] Y.-Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Proc. ECCV*, 2020, pp. 86–103, doi: 10.1007/978-3-030-58610-2_6.

[49] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2185–2194, doi: 10.1109/CVPR46437.2021.00222.

[50] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16312–16321, doi: 10.1109/CVPR46437.2021.01605.

[51] S. Tariq, S. Lee, and S. S. Woo, "A convolutional LSTM based residual network for deepfake video detection," 2020, *arXiv:2009.07480*.

[52] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2274–2283.

[53] S. Tariq, S. Lee, and S. Woo, "One detector to rule them all: Towards a general deepfake attack detection framework," in *Proc. Web Conf.*, Jun. 2021, pp. 3625–3637, doi: 10.1145/3442381.3449809.

[54] B. Chen, T. Li, and W. Ding, "Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM," *Inf. Sci.*, vol. 601, pp. 58–70, Jul. 2022, doi: 10.1016/j.ins.2022.04.014.

[55] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma, "Delving into the local: Dynamic inconsistency learning for DeepFake video detection," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 1, pp. 744–752, doi: 10.1609/aaai.v36i1.19955.

EMAN ALATTAS received the B.Sc. and master's degrees in computer science from King Saud University, Riyadh, in 2006 and 2011, respectively. She is currently pursuing the Ph.D. degree with The University of Sheffield, U.K. She is a Lecturer with King Abdulaziz University, Jeddah, Saudi Arabia. Her research interests include artificial intelligence, machine learning, deep learning, computer vision, and security.

JOHN CLARK (Member, IEEE) received the M.A. degree in mathematics from the University of Oxford, in 1985, the M.Sc. degree in applied statistics, Oxford, in 1986, and the Ph.D. degree in computer science from the University of York, in 2001. He was with the Computer Security Division, Logica, in 1987. In 1992, he joined the University of York. In 2005, he was a Professor of critical systems. Since 2017, he has been a Professor of computer and information security with The University of Sheffield, U.K., where he leads the Security of Advanced Systems Research Group. His research interests include high-integrity systems, safety, reliability, security, and the use of non-standard computation and machine learning, including quantum approaches. He is a Professional Member of the British Computer Society (BCS). He has received multiple best paper awards, medals in Human Competitive Evolutionary Computation, and held a Royal Society Wolfson Research Merit Award. He has chaired and served on numerous EPSRC panels, managed over £7M in research funding as a Principal Investigator, and supervised around 40 Ph.D. students.

BASSMA ALSULAMI received the Ph.D. degree in computer science from Howard University, USA, with a specializing in cybersecurity and wireless networking. She is currently an Associate Professor of computer science. Her research interests include cybersecurity, machine learning, and the application of AI in education, with multiple publications in international journals and conferences.

SALMA KAMMOUN JARRAYA received the Ph.D. degree in computer science from Sfax University, Tunisia. From 2010 to 2014, she was with the University of Sfax, Tunisia, where she is also a member of the Mir@cl Laboratory. She was working as an Associate Professor with the Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University. She is currently working with the Education and Advancement Department, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Her research interests include computer vision, video and image processing, as well as data mining, and knowledge discovery in images and video. She was a member of many international projects. She served on the technical conference committees and a reviewer in many international conferences and journals.