RESEARCH

Deepfake Detection Using Spatiotemporal Transformer

Bachir Kaddar[1*], Sid Ahmed Fezza[2], Zahid Akhtar[3], Wassim Hamidouche[4], Abdenour Hadid[5] and Joan Serra-Sagristà[6]

Abstract

Recent advances in generative models and the availability of large-scale benchmarks have made the process of deepfake video generation and manipulation easier. Nowadays, the number of new hyper-realistic deepfake videos used for negative purposes is dramatically increasing, thus creating the need for effective deepfake detection methods. Although many existing deepfake detection approaches, in particular CNN-based methods, show promising results, they suffer from a number of drawbacks. In general, poor generalization results have been obtained under unseen real scenes or new deepfake generation methods. The crucial reason for the above defect is that CNN-based methods focus on the spatial local artifacts, which are unique for every manipulation method. Therefore, it is hard to learn the general forgery traces of different manipulation methods without considering the dependencies that extend beyond the local receptive field. To address this problem, this paper proposes a framework which combines Convolutional Neural Network (CNN) with Vision Transformer (ViT) to improve the detection accuracy and enhance generalizability on videos with various content. Our method, called "HCiT", exploits the advantages of CNN to extract local meaningful features with the ViT's self-attention mechanism to explicitly learn discriminative global contextual dependencies in a frame-level image. In this hybrid architecture, the high-level feature maps extracted from the CNN are feeded into the ViT model that determines whether a specific video is fake or real. Experiments were performed on Faceforensics++, DeepFake Detection Challenge preview and Celeb datasets, and the results show that the proposed method significantly outperforms the state-of-the-art methods. In addition, the HCiT method shows a great capacity for generalization on datasets covering various techniques of deepfake generation. We then present a detailed ablation study to investigate the effectiveness of the pure ViT and the key components of our hybrid architecture in the deepfake detection setting.

Keywords: DeepFake video; detection; convolutional neural network; vision transformer; hybrid

1 Introduction

The development of new deep generative models, such as Autoencoders (AEs) [1] or Generative Adversarial Networks (GANs) [2], in addition to the availability and free access to a large amount of public datasets [3–5] have made the process of creating convincing fake videos, known as *deepfakes*, easier and faster. As a result, the deepfake technology has been significantly improved and the number of new spoofed video content increases dramatically, making it more difficult to distinguish between real and synthesized videos, even by a human observer [6, 7].

Generally, the methods to generate deepfake videos fall into four main categories [8]: (*i*) reenactment, (*ii*) swapping,(*iii*) editing and (*iiii*) synthesis. In face reenactment [9], the fake video is obtained by synthesizing the facial expressions and movements of the person in the target video using the face of person in the source video. Different from face reenactment technologies, in face swapping [10], or face replacement, the face of person in the target video is entirely replaced by the face of other person in the source video. An editing or enchantment deepfake is where the attributes of the target video such as facial hair, age, weight, beauty, and ethnicity, are added, altered, or removed. Finally, synthesis is where the deepfake is created with no target as a basis. It can be used to create fake personas online or generate characters for movies and games. Thus, the first two categories of methods only deal with the facial expressions manipulation, while the two latter category concerns identity manipulation. Despite their many interesting and positive applications such as film-making and virtual reality [11], the growing harms and malicious uses of deepfake technology have various negative impacts on privacy, social stability and national security [12]. Consequently, to address these challenging security concerns, in the last few years, extensive research efforts have been devoted to the development of new deepfake content detection methods [13, 14]. For instance,

*Correspondence: bachir.kaddar@univ-tiaret.dz
[1]University of Ibn Khaldoun-Tiaret, Tiaret, Algeria
Full list of author information is available at the end of the article

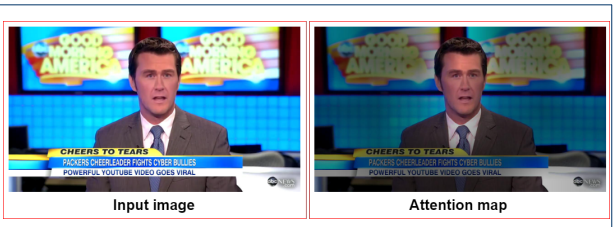


Figure 1: The proposed HCiT method for deepfake video detection. In HCiT (Convolutional Neural Network (CNN)-Vision Transformer (ViT)), the features extracted by the CNN module are used as inputs by ViT for the binary classification task.

early studies attempted to identify fake videos by detecting visual artifacts present in the first generation of fake videos, such as eye color, missing reflections, and missing details in the teeth areas [15]. However, the fast progress on AEs [1] and GANs [2] have significantly improved the quality and realism of generated fake videos [16,17], making the distortion-based methods ineffective. On the other hand, with growing availability of large-scale public deepfake datasets Faceforensics++ (FF++) [3], DeepFake Detection Challenge preview (DFDC-p) [4], Celeb-DF [5], more recent works have focused on deep learning-based approaches to automatically extract discriminating features characterizing forged content [18]. These detection methods, which are essentially based on CNNs, extract spatio-temporal features for forgery videos detection. The temporal methods mainly use Recurrent Neural Networks (RNNs) to learns from temporal sequences [19]. These methods leverage temporal inconsistencies, discrepancies and discontinuities across adjacent frames to distinguish between fake and real videos [20, 21]. For the methods based on spatial artifacts, a frame-by-frame analysis is performed using CNNs models. They can be implemented either by deep or shallow classifiers. In general, these methods have achieved very good results using popular network architectures such as VGG16 [22], ResNet50 [23] and particularly Xception [24].

Though CNN-based techniques have demonstrated good performance, most of them fail to generalize to unseen forgeries or to different classes of deepfake generation methods [25]. These methods can only detect the type of deepfake on which they are trained [26], and tend to overfit to manipulation-specific artifacts. The main reason for the above defect is that CNN-based methods focus on the spatial local artifacts, which are unique for every manipulation method [27]. Therefore, it is hard to learn the general forgery traces of different manipulation methods based only on the local context feature without considering the dependencies that extend beyond the local receptive field. At the same time, deepfake technology is getting better every day and new types of deepfake manipulations emerge quickly, where the differences between fake and real videos are becoming more subtle and fine-grained, making it difficult for existing CNN-based methods to detect them. The core for tackling the high realistic fake videos is to learn subtle yet discriminative features. Therefore, there is an urgent need to develop deepfake detection methods with a high generalization capability, which deal with as many kinds of deepfake manipulations as possible.

Recently, self-attention-based architectures, in particular transformers [28], have become the standard for natural language processing (NLP) tasks. Following this success, and over the last few years, the field of computer vision has been revolutionized by the emergence of methods based on self-attention. Among them, ViT has demonstrated impressive performance in image classification tasks [29]. ViT is the first purely self-attention-based network that achieved state-of-the-art performance in image recognition. The excellent reported results highlight the importance of using the attention mechanism for the image processing task [28]. However, despite recent progress and the great success of the ViT model, only a few ViT-based deepfake detectors have been proposed [30–32]. Therefore, it is urgently needed to investigate whether the transformer-based method works well for deepfake detection. In addition, there is still room for another ViT-based deepfake method which provides much better performance and demonstrates high generalization ability.

In this paper, we propose a hybrid architecture for deepfake detection using a concord of CNN and ViT models, named HCiT. This HCiT hybrid architecture utilizes advantages of CNN to extract local information, i.e., strengthening locality, with the advantages of transformer introducing the self-attention mechanism, which aggregates the global information from the entire input sequence. Our idea consists in introducing a self-attention mechanism on the high-level features previously learned for the task of image classification and adapting them for the deepfakes detection relying on a knowledge-transfer process. Thus, the high features maps extracted from CNN are used as input to the ViT encoder to detect whether the image is fake or real. We show that such a design allows our framework to preserve the advantages of ViT and also benefit deepfake videos detection. This CNN-ViT combination provides competitive performance and better convergence, as well as a high generalization capacity on videos with various content compared to CNNs models.

The rest of this paper is organized as follows. Section 2 describes previous works on the detection of deepfake videos. Section 3 describes in detail the proposed HCiT method. The experimental results are presented in Section 4. Finally, conclusions are given in Section 5.

2 Related work

Different approaches have been proposed in the literature to detect deepfake videos. These methods can be categorized into two main groups [33]: (*i*) handcrafted feature-based techniques and (*ii*) deep learning-based techniques.

Handcrafted feature-based techniques involve manual selection of a specific set of features to classify real and fake videos [15]. The analysis of these features, i.e., their presences/absences and intensity, allows to distinguish forged videos from real ones. In this context, the early studies involve looking for visual artifacts and inconsistent features in the first generation of fake videos. An overview of such methods can be found in [34], [35].

For instance, the authors of [15], proposed fake detectors based on simple visual artifacts such as eye color, missing reflections, and missing details in the teeth areas. Some works detect deepfakes by searching for specific artifacts with edge detectors, quality measures, and frequency analysis [36], [37], [38]. Other proposed methods are used for performing statistical analysis on the residuals of an image to detect forgery [39], [40], [41], validate noise priors from metadata [42], or learn specific manipulation traces, such as recoloring [43] or recompression [44]. [45], [39], [46] aim to detect forgeries by utilizing intrinsic statistics (e.g., frequency domain characteristics) of images. Zhang et al. [47] used the bag of words method to extract a set of compact features and fed them into various classifiers for discriminating swapped face images from the genuine.

In [48], the authors found that GANs leave unique fingerprints and show how it is possible to classify the generator given the content. Several works have also focused on the detection of these subtle features, patterns, and possible inconsistencies within RGB frames of the video ley by the GAN models [49], [3], [50]. In [51], the authors proposed a detection system based on colour features and a linear Support Vector Machine (SVM) for the final classification, achieving a final 70.0% AUC when evaluating with the NIST MFC2018 dataset [52]. However, one the approaches cited above are not robust against unseen simple image perturbation attacks such as noise, blur, cropping or compression, and the models need to be re-trained again. In [53], the authors used the facial expression in order to distinguish a fake speaking pattern from natural one. Yang et al. [54] utilized the inconsistency in head pose to detect fake video contents. In [55] and [56], the authors proposed an approach based on correlation between the speech and the landmarks around the mouth to detect deepfake videos.

However, the fast progress on GANs [2] and AEs [1] have significantly improved the quality and realism of generated fake videos [16, 17], making the distortion-based methods ineffective.

Deep learning-based techniques are data-driven approaches using large deepfake datasets in a supervised configuration [3, 18]. Often these types of approaches rely on CNN models.

In [57], [58], [3], it was shown that these fake detectors tend to perform better than traditional image forensic tools and have achieved very good results using popular network architectures such as Xception [24], [59] [50] [60] [61]. Other popular Deepfake detection approaches include Two-Stream [62], MesoNet [50], FWA [63], VA [15], Multi-task [64], capsule [65] and DSP-FWA [66].

For the deep learning-based methods, two main categories can be distinguished: (*i*) methods that employ spatio-temporal features, and (*ii*) those that are based on visual artifacts.

The first category of methods mainly use CNN combined with RNNs to take advantages of the temporal information [19]. These methods look for temporal inconsistencies and discrepancies across frames to distinguish between fake and real videos [20, 21].

For instance, Guera and Delp [21] proposed a recurrent neural network incorporating temporal information to detect Deepfake videos. A pre-trained CNN was employed to extract framelevel features, which were then fed into the LSTM to create a temporal sequence descriptor. Likewise, [20] proposed a recurrent convolutional model (RCN) based on the integration of the convolutional network DenseNet [67] and the gated recurrent unit cells to be trained end-to-end. Their proposed detection approach was evaluated on FaceForensics++ dataset achieving impressive results. In [68], the authors refined the network's focus by monitoring the frames' optical flow. In [69], the authors focused on physical characteristics and proposed to model the eye blinking related to fake faces with a CNN/RNN model to expose DeepFake videos. However, this detection technique show limitation when images with closed eyes are incorporated in training.

For the methods based on visual-spatial artifacts, a frame-by-frame analysis is performed using CNNs models. They can be implemented either by deep or shallow classifiers. In general, these methods have achieved very good results using popular network architectures such as VGG16 [22] and ResNet50 [23]. For

Table 1: Datasets for Deepfake detection : FF++, DFDC-p, and Celeb-DF.

Dataset	Real		Forged	
	Video	Frame	Video	Frame
FF++ [3]	1,000	509.9k	4,000	1,830.1k
DFDC-p [4]	1,131	488.4k	4,113	1,783.3k
Celeb-DF [5]	590	225.4k	5,639	2,116.8k

instance, Rossler *et al.* [3] used the XceptionNet model that was trained on the FF++ dataset to distinguish fake videos from the real ones. Zhou et al. proposed a two-stream neural network for this task [62]. Nguyen *et al.* [58] proposed to use the capsule networks that takes features obtained from the VGG19 network for detecting manipulated videos. The capsule network that can detect various kinds of deepfakes. In [50], Afchar *et al.* proposed MesoNet method that is a relatively shallow CNN architecture to focus on the mesoscopic properties of images. Specifically, they started with deep complex architecture and have gradually simplified it, up to the one producing the same results but more efficiently.

3 Proposed deepfake detection method

Motivated by the recent success of ViT in image classification tasks [29], it is therefore quite natural to attempt applying ViT for the deepfake detection problem. However, applying a pure ViT architecture requires a vast amount of training data and a high number of training iterations to achieve comparable performance with state-of-the-art CNNs. Transformer-based models under-perform CNNs and do not generalize well when trained on insufficient data [70]. In addition, the features learned by ViT contain less low-level information [71]. Therefore, in order to exploit the self-attention mechanism of ViT and at the same time to avoid the above limitations, the idea behind the proposed method is to consider a hybrid architecture using the CNN model in conjunction with ViT, named HCiT. In the HCiT model, a CNN block is introduced to help learning low-level features (e.g., corners and edges) combined with the ViT model which establishes long-range dependencies, i.e., aggregates global information from the entire input sequence.

As illustrated in Figure 2, the HCiT deepfake detection method is composed of three main components: (*i*) face cropping, (*ii*) features extraction backbone and (*iii*) ViT-based binary classification. Each component of the proposed HCiT method is described in the following sections.

3.1 Face cropping

Let's consider a video sequence X that consists of T frames, denoted as $X^{(t)}$, $t = 1, 2, \ldots, T$. Each $X^{(t)} \in \mathbb{R}^{H \times W \times C}$ is a 2D image, where $(H \times W)$ is the image resolution and C is the number of color channels. Before feeding the sequence to the CNN, each frame $X^{(t)}$ is first pre-processed. For this we used the widely known Dlib package [72] to perform face detection and landmark localization. Then, face alignment is performed followed by face cropping to a resolution of 224×224. The cropping step helps to avoid the contribution of background and other database biases, which can significantly increase detection accuracy by focusing only on the location of manipulations instead of using the full frame [3].

Finally, the T cropped face regions denoted as $X'^{(t)}$, $t = 1, 2, \ldots, T$, are passed to the CNN model for feature extraction.

3.2 Feature extraction backbone

As mentioned previously, instead of using raw image patches, the ViT input sequence can be formed from a CNN feature maps [29]. This is to take advantage of CNNs to extract low-level features and avoid training ViT with large amounts of data, which is computationally expensive.

Thus, after obtaining cropped face images, a CNN network, Xception network architecture [24], is used as the feature extraction backbone network. To achieve this goal, we removed the fully-connected (FC) layer at the top of the Xception network to directly generate a 2D deep representation of each cropped face image, i.e., feature maps. The output of the Xception consists of feature maps f of size $2048 \times 7 \times 7$, whose the spatial dimensions are flattened and projected to the ViT dimension.

It is important to note that the Xception network is initialized with weights pre-trained on ImageNet [73] and then fine-tuned on deepfake dataset.

3.3 ViT-based deepfake detection

In vision tasks, the locality of CNNs impairs the ability to capture long-range dependencies [74]. Therefore, to entirely dispense with the convolutional inductive bias, we rely on the self-attention mechanism of ViT model. The self-attention layers of ViT aggregate global information across the entire input sequence. This allows the model correctly focuses on more relevant parts, i.e., the discriminative features, for deepfake video detection. Thus, the feature maps resulting from the previous step, which consist of a sequence of 2048 flattened 2D patches of size 7×7, are fed into the transformer encoder, as illustrated in Figure 1. Similar to the vision transformer models [29], we append an extra learnable classification token at the beginning of the input sequence of embedded feature. In addition, we added position embeddings to the feature embedding in order to

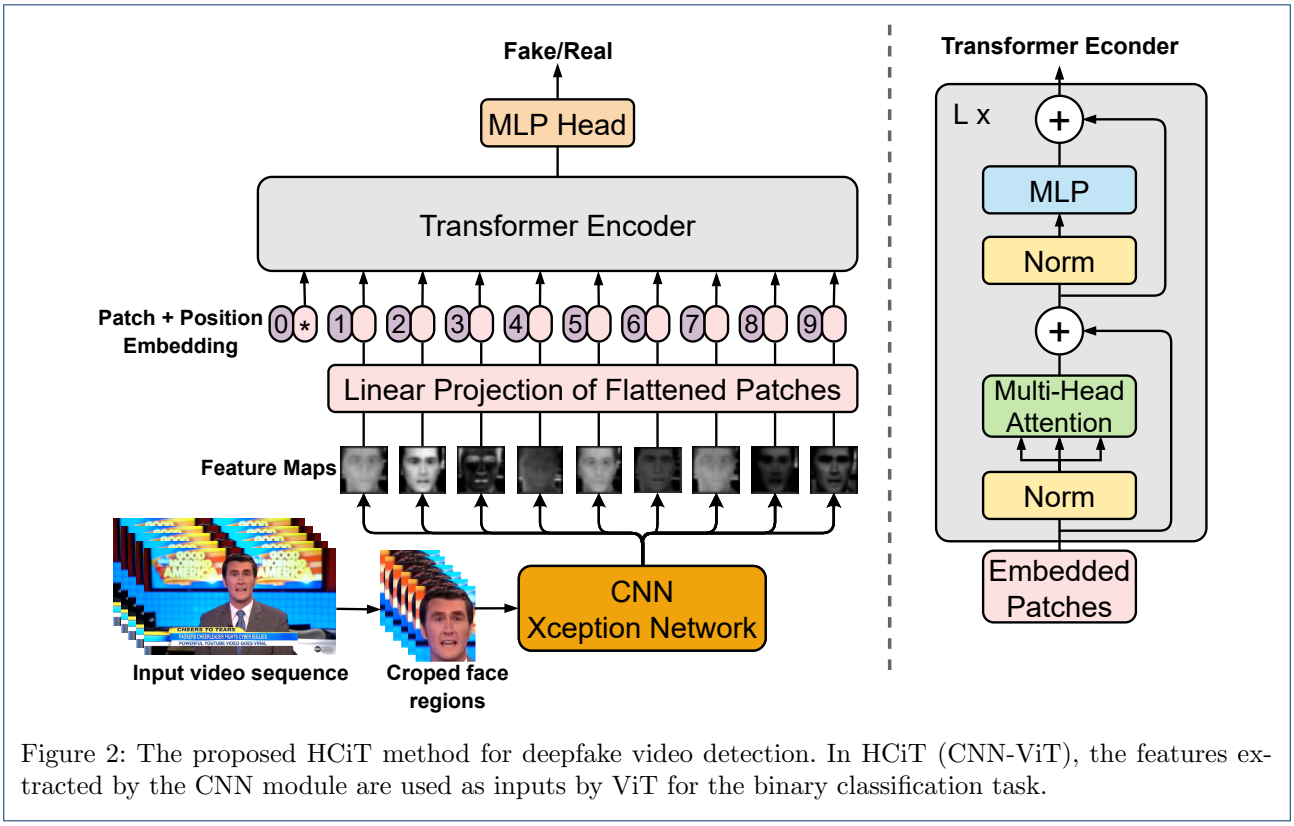


Figure 2: The proposed HCiT method for deepfake video detection. In HCiT (CNN-ViT), the features extracted by the CNN module are used as inputs by ViT for the binary classification task.

Table 2: The performance of ViT-B/32 model on DeepFake, FaceSwap, Face2Face and NeuralTextures datasets using different hyperparameters tuning.

Input image size	Batch size		Learning rate		Accuracy (%)			
	16	32	1e-3	1e-4	DeepFake	FaceSwap	Face2Face	NeuralTextures
224 × 224	✓			✓	93.10	94.97	89.31	78.30
224 × 224		✓		✓	**95.71**	95.64	87.50	77.39
224 × 224		✓	✓		89.31	90.55	82.84	69.84
256 × 256	✓			✓	92.24	95.61	88.36	76.65
256 × 256		✓		✓	95.49	94.62	87.50	77.61
256 × 256		✓	✓		81.32	87.64	83.72	63.51
288 × 288	✓			✓	94.46	**95.76**	87.07	79.74
288 × 288		✓		✓	94.69	95.56	**90.69**	**79.79**
288 × 288		✓	✓		88.29	91.42	82.99	65.40

retain positional information. Finally, the resulting sequence of feature embedding serves as an input to the transformer encoder. The latter consists of alternating layers of Multihead Self-Attention (MSA) and Multi-Layer Perceptron (MLP) blocks. Each layer consists of one normalization layer, a multi-attention layer of four heads, one skip connection, another normalization layer, MLP head with two linear layers with 64×2 and 64 hidden units, respectively, and other skip connection. The output from transformer block is normalized and flatten, then used as input to the MLP. Finally, a dense layer is used as final classifier with MLP head. The MLP head has two linear layers and the ReLU nonlinearity. The first layer has 2048 channels and the last layer has two channels which represent the class of fake or real faces. Softmax is applied on the MLP head output to squash the weight values between 0 and 1 to predict whether the video is real or manipulated.

4 Experiments

In this section, we discuss experiments carried out to evaluate the performance of our proposed HCiT method. First, the used datasets are described. Second, the efficiency of the pure-ViT model for deepfake detection is investigated. Then, the proposed solution is compared to existing approaches. Finally, an ablation study is conducted.

Table 3: The performance of ViT-B/32 model on Faceforensics++ dataset (mixed DeepFake, FaceSwap, Face2Face and NeuralTextures datasets) using different hyperparameters tuning.

ViT	Batch size					Learning rate		Accuracy	Precision	Recall	F1 score
Model	16	32	64	128	256	1e-3	1e-4				
ViT-B/32	✓					✓		60.49	61.37	55.12	58.08
ViT-B/32	✓						✓	67.69	81.33	45.51	58.36
ViT-B/32		✓				✓		63.27	67.47	47.37	55.66
ViT-B/32		✓					✓	73.50	78.50	64.24	70.66
ViT-B/32			✓			✓		64.92	69.57	51.78	59.37
ViT-B/32			✓				✓	**75.17**	80.97	**64.70**	**71.93**
ViT-B/32				✓		✓		68.63	76.15	52.77	62.34
ViT-B/32				✓			✓	71.98	83.49	53.60	65.29
ViT-B/32					✓	✓		66.17	86.72	34.99	49.86
ViT-B/32					✓		✓	71.34	**90.76**	44.41	59.64

Table 4: The performance of ViT-B model on DeepFake, FaceSwap, Face2Face and NeuralTextures datasets using 32×32 and 16×16 input patch.

ViT	Batch size			Learning rate		Accuracy (%)			
Model	16	32	64	1e-3	1e-4	DeepFakes	FaceSwap	Face2Face	NeuralTextures
ViT-B/16	✓			✓		91.59	84.91	75.00	65.94
ViT-B/16	✓				✓	**97.12**	**97.91**	**93.82**	**82.61**
ViT-B/16		✓		✓		95.49	96.29	91.42	80.23
ViT-B/16		✓			✓	96.14	97.60	90.55	79.86
ViT-B/16			✓	✓		95.08	97.39	92.48	80.58
ViT-B/16			✓		✓	97.02	95.90	88.09	76.19
ViT-B/32	✓			✓		83.33	78.23	73.06	50.79
ViT-B/32	✓				✓	93.10	94.97	89.31	78.30
ViT-B/32		✓		✓		89.31	90.55	82.84	69.84
ViT-B/32		✓			✓	95.71	95.64	87.50	77.39
ViT-B/32			✓	✓		90.03	93.67	83.85	73.36
ViT-B/32			✓		✓	93.30	94.42	90.62	77.75

Table 5: Performance comparison in classification accuracy on the Faceforensics++ (FF++), DeepFake Detection Challenge preview (DFDC-p) and Celeb-DF datasets. In FF++ dateset, the four manipulation methods are: DeepFake (DF), FaceSwap (FS), Face2Face (F2F) and NeuralTextures (NT).

Method	FF++				DFDC-p	Celeb-DF
	DeepFake	FaceSwap	Face2Face	NeuralTextures		
Xcept. (Full) [24]	74.55	70.87	75.91	73.33	61.24	N/A
Xcept. (Face) [24]	94.92	90.29	86.86	80.67	85.50	71.6
MesoNet [50]	87.27	61.17	56.20	40.67	74.46	54.8
Bayer *et al.* [75]	50.84	50.59	51.13	49.17	50.62	N/A
EfficientNet-b5 [76]	90.94	91.57	89.75	82.15	80.78	N/A
Inception Res.V1 [77]	79.51	78.20	77.89	63.87	59.83	N/A
Conv-LSTM [21]	52.38	61.72	60.63	58.80	57.55	N/A
CViT [30]	93.00	69.00	46.00	60.00	87.25	73.85
HCiT	**96.00**	**97.82**	**95.85**	**86.29**	**89.73**	**76.03**

4.1 Dataset

The experiments were carried out with two commonly used datasets, namely the Faceforensics++ (FF++) [3] and the DeepFake Detection Challenge preview (DFDC-p) [4]. FF++ dataset consists of 1000 original video sequences that have been manipulated with four methods: DeepFake (DF), FaceSwap (FS), Face2Face (F2F) and NeuralTextures (NT). The DFDC-p dataset provides more than 5K videos featuring two facial modification algorithms. For both datasets, we used 70:15:15 ratios for training, validation, and test. Celeb-DF is comprised of 590 real videos and 5,639 Deepfake videos, in which 6,011 videos are used for training and 518 videos are for testing. We summarize the basic information of these datasets in Table 1.

4.2 Performance Evaluation Methodology

To evaluate the performance of our model, we employ a number of classification metrics. We choose the overall accuracy, precision, recall and the F1 score [78]. Additionally, a graphical representations,i.e., Receiver

Operator Characteristic (ROC) curve, The area under the curve (AUC), and Confusion matrix, are used to provide a clear visual illustration of the overall performance.

4.3 The effectiveness of pure-ViT for deepfake detection

Before assessing our method, in this section, the pure-ViT model is extensively evaluated for deepfake detection using different hyperparameters tuning.

Typically, the standard ViT receives a sequence of token embeddings as input. For a 2D image, ViT reshapes the image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where $(H \times W)$ is the resolution of the image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches (i.e., the input sequence length). These patches are flattened and mapped to latent embeddings with a size of D [29]. In addition, an extra class token and position embeddings are added to the sequence.

For this experiment, we consider ViT-B/32, where B and 32 mean the base variant with an input patch size of 32×32, pre-trained on the public ImageNet-21k dataset. The ViT-B/32 is fine-tuned on FF++ dataset. The number of epochs is fixed to 20 and only the best model is considered.

Table 2, Table 3 Table 4 summarize the performance of ViT-B on FF++ dataset using different hyperparameters tuning. The effect of input image size, batch size, patch size on the performance of ViT model and the impact of learning rate will be studied for deepfake detection. Input image sizes [224×224, 256×256, 288×288], batch sizes [16, 32, 64, 128, 256], patch sizes [16, 32] and learning rates [0.001, 0.0001], are used.

TABLE 2 shows the accuracy results on Deepfake, FaceSwap, Face2Face, and NeuralTextures. As we can se, using learning rate of 0.001, whatever the dataset, input image size and batch size used, the lowest performance was achieved. The highest performance was using the learning rate of 0.0001 with significant improvement.

In particular, on DeeFake dataset, for a learning rate of 0.0001, the ViT model achieves the highest accuracy using a batch size of 32, i.e., 224×224 (95.71 %), 256×256 (95.49 %), 288×288 (94.69 %), better than using a batch size of 16 (+2.61), (+3.25 %) and (+0,23 %), respectively. The highest performance was from using an input image size of 224×224, the largest batch size of 32 and a learning rate of 0.0001 (95.71%). On FaceSwap dataset, Face2Face, and NeuralTextures datasets, best performance is obtained using input image size 288 × 288, batch size of 32, and learning rate 0.0001, with an accuracy of (95.76 %), (90.69 %), (79.79 %), respectively.

Table 3 shows the results achieved on all FF++ dataset, i.e., Deepfake, FaceSwap, Face2Face, and NeuralTextures datasets are mixed, using input image size of 224×224. From the results, ViT-B/32 achieves the best performance, with 75.17% accuracy, 80.97% precision, 64.70% recall and 71.93% F1 score, using batch size of 64 and and learning rate 0.0001. Considering the Precision metric, we can observe that, whatever the learning rate, more the batch size increase, best the performance is.

Table 4 shows the accuracy results on Deepfake, FaceSwap, Face2Face, and NeuralTextures, using patch size of 16×16, i.e., ViT-B/16, and 32×32, i.e., ViT-B/32. We perform additional regularization on the parameters using different batch sizes and learning rates. Table 4 shows the results. As one can obsrve, the performance is significantly affected by the input patch choice. In prticular, the ViT-B/16 model outperforms ViT-B/32 on all datasets. The main reason for this is that small input patch, i.e., 16×16, alows learning local and long-range features within the input image more effectively. However, this requires substantially more memory resources usage to train.

4.4 Evaluation of HCiT

4.4.1 Implementation details

Our HCiT method combines the Xception and ViT models. First, we have initialized all the layers of Xception model with the ImageNet weights, then we retrained the Xception model on the considered deepfake dataset for 20 epochs. Next, we removed the FC layer at the top of the Xception network to directly obtain the feature maps. Finally, we combined this Xception model with ViT to form the HCiT model.

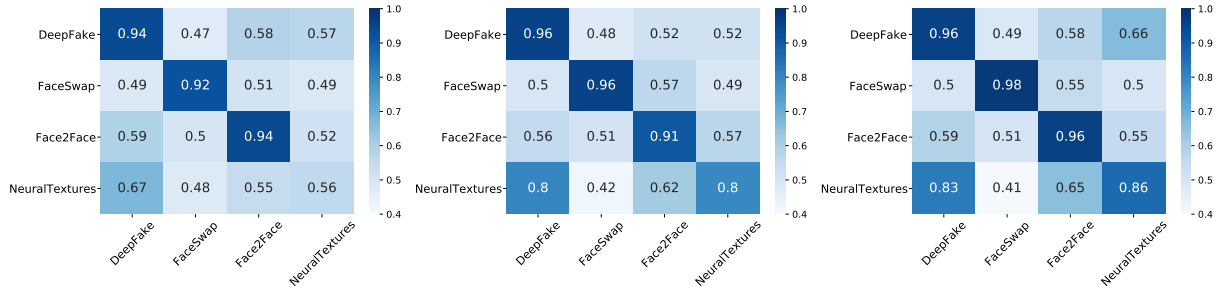
Subsequently, the HCiT model was trained, i.e., Xception and ViT were jointly trained in end-to-end manner, with cross-entropy loss, a learning rate of 0.0001 and a batch size of 32. We used the standard Adam optimizer [79] for training the HCiT model for 20 epochs.

4.4.2 Baseline methods

We compared the proposed HCiT method with eight state-of-the-art deepfake detection methods on the FF++ and DFDC-p datasets, including Xception using full image resolution (noted as Xcept. (Full)) [24], MesoNet [50], Xception using cropped face (noted as Xcept. (Face)) [24], Bayer *et al.* method [75], EfficientNet-b5 [76], Inception ResNet V1 [77], ConvLSTM [21] and CViT [30]. The results in terms of classification accuracy are reported in Table 5.

4.4.3 Results

From Table 5, we can see that the proposed method achieves much better performance than the baseline



(a) Xception (b) ViT (c) HCiT

Figure 3: Accuracy (%) for cross-dataset evaluation on FF++ : DeepFake, FaceSwap, Face2Face and NeuralTextures.

Table 6: Ablation study of HCiT method conducted on Faceforensics++ dataset.

Method	Training on				Testing on	
	DF	**FS**	**F2F**	**NT**	**DFDC-p**	**Celeb**
Xception	✓				55.01	53.75
		✓			37.74	14.80
			✓		16.31	21.39
				✓	35.52	37.88
ViT	✓				39.96	49.23
		✓			21.71	17.63
			✓		25.67	19.47
				✓	53.76	65.10
HCiT	✓				57.02	54.03
		✓			48.45	21.47
			✓		28.33	24.26
				✓	55.04	68.81

methods on all datasets. The Xception model using cropped face outperforms the one using full image resolution. This is because the cropping operation allows the model to focus on the manipulated region and avoid the contribution of background and other biases. The EfficientNet-b5 provides an acceptable results on both databases, with a slight decrease on the DFDC-p dataset. Some methods work well on some manipulation methods and not on others, demonstrating a lack of generalization. However, our method performs well on all types of manipulation, this is particularly noticeable for FaceSwap and Face2Face sub-datasets, where the HCiT method greatly outperforms the other methods.

4.4.4 Cross-dataset evaluations

To evaluate the generalizability of our HCiT method, we performed cross-data evaluation by using a model trained on one dataset and testing it on the other datasets. The goal here is to assess the HCiT's ability to predict a new type of face manipulation that was not used in the training.

In order to give an insight on how the model will generalize to an independent dataset, the considered datasets are classified into two sub-categories, according to the technique used to produce the fake videos : *i*) identity modification (e.g., DeepFake and FaceSwap), and *ii*) expression modification (e.g., Face2Face and NeuralTextures).

As shown in Figure 3, for the prediction accuracy between different datasets, but on the same sub-category, there is a significant drop in performance for the Xception model, in particular, on the methods based on identity modification. While HCiT outperforms Xception model, for example FaceSwap (+2%) and DeepFake (+1.28%). For methods based on expression modification, our method has the highest prediction accuracy and outperforms Xception model by a large margin, NeuralTextures (+2.91%) and Face2Face (+10.62%). On the other hand, when the models are evaluated on unrelated manipulation methods, which is more challenging, HCiT show better results, although the performance of both methods appear to be the same in some cases. For instance, when trained on NeuralTextures, our model achieves a prediction

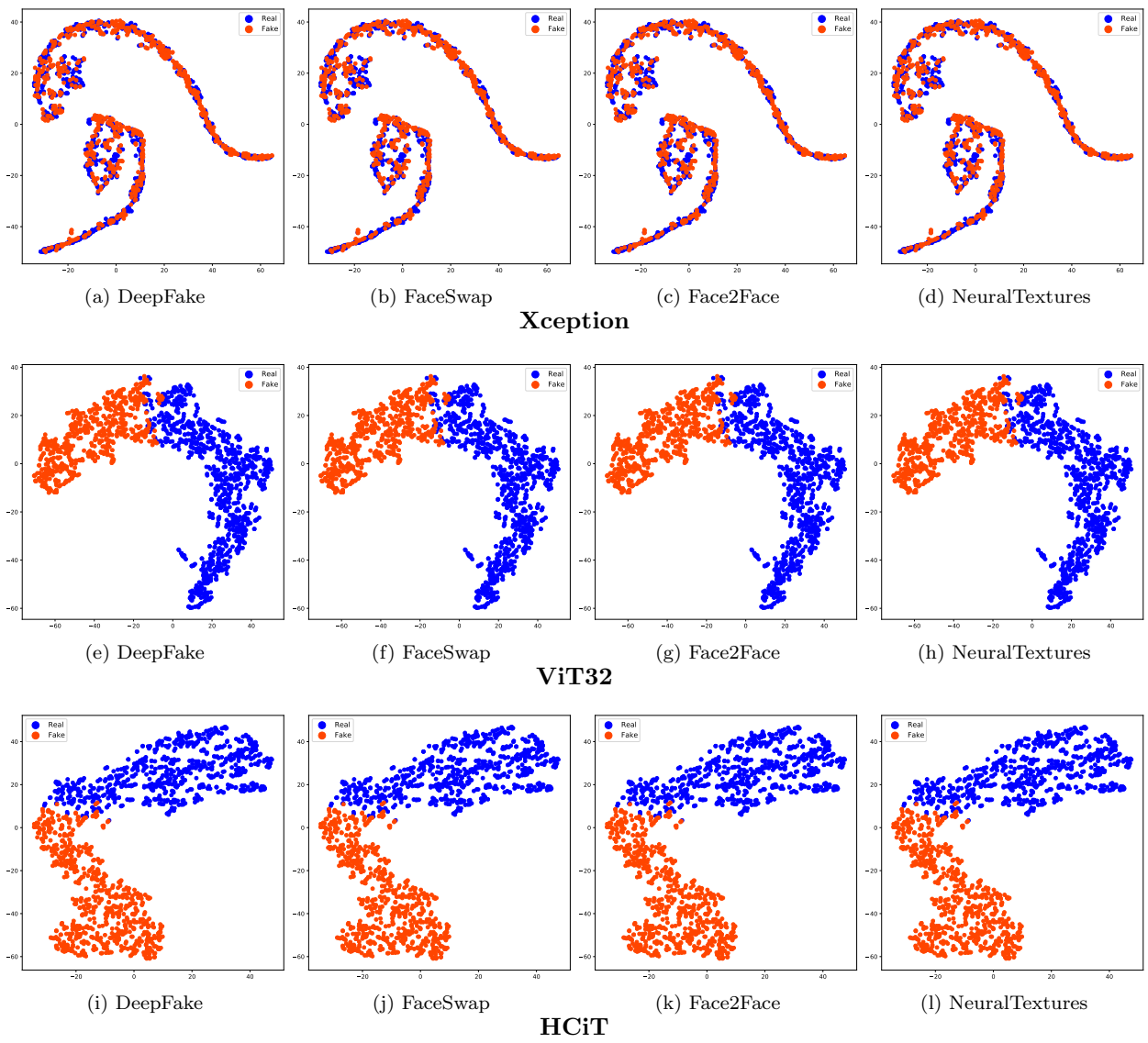


Figure 4: t-SNE for different models using Faceforensics++ (FF++) dataset.

accuracy of 82.71% on DeepFake, which represents an increase of +15.29% compared to Xception. However, the performance drop significantly when tested on FaceSwap dataset. When trained on Face2Face, both models perform slightly similar on DeepFake and FaceSwap datasets.

These results demonstrate that the proposed HCiT method is not limited by the database on which it was trained and shows a strong capacity for generalization.

Furthermore, we demonstrate the generalization capabilities of our proposed model on both DFDC and Celeb datasets. In the experiment setting, we use videos in each face manipulation method listed in the FF++ dataset, i.e., DF, FS, F2F, NT, for training while testing on Celeb-DF, DFDC. We include the results of Xception and ViT for comparison. The results are shown in Table 6. We can observe that our proposed method consistently outperforms all compared opponents on all unseen datasets. In particular, testing on DFDC dataset, the accuracy of our method exceeds Xception method by +2.01, +10.71, +12.02, +18.52. Different from Xception which only utilizes the local feature information, our model combine both local and global information for a more comprehensively feature presentation, so that more kinds of artifacts of the faked face can be captured. Compared with ViT which also takes global information into consideration, our method demonstrates superior performance on all

unseen datasets. This proves the effectiveness of our proposed method.

On the other hand, we observe that cross-validating models trained with both DF and NT features can obtain more generalized results than models trained with other datasets, i.e., FS and F2F. The main reason for this is that both DF and NT manipulation methods produce less forged artifacts which reduces difference between real and fake video. Training on such datasets, model will best generalize for unseen datasets with more distinction between unforged and munipulated images.

4.4.5 Feature Space Distribution

To gain more insight of the representation learning capabilities of Xception, ViT-B/32, and the HCiT on Faceforensics++ dataset, we visualize the top layer features obtained from each model using t-Distributed Stochastic Neighbor Embedding (t-SNE). t-SNE is a relatively new method of dimension reduction particularly suitable for non-linear and high-dimensional datasets. It is one of the leading techniques for data visualization and clustering. This method finds lower dimensional embeddings of data points while minimizing distortions in distances between neighboring data points. By construction, t-SNE discards information about large scale structure of the data.

The 2D visualization by t-SNE of the feature space of Xception, ViT-B/32, and HCiT is illustrated in Figure 4. Each point represents a feature, and its colour is the corresponding label, i.e., real (dark blue) and fake (light blue) videos. The four top figures are the visualization of features extracted using the Xception model from the lobal-average-pooling2d layer, the four middles figure are the visualization of features extracted using the ViT-B/32 model from the last dense layer, while the four bottom figures are the visualization of features extracted using the HCiT model from the the last dense layer. These features are fed into t-SNE with the corresponding labels.

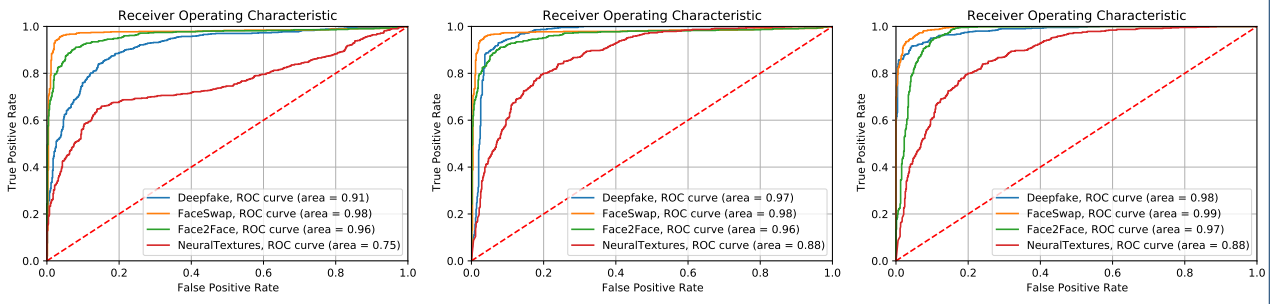
It can be seen that the features generated by our HCiT exhibits better distinct split distribution between real and fake videos compared to the Xception and ViT-B/32. We can see the features distribution of the HCiT is more concentrated, while the real and forged images of FF++ dataset can be easily separated in the feature space. It shows a clear separation between real and fake videos from the Faceforensics++ dataset. This observation demonstrates that the feature vectors from our HCiT model are high level representations and are able to achieve best detection performance. It is obvious that real and forged images are highly distinguishable from each other. This figure also shows that high level representations features are obtained given the ViT-B/32. However, the features are more scattered in the 2D space compared to our HCiT. Considering the Xception model, For both Face2Face and NeuralTextures dataset, features of both real and fake videos are relatively similar and are grouped closely. We deduced that the Xception is not able to extract discriminative features that can well characterize both real and fake videos, thus considerable amounts of confusion are still present, as seen by the mixture of point represents both classes.

4.5 Performance Evaluation based on ROC Curve and AUC Metric

In order to further analyze the performance of different descriptors, we conduct ROC curve-based evaluation. ROC analysis is used in this study to quantify how accurately each descriptor can discriminate between two videos, typically referred to as “fake” and “real”. The ROC curve shows the trade off between the true positive fraction (TPF) and false positive fraction (FPF). Doing this for each descriptor on different datasets, gives the plots in Fig 5.(a), Fig 5.(b), and Fig 5.(c). ROC curve corresponding to best descriptor are located progressively closer to the upper left hand corner in “ROC space”. Intermediate ROC curve corresponds to a moderate discriminative ability of descriptor. An ROC curve lying on the diagonal line reflects the performance of a detection descriptor is poor, i.e., a descriptor which yields the positive or negative results unrelated to the true videos classes.

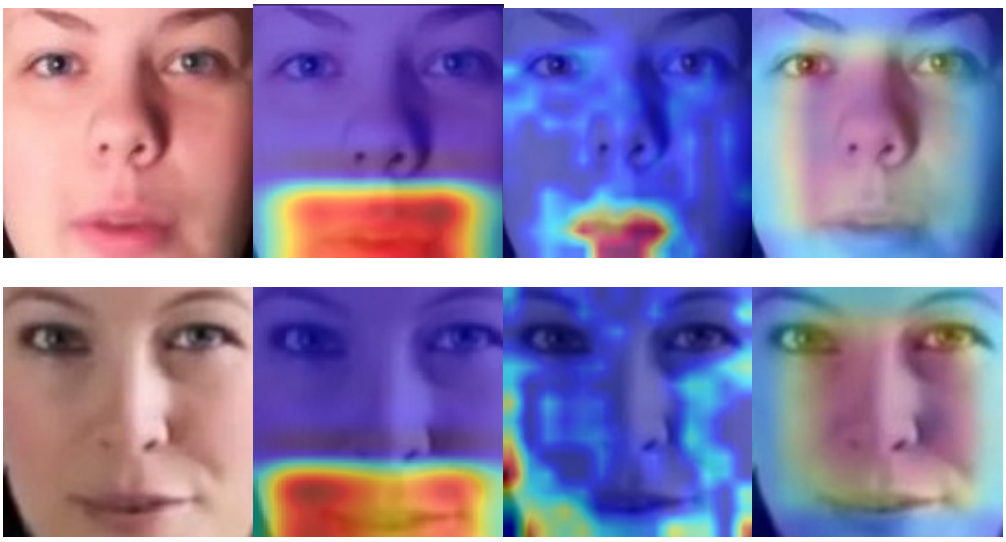
In addition, we used derived summary measure of accuracy from the ROC curve, the AUC, to determine more precisely the inherent ability of the descriptor to discriminate between the fake and real videos. The AUCs for all of the prediction descriptors is summarized in legend of Figures 5. The higher the AUC, the better the classification performance of a model. The maximum AUC=1 means that the descriptor is perfect in the differentiation between the fake and real video. This happens when the distribution of descriptor results for the fake and real do not overlap. AUC =0.5 means the chance discrimination that curve located on diagonal line in ROC space. The minimum AUC should be considered a chance level, i.e., AUC = 0.5 while AUC = 0 means descriptor incorrectly classify all videos with fake as negative and all videos with real as positive. Thus, using this as a measure of a detection performance, one can compare descriptors or judge whether the use of various combination of image components can improve their accuracies. The accuracy of descriptors with AUCs between 0.50 and 0.70 is low; between 0.70 and 0.90, the accuracy is moderate; and it is high for AUCs over 0.90.

As one can observe from Fig 5.(a), Fig 5.(b), and Fig 5.(c), HCiT tends to perform the best among all



(a) Xception (b) ViT (c) HCiT

Figure 5: Receiver Operating Characteristic(ROC) curves using Faceforensics++ (FF++) dataset.



(a) Fake image (b) Xception (c) ViT (d) HCiT

Figure 6: Sample extracted fake face images from FaceForensics++ dataset with their corresponding Grad-CAM heatmaps, representing the facial features most useful for Xception, ViT and HCiT model fake detector.

the other models with a healthy ROC curves pushed towards the top-left side. In particular, on Deepfake, FaceSwap, Face2Face, and NeuralTextures, the HCiT achieves an AUC = 0.98 and 0.99, 0.97, and 0.88, respectively. ViT get slightly similar results and achieves an AUC = 0.97 and 0.98, 0.96, and 0.88, respectively. On the other hand, Xception have the worst performance compared with others models and achieves an AUC = 0.91 and 0.98, 0.96, and 0.75, respectively.

4.6 Attention Analysis with Activation Maps

In this section, we delve into the activation maps of the trained Xception, ViT, and HCiT models to find out their attention when performing the deepfake detection, and verify if the attention mechanizme introduced by the ViT has an influence to it. Specifically, we chose to visualize the activation maps of the last convolution output before the fully connected layer, so that the spatial location of the activations are preserved. The resultant aggregated map will have high values for locations that are either highly activated or gives high contribution to the classifier. This map is then resized to the original image's dimensions and superimposed onto the image, whereby we will be able to visualize the model's attention on the image that led to the classification result. Figure 6 shows two examples of the classified test images and their respective activation regions. Our analysis found that using HCiT, the activation maps shows that that the main attention is often drawn to entirely specific facial regions, i.e., the eye right, the eye left, the nose, and the mouth, highlight-

Table 7: Ablation study of HCiT method conducted on Faceforensics++ dataset.

Method	Pre-train		Accuracy (%)			
	Xception	ViT	DeepFake	FaceSwap	Face2Face	NeuralTextures
(1) ViT	✗	✗	73.32	65.62	69.62	58.64
(2) ViT	✗	ImageNet	95.71	95.76	90.69	79.79
(3) Xcep.+ViT	✗	✗	82.12	77.76	74.70	64.46
(4) Xcep.+ViT	ImageNet	✗	93.60	95.42	94.25	83.06
(5) Xcep.+ViT (**HCiT**)	ImageNet + FF++	✗	**96.00**	**97.82**	**95.85**	**86.29**

ing the importance of this combi-nation on the final detection performance. For ViT model, the activation maps shows that the main attention is drawn on some partially regions of the face. Considering the Xception model, we can see that the initial layers aren't concentrating exactly on the glasses, but we can also see that as we reach the final layers, they're able to focus on the sunglasses.

4.7 Ablation study

To evaluate the efficiency of our proposed method, we conducted an ablation study. We therefore evaluated the performance of each component of the HCiT method using different training strategies. Specifically, we investigated the performance of five model variants for deepfake video detection as follows:

(1) Pure-ViT trained from scratch,
(2) Pure-ViT pre-trained on imageNet,
(3) Xception+ViT: hybrid model with both trained from scratch,
(4) Xception+ViT: hybrid model with Xception pre-trained on imagenet and ViT from scratch,
(5) Xception+ViT (HCiT): hybrid model with Xception fine-tuned on deepfake dataset and ViT from scratch.

As shown in Table 7, direct application of a pure-ViT model trained from scratch for deepfake detection cannot produce a satisfactory result. Indeed, ViT requires a large amount of training data to obtain high performance, which is not the case with the adopted deepfake datasets. When transfer learning is performed with pre-trained ViT model on ImageNet, significant improvements are achieved. This confirms findings of the literature, that the use of pre-trained networks has been an effective strategy to deal with limited data training. Now the combination of Xception and ViT models, both trained from scratch, works better than pure-ViT model trained from scratch. Because, in this hybrid model, Xception helps to learn local structures. However, the result is far from satisfactory with this third model. To remedy this, we used the same model, except that Xception is pre-trained on ImageNet. This clearly increases the results. Finally, when using the hybrid model with the Xception pre-trained on ImageNet and fine-tuned on deepfake dataset, the best results can be achieved.

5 Conclusion

In this work, an efficient deepfake detection method, called HCiT, was presented. Instead of using a pure-ViT model applied directly to sequences of image patches, we have feed the feature maps into ViT. These features maps were extracted from Xception fine-tuned on deepfake dataset. We have thus shown that the combination of Xception and ViT makes it possible to exploit the strengths of the two architectures while avoiding their respective limitations.

The HCiT method has been extensively evaluated and the obtained results demonstrate clearly the effectiveness of our proposed approach. Regarding the obtained performance, the hybrid model improves upon pure-ViT and outperforms the state-of-the-art deepfake detection methods on all considered databases. In addition, the model significantly improves the performance on cross databases.

Author details

[1]University of Ibn Khaldoun-Tiaret, Tiaret, Algeria. [2]National Higher School of Telecommunications and ICT, Oran, Algeria. [3]State University of New York Polytechnic Institute, New York, USA. [4]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France. [5]Sorbonne Center for Artificial Intelligence, Sorbonne Univertsity Abu Dhabi, Abu Dhabi, UAE. [6]ring, Universitat Autònoma de Barcelona, Bellaterra, Spain.

References

1. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning, pp. 1096–1103 (2008)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27**, 2672–2680 (2014)
3. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–11 (2019)
4. Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
5. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
6. Ajder, H., Patrini, G., Cavalli, F., Cullen, L.: The state of deepfakes: Landscape, threats, and impact. Amsterdam: Deeptrace (2019)
7. Korshunov, P., Marcel, S.: Subjective and objective evaluation of deepfake videos. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2510–2514 (2021). IEEE

8. Masood, M., Nawaz, M., Malik, K.M., Javed, A., Irtaza, A.: Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. arXiv preprint arXiv:2103.00484 (2021)
9. Thies, J., Zollhöfer, M., Theobalt, C., Stamminger, M., Nießner, M.: Headon: Real-time reenactment of human portrait videos. ACM Transactions on Graphics (TOG) **37**(4), 1–13 (2018)
10. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. In: ACM SIGGRAPH 2008 Papers, pp. 1–8 (2008)
11. Marr, B.: The best (and scariest) examples of ai-enabled deepfakes. Re-trieved from https://www. forbes. com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-aienabled-deepfakes (2019)
12. Chesney, B., Citron, D.: Deep fakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev. **107**, 1753 (2019)
13. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. arXiv preprint arXiv:2001.00179 (2020)
14. Mirsky, Y., Lee, W.: The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR) **54**(1), 1–41 (2021)
15. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92 (2019). IEEE
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.-k.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
17. Bińkowski, M., Sutherland, D.J., Arbel, M.-c., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
18. Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2017). IEEE
19. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 3154–3160 (2017)
20. Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., Natarajan, P.: Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**(1) (2019)
21. Güera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2018). IEEE
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pat-tern Recognition, pp. 770–778 (2016)
24. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
25. Neves, J.C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., Fierrez, J.: Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. IEEE Journal of Selected Topics in Signal Processing **14**(5), 1038–1048 (2020)
26. Fernando, T., Fookes, C., Denman, S., Sridharan, S.: Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. arXiv preprint arXiv:1911.07844 (2019)
27. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security **13**(11), 2691–2706 (2018)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
29. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
30. Wodajo, D., Atnafu, S.: Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126 (2021)
31. Wang, J., Wu, Z., Chen, J., Jiang, Y.-G.: M2tr: Multi-modal multi-scale transformers for deepfake detection. arXiv preprint arXiv:2104.09770 (2021)
32. Heo, Y.-J., Choi, Y.-J., Lee, Y.-W., Kim, B.-G.: Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353 (2021)
33. Pino, S., Carman, M.J., Bestagini, P.: What's wrong with this video? comparing explainers for deepfake detection. arXiv preprint arXiv:2105.05902 (2021)
34. Redi, J.A., Taktak, W., Dugelay, J.-L.: Digital image forensics: a booklet for beginners. Multimedia Tools and Applications **51**(1), 133–162 (2011)
35. Farid, H.: Photo Forensics. MIT press, ??? (2016)
36. Agarwal, A., Singh, R., Vatsa, M., Noore, A.: Swapped! digital face presentation attack detection via weighted local magnitude pattern. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 659–665 (2017). IEEE
37. Akhtar, Z., Dasgupta, D.: A comparative evaluation of local feature descriptors for deepfakes detection. In: 2019 IEEE International Symposium on Technologies for Homeland Security (HST), pp. 1–5 (2019). IEEE
38. Mo, H., Chen, B., Luo, W.: Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, pp. 43–47 (2018)
39. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security **7**(3), 868–882 (2012)
40. Cozzolino, D., Poggi, G., Verdoliva, L.: Splicebuster: A new blind image splicing detector. In: 2015 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2015). IEEE
41. Cozzolino, D., Verdoliva, L.: Noiseprint: a cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2019)
42. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 101–117 (2018)
43. Barni, M., Nowroozi, E., Tondi, B.: Detection of adaptive histogram equalization robust against jpeg compression. In: 2018 International Workshop on Biometrics and Forensics (IWBF), pp. 1–8 (2018). IEEE
44. Mandelli, S., Bonettini, N., Bestagini, P., Lipari, V., Tubaro, S.: Multiple jpeg compression detection through task-driven non-negative matrix factorization. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2106–2110 (2018). IEEE
45. Pan, X., Zhang, X., Lyu, S.: Exposing image splicing with inconsistent local noise variances. In: 2012 IEEE International Conference on Computational Photography (ICCP), pp. 1–10 (2012). IEEE
46. Goljan, M., Fridrich, J.: Cfa-aware features for steganalysis of color images. In: Media Watermarking, Security, and Forensics 2015, vol. 9409, pp. 279–291 (2015). SPIE
47. Zhang, Y., Zheng, L., Thing, V.L.: Automated face swapping and its detection. In: 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), pp. 15–19 (2017). IEEE
48. Yu, N., Davis, L.S., Fritz, M.: Attributing fake images to gans: Learning and analyzing gan fingerprints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7556–7566 (2019)
49. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)
50. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7 (2018). IEEE
51. McCloskey, S., Albright, M.: Detecting gan-generated imagery using color cues. arXiv preprint arXiv:1812.08247 (2018)
52. Guan, H., Kozak, M., Robertson, E., Lee, Y., Yates, A.N., Delgado, A., Zhou, D., Kheyrkhah, T., Smith, J., Fiscus, J.: Mfc datasets:

Large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 63–72 (2019). IEEE
53. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting world leaders against deep fakes. In: CVPR Workshops, vol. 1 (2019)
54. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265 (2019). IEEE
55. Korshunov, P., Marcel, S.: Speaker inconsistency detection in tampered video. In: 2018 26th European Signal Processing Conference (EUSIPCO), pp. 2375–2379 (2018). IEEE
56. Korshunov, P., Halstead, M., Castan, D., Graciarena, M., McLaren, M., Burns, B., Lawson, A., Marcel, S.: Tampered speaker inconsistency detection with phonetically aware audio-visual features. In: International Conference on Machine Learning (2019)
57. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 384–389 (2018). IEEE
58. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311 (2019). IEEE
59. Ding, X., Raziei, Z., Larson, E.C., Olinick, E.V., Krueger, P., Hahsler, M.: Swapped face detection using deep learning and subjective assessment. EURASIP Journal on Information Security **2020**(1), 1–12 (2020)
60. Do, N.-T., Na, I.-S., Kim, S.-H.: Forensics face detection from gans using convolutional neural network. ISITC **2018**, 376–379 (2018)
61. Tariq, S., Lee, S., Kim, H., Shin, Y., Woo, S.S.: Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, pp. 81–87 (2018)
62. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831–1839 (2017). IEEE
63. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
64. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876 (2019)
65. Nguyen, H.H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019)
66. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)
67. Zhu, Y., Newsam, S.: Densenet for dense flow. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 790–794 (2017). IEEE
68. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
69. Li, Y., Chang, M.-C., Lyu, S.: In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
70. Yuan, K., Guo, S., Liu, Z., Zhou, A., Yu, F., Wu, W.: Incorporating convolution designs into visual transformers. arXiv preprint arXiv:2103.11816 (2021)
71. Shao, R., Shi, Z., Yi, J., Chen, P.-Y., Hsieh, C.-J.: On the adversarial robustness of visual transformers. arXiv preprint arXiv:2103.15670 (2021)
72. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research **10**, 1755–1758 (2009)
73. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee
74. d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. arXiv preprint arXiv:2103.10697 (2021)
75. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10 (2016)
76. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR
77. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)
78. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19, pp. 1015–1021 (2006). Springer
79. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)