Hybrid Transformer Network for Deepfake Detection

Sohail Ahmed Khan
sohail.khan@uib.no
MediaFutures
Bergen, Norway

Duc-Tien Dang-Nguyen
ductien.dangnguyen@uib.no
MediaFutures
Bergen, Norway

ABSTRACT

Deepfake media is becoming widespread nowadays because of the easily available tools and mobile apps which can generate realistic looking deepfake videos/images without requiring any technical knowledge. With further advances in this field of technology in the near future, the quantity and quality of deepfake media is also expected to flourish, while making deepfake media a likely new practical tool to spread mis/disinformation. Because of these concerns, the deepfake media detection tools are becoming a necessity. In this study, we propose a novel hybrid transformer network utilizing early feature fusion strategy for deepfake video detection. Our model employs two different CNN networks, i.e., (1) XceptionNet and (2) EfficientNet-B4 as feature extractors. We train both feature extractors along with the transformer in an end-to-end manner on FaceForensics++, DFDC benchmarks. Our model, while having relatively straightforward architecture, achieves comparable results to other more advanced state-of-the-art approaches when evaluated on FaceForensics++ and DFDC benchmarks. Besides this, we also propose novel face cut-out augmentations, as well as random cut-out augmentations. We show that the proposed augmentations improve the detection performance of our model and reduce overfitting. In addition to that, we show that our model is capable of learning from considerably small amount of data.

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**.

KEYWORDS

deepfake detection, face forensics, attention mechanisms, image analysis, feature fusion, misinformation detection, transformers

1 INTRODUCTION

The availability of huge image/video datasets and affordable compute resources, has resulted in swift progress in the field of deep learning research, specifically in the subarea of Generative Adversarial Networks (GANs) [12]. This progress has made it almost effortless to generate realistic synthetic content even for non-technical computer users. The synthetic content generated using deep learning models (i.e., GANs) is called Deepfake media. Deepfake media can be in the form of images, videos, text and audios. However, out of all the different categories of deepfake media, the visual deepfake media is the most common form of fake/synthetic content we encounter nowadays. The number of deepfake media generation techniques is growing exponentially. The newer generation techniques are able to generate extremely plausible synthetic content, and it is becoming more and more challenging to detect the generated fake media.

The most popular form of facial deepfake media we encounter at present is generated using face swapping method, in which the face of a person (target) is swapped with the face of another person (source). There are 4 different types of facial deepfake media, i.e., (1) Face Swapping, (2) Face Re-enactment, (3) Face Editing and (4) Face Synthesis [18]. In this paper we focus on detecting facial deepfake media, specifically the media generated using face swapping and face re-enactment techniques.

Using ensembled or fusion based models tend to achieve exceptional results as compared to single models [8, 9, 22, 28]. We therefore propose to employ two different CNN models as feature extractors along with a transformer architecture (Vision Transformer [10]). We expect that by fusing features extracted using different feature extractors will result in diverse feature spaces, which will help the transformer to learn diverse set of features. Transformer architectures are capable of simultaneously learning meaningful associations from long input sequences. We therefore choose transformer [10] to learn joint feature space, instead of the classical way of using a fully connected layer to combine different feature sets. Besides this, hybrid (having CNN as feature extractor instead of using simple patch embeddings) transformer models tend to achieve even better results.

We choose XceptionNet as one of the feature extractors as it has been widely employed in deepfake detection domain [23]. The second CNN which we choose is the EfficientNet B4 model, which also achieved exceptional results on ImageNet benchmark. We do not freeze the feature extractors during training, i.e., we train both feature extractors, as well as the transformer architecture in an end-to-end manner using a single loss function.

The contributions of this paper are three fold, (1) we propose a novel hybrid transformer architecture which learns from joint feature sets extracted by two different CNN feature extractors, (2) we show that image augmentations we generate using our face pre-processing module, combined with other affine transformation based augmentations, improve the performance of the detection models while reducing overfitting, and (3) we show that while having a simple and easy to implement architecture, our model achieves comparable results to other more complex state-of-the-art approaches while being trained on comparably smaller number of data samples.

This paper is structured as follows, in section 2 we present a brief literature review, in section 3 we describe the augmentations we employ, the proposed face pre-processing module, our model architecture, the datasets used to train our models, and implementation details, in section 4 we compare the achieved results with other deepfake detection baselines, and in section 5 we conclude our study and propose future research directions.

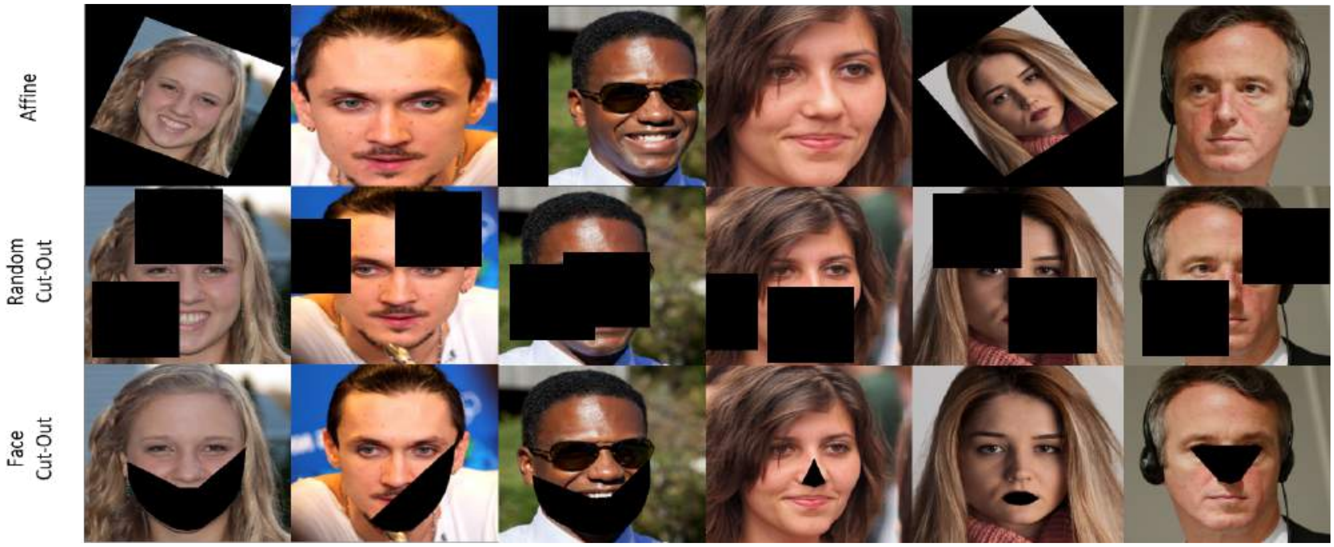

Figure 1: Our face pre-processing module is responsible for applying 3 different types of image augmentations, (1) Affine Transformations, (2) Random cut-out augmentations, and (3) Face cut-out augmentations. More details about the augmentations we use are given in upcoming sections. We show that the employed image augmentations improve detection performance while reducing overfitting. Photos acquired from Flickr Faces HQ dataset [15].

2 RELATED WORK

Recent works on deepfake media detection mostly employ CNN based architectures along with other strategies (e.g., multimodal features, recurrent networks, transformer models etc) to detect deepfake images/videos. Unlike the previous research studies, in this paper we propose a novel strategy to simultaneously learn from joint feature spaces extracted using two different CNN feature extractors using a transformer architecture while employing heavy image augmentations.

Rossler *et al.* in [23] proposed a simple deepfake detection technique based on the XceptionNet [4] CNN model pre-trained on the imagenet dataset. Authors fine-tune the generic XceptionNet on their FaceForensics++ dataset while reporting excellent performance scores the four subsets of the FaceForensics++ dataset, namely, (1) FaceSwap, (2) Face2Face, (3) DeepFakes, and (4) NeuralTextures [23]. The proposed model achieved excellent results on uncompressed videos, however, lost performance when tested on compressed videos.

In [28] Zhu *et al.* propose to utilize 3D facial details to detect deepfakes. Authors find that merging the 3D identity texture and direct light is significantly helpful in detecting deepfakes. They employ the XceptionNet CNN model for feature extraction. A face cropped image and its 3D detail is used to train the detection model. They also perform a detailed analysis of a number of different feature fusion strategies. The proposed technique was trained on FaceForensics++ dataset and evaluated on (1) FaceForensics++, (2) Google Deepfake Detection Dataset, and (3) DFDC datasets. Authors report promising results on all of the three datasets along with better generalization capability than the previously proposed techniques.

Qi *et al.* in [22] propose a novel deepfake detection technique which they call, DeepRhythm. The proposed technique works by analyzing the heartbeat rhythms of a person in a given video. They employ photoplethysmography (PPG) to analyze minute changes in the skin tone inherent with the blood pumping visible on the human faces. Authors evaluate the proposed technique on FaceForensics++, and DFDC datasets and report excellent performance results.

In [27] Xuan *et al.* proposed a more general deepfake media detection technique by employing image augmentations, for example, gaussian blur and gaussian noise to preprocess images in order to remove low level high frequency GAN artifacts present inside the generated images. They then trained a forensic convolutional neural network model on the preporcessed images. They established that by using image augmentations on both real and fake images, destroy the low level noise cues, while forcing the forensics model to learn more meaningful features. Through experimentation authors establish the effectiveness of their proposed technique on deepfake media detection.

Sabir *et al.* in [24] proposed a recurrent convolutional network for deepfake media detection. The DenseNet convolutional neural network was employed along with a gated recurrent neural network to exploit temporal inconsistencies present between frames of a deepfake video. The proposed technique was evaluated on the FaceForensics++ [23] dataset showing promising performance.

Following a similar path, Guera and Delp in [13] proposed to employ a convolutional neural network along with a long short term memory (LSTM) network for deepfake video detection. Authors tried to exploit inter-frame discrepancies inherent to most deepfake videos. The CNN was tasked with extracting frame-level features, which were then fed to the LSTM network to learn the temporal features. Authors evaluated the model on their own dataset showing exceptional performance.

Nguyen *et al.* in [21] proposed a capsule network based model to detect deepfake media. The proposed model was evaluated on

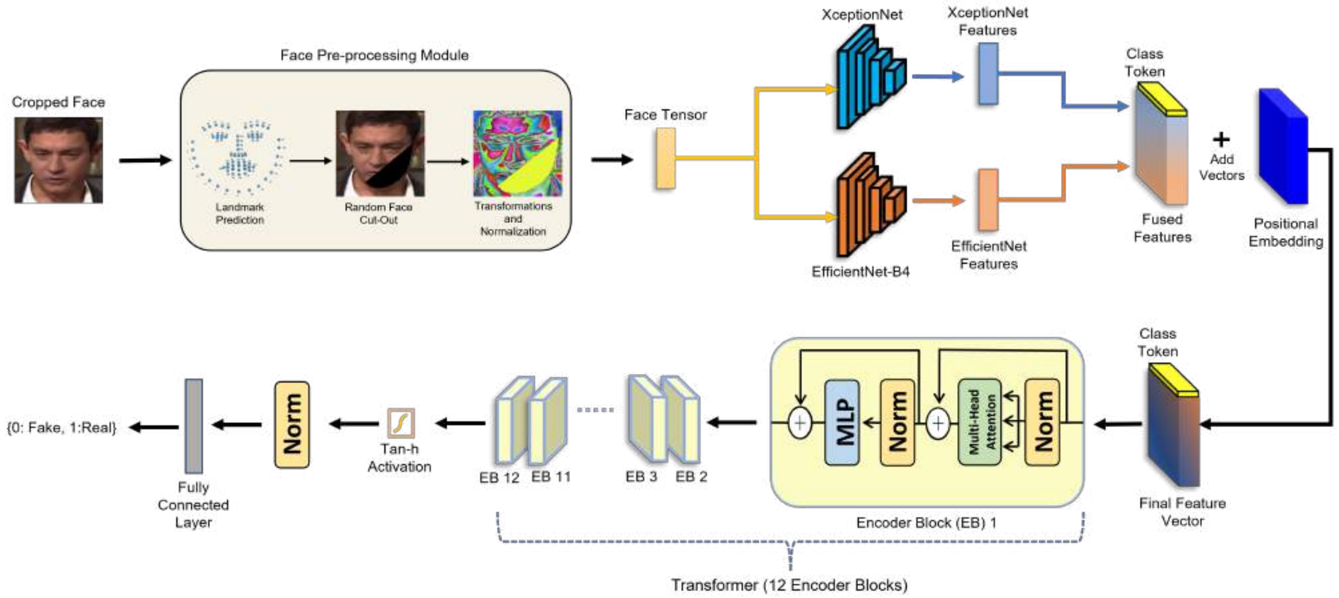

Figure 2: Our hybrid transformer network comprising of two different CNNs, (1) XceptionNet, (2) EfficientNet-B4 and a transformer. Our model is trained on the cropped face images. The face image is fed to the face pre-processing module which applies face cut-out/random cut-out and affine transformation augmentations randomly. The augmented face image is then fed to the CNN feature extractors. The extracted features are then fused together by concatenation. A BERT style [*class*] token and learnable positional embedding are added to the concatenated features which are then fed to the transformer. Both feature extractors and the transformer are trained in an end-to-end manner using a single loss function.

four different deepfake detection datasets containing a wide range of synthetic images and videos. The proposed method achieved excellent results as compared to other methods on all datasets.

Again in [20], Nguyen *et al.* proposed a different strategy employing a Y-shaped encoder-decoder model. Authors trained the model by following a multi-task learning based technique which was able to classify and generate a segmentation mask of the tampered regions within manipulated images/videos. The proposed model was evaluated on the FaceForensics and FaceForensics++ [23] datasets achieving promising results even when finetuned on a small number of images.

In [5] Ciftci *et al.* proposed a novel deepfake detection technique employing biological signals (i.e., photoplethysmography or PPG signals) to train a CNN and a support vector machine (SVM). The predictions from the CNN and SVM were fused to get final classification label. The proposed model achieved promising results when evaluated on several different deepfake detection datasets including, Face Forensics, Face Forensics++ [23], and CelebDF [17] datasets.

In [1] Afchar *et al.* proposed two different CNN models, which they called, (1) Meso-4 and (2) MesoInception-4, both containing a small number of layers focusing on mesoscopic image features. The proposed model was tested on an existing deepfake detection dataset, as well as, a custom dataset collected by the authors. The model achieved excellent results on both datasets.

In [16] proposed a novel video transformer network for deepfake media detection, capable of learning from new data in an incremental manner. The proposed video transformer model was trained on multimodal data (i.e., face cropped images, and UV texture maps). The proposed models achieved excellent results on a number of different deepfake detection benchmarks.

3 METHODOLOGY

In this section, we briefly introduce, (1) image augmentations we employ to train our models, (2) face pre-processing module which is responsible for applying the image aumentations and normalizing the image, (3) the two CNN backbones we employ to extract image features, and (4) the transformer model.

3.1 Augmentations

We use two different image cropping (face cut-out, random cut-out) based augmentations, which are applied along with three different affine transformation (rotation, translation, scaling) based augmentations.

In face cut-out augmentations, we crop out specific face part from the image in a random order using facial landmarks, as shown in figure. Whereas, in case of random cut-out augmentations, we crop out random square shaped region from a given face image. Both of these augmentations are separately applied to train two different models. However, the affine transformation based augmentations are applied along with both of these cut-out based augmentations. The pipeline shown in figure 2, applies face cut-out augmentation along with affine transformation based augmentations to a given face image.

3.2 Face Pre-processing Module

The face pre-processing module is responsible for applying augmentations and normalizing a given input face image. In case of face cut-out augmentations, the face pre-processing module takes a face cropped image as input, predicts 81 face landmarks using OpenCV's facial landmarks predictor. It then selects a number of different landmarks randomly to cut a specific face part out from the image.

The face pre-processing module in case of random cut-out augmentations, takes a face cropped image, randomly crop outs two squared shaped regions from the image. The purpose of using heavy cut out augmentations is that we want to prevent our model from overfitting. If we do not use augmentations, the model will memorize the training data and will not be able generalize well on the test set, as can be seen from the results in table 2.

3.3 XceptionNet

XceptionNet architecture with depth-wise separable convolutions was proposed by François Chollet in [4]. For deepfake detection, Rossler *et al.* employed XceptionNet in [23]. They showed that the XceptionNet architecture achieved exceptional results on FaceForensics++ dataset. Because of the excellent performance of XceptionNet on deepfake detection, in this paper we use XceptionNet as one of the feature extractors.

3.4 EfficientNet-B4

EfficientNets are a new set of state-of-the-art convolutional neural network models for image classification proposed in [25]. For deepfake detection EfficientNet architectures achieved promising results on DFDC dataset. In-fact, the winning model of the DFDC was an ensemble of 5 EfficientNet CNNs [8]. Furthermore, because of the memory constraints we needed a smaller but effective model, making EfficientNet-B4 the best choice for our study.

3.5 Hybrid Image Transformer

Table 2 presents a performance comparison of our model trained under different augmentation strategies. We use heavy image augmentations for example, rotation, horizontal flipping, translation, scaling, face cut-outs and random cut-outs. During experimentation we found that the model overfits severely when no image augmentations are employed. Random cut-out augmentations along with affine transformation based augmentations give best results.

We employ BERT [7, 10] styled transformer to learn the joint features extracted using the two CNNs. Since Transformers have proven to be excellent in simultaneously learning meaningful properties from long sequences because of their bidirectional representation learning capability, we employ these models to learn the joint features extracted by two CNNs. We believe that rather than utilizing the early feature fusion followed by a fully connected layer to learn from the joint feature space, as it is widely done while using multiple CNNs; using a transformer to learn the joint feature space will yield better results.

In addition to this, ensembled/fusion based models are proven to achieve better results when compared with single models specifically in deepfake detection, as we can infer from Facebook's Deepfake Detection Challenge, in which most of the top ranked models employed ensembled/fusion based networks [8].

We utilize Transformer architecture pretrained on ImageNet[1]) on features extracted using an early fusion based strategy in which two different CNN models as introduced above. The extracted features are concatenated, a classification token (similar to BERT's *[class]* token) is added at the start, and assigned a positional embedding before being fed to the transformer architecture. Our model comprises of 12 encoder blocks and 12 attention heads (following the architecture of ViT-Base-16 [10]). The weights for the pretrained CNN models are obtained from Ross Wightman[2]. Both the feature extractors, and the transformer are trained in an end-to-end manner using a single loss function i.e., we do not freeze the weights of feature extractors while training.

3.6 Dataset

We train and evaluate our models on FaceForensics++ [23] and DFDC [9] datasets.

3.6.1 ***FaceForensics++.*** This dataset comprises of four subsets namely, (1) FaceSwap, (2) Face2Face, (3) Deepfakes, and (4) Neural Textures. The videos we used to train our models are high quality (c23). We create train, validation, and test sets according to the instructions in FaceForensics++ dataset paper [23]. There are 720 train videos, 140 validation videos, and 140 test videos. We train our models on 200K images (100k real and 100k fake), which is less than half of the size of training set Rossler *et al.* used to train their XceptionNet model in [23]. Out of the 200K training images, 160K images are used for training and the remaining 40K images are used for validation.

For fake videos, we extract 50 face frames starting from the beginning of each video, whereas, for the real videos, we extract 150 frames from each video. We do this to balance the real and fake test sets. The exact amount of images from each dataset used to train, validate and test our models as given in table 3.

To evaluate our models we extract 16 face frames from 140 test videos which results in a total of 2100 images as mentioned in the table 3. To assign a classification label to any test video, we take 16 frames and feed to our model one by one. The final prediction is made after averaging the predictions obtained from the model for each frame.

3.6.2 ***Deepfake Detection Challenge (DFDC).*** This dataset comprises of around 124K videos. To train our models we use only around 8K videos. The amount of fake videos in DFDC dataset is more than the real videos, and thus to balance the real and fake image samples, we extract 50 frames from each fake video, whereas, from each real video we extract 150 frames. This results in around 265K real and fake frames, from which we only use 48K images to train our model and 12K images for validation purposes. So in total, we only used 60K real and fake face frames from DFDC dataset to train and validate our models.

[1] https://github.com/lukemelas/PyTorch-Pretrained-ViT

[2] https://github.com/rwightman/pytorch-image-models

Table 1: Performance (accuracy) comparison of a number of different deepfake detection baseline models on FaceForensics++ dataset. Each of the mentioned model was trained on all subsets of the FaceForensics++ dataset at once. Best results are highlighted.

Approach	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Pristine	Cumulative
Steg. Features + SVM [11]	68.80%	67.69%	70.12%	69.21%	72.98%	70.97%
Cozzolino *et al.* [6]	75.51%	86.34%	76.81%	75.34%	78.41%	78.45%
Bayar and Stamm [2]	90.25%	93.96%	87.74%	83.69%	77.02%	82.97%
Afchar *et al.* [1]	89.55%	88.60%	81.24%	76.62%	82.19%	83.10%
Rossler *et al.* [23]	97.49%	97.69%	96.79%	**92.19%**	95.41%	95.73%
Qi *et al.* [22]	**99.70%**	**98.90%**	97.80%	-	-	-
Ours (Face cut-out)	97.85%	97.85%	96.42%	90.71%	95.00%	95.57%
Ours (Random cut-out)	98.57%	98.57%	**97.85%**	92.14%	**97.85%**	**97.00%**

Table 2: Performance comparison of our models trained with and without image augmentations. In this table, DF refers to Deepfakes, F2F refers to Face2Face, FS refers to FaceSwap, NT refers to NeuralTextures, and P refers to Pristine.

Model	DF	F2F	FS	NT	P	Agg.
No Augs	95.71%	93.57%	92.85%	85.00%	96.42%	92.71%
Face cut-out	97.85%	97.85%	96.42%	90.71%	95.00%	95.57%
Random cut-out	**98.57%**	**98.57%**	**97.85%**	**92.14%**	**97.85%**	**97.00%**

Table 3: Number of frames used to train models on different datasets: Pristine, FaceSwap, Deepfakes, Face2Face, Neural Textures.

Dataset	Training	Validation	Test
Pristine	100K	20K	2.24K
FaceSwap	25K	5K	2.24K
Deepfakes	25K	5K	2.24K
Face2Face	25K	5K	2.24k
Neural Textures	25K	5K	2.24K

We use 400 test videos provided with the DFDC dataset to evaluate our model. For evaluation, we use the same strategy we used for evaluating the models on FaceForensics++ dataset, i.e., we extract 16 face frames from each test video and feed to our model one by one. The final prediction about a video is made after averaging the individual frame predictions.

3.7 Implementation Details

For face detection and cropping, we use OpenCV[3]. For custom image augmentations such as, rotation, flipping, translation and cutouts we employ ImgAug[4] library. We use Ross Wightman’s github repoitory to download CNN weights pretrained on ImageNet. We borrow code and weights of vision transformer pretrained on ImageNet from Luke Melas [5].

To train our models we use SGD with a momentum ranging from 0.6 to 0.9, with a learning rate of 3×10^{-3}. We stop training when the validation loss keeps on increasing for 3 consecutive epochs, or the training accuracy approaches to 100% (to prevent severe overfitting). We train the two CNNs and the transformer in an end-to-end manner, and optimize them through a single binary cross entropy loss function.

We resize images to $[3, 224, 224]$ dimensions. Having higher resolution images yield better results but because of memory constraints we choose to use this image resolution to train and evaluate our models. The input image is fed to the CNN feature extractors, which after extracting features, each of the feature extractors return features of dimension $[1, 162, 768]$. The obtained features are then concatenated to get final features of dimension $[1, 324, 768]$. We append a BERT style $[class]$ token at the start of the extracted features resulting making the dimension $[1, 325, 768]$. A learnable positional embedding is added to these features through element wise addition. The resulting features are then fed to the transformer.

4 RESULTS

In this section we will present and compare the results our model achieved on the FaceForensics++ and DFDC dataset. We trained our model on FaceForensics++ dataset under three different settings e.g., (1) without image augmentations, (2) with face cut-out augmentations, and (3) with random cut-out augmentations. We found that the model trained using random cut-out augmentations outperformed the other two variants, and thus we further trained this model on DFDC dataset as well. Performance comparison of our model under three augmentation strategies can be seen in table 2. As we understand, the reason behind the excellent performance of random cut-out augmentation is because it cuts out most parts of

[3] https://opencv.org/

[4] https://imgaug.readthedocs.io/en/latest/

[5] https://github.com/lukemelas/PyTorch-Pretrained-ViT

Table 4: We compare the performance (accuracy) of our model with the XceptionNet proposed by Rossler *et al.* in [23]. In this table, DF refers to Deepfakes, F2F refers to Face2Face, FS refers to FaceSwap, NT refers to NeuralTextures.

Model	DF	F2F	FS	NT	Agg.
Rossler *et al.* [23]	92.48%	91.33%	92.63%	85.98%	90.60%
Ours (Random cut-out)	**95.00%**	**95.00%**	**95.71%**	**90.00%**	**93.92%**

Table 5: Number of training and validation images used by different deepfake detection techniques. The number of train, validation, and test sets of other studies in this table are rough estimates, as the the authors do not specify exact number of images they used to train their models.

Approach	Train	Validation	Test
Rossler *et al.* [23]	388K	70K	70K
Zhu *et al.* [28]	360K	70K	70K
Ours	200K	40K	11.2K

the image in a random order preventing the model from memorizing face images.

It should also be noted that we trained our models on smaller number of samples from the FaceForensics++ dataset (given in table 3) as compared to other approaches, e.g., Rossler *et al.* [23] train their model on around 388K images, Zhu *et al.* [28] train their models on around 360K images. We compare our models with the baseline results provided in the original FaceForensics++ paper [23], since they provide performance scores of each model on every subset of the FaceForensics++ dataset. We also present the results achieved by our models on DFDC dataset in table 6.

Table 1 presents a detailed comparison of the obtained accuracy scores on all of the subsets of FaceForensics++ dataset. Our model trained using heavy image augmentations achieves comparable results to the baseline techniques listed in Table 1 and new state-of-the-art deepfake detection techniques[22, 28]. The reason for choosing the techniques in table 1 for comparison is that these techniques are also evaluated in the original FaceForensics++ dataset paper, and carry out the training in the same manner as we do in this paper. Testing is done in a different manner than these approaches i.e., we test our models on 16 face frames per video, whereas these approaches test there models on 100 face frames per video. For validation, the approaches in FaceForensics++ paper use 100 face frames per video, whereas, we split the train and validation set in 80:20 ratio i.e., 200K images for training and 40K images for validation.

Some of the other deepfake detection state-of-the-art techniques [5, 22, 28] achieve better results than our model, however, those techniques are quite complex to implement, and in most cases use more data than we use to train the models.

In table 4 we present a comparison of our model with XceptionNet [23] model proposed by Rossler *et al.* We train and evaluate our model on each of the four subsets of FaceForensics++ dataset separately. For training we only use 5000 fake images (from each subset) and 5000 real images. While Rossler *et al.* trained their models on 50 videos or 13500 images since they use 270 images from each video for training. We show that our model even being trained on less data, improves detection performance. In table 5 we compare the size of training set Rossler *et al.* [23] and Zhu *et al.* [28] use to train their respective models.

Table 6: Performance (accuracy) comparison of a number of different deepfake detection baseline models on DFDC dataset. Best results are highlighted.

Approach	Dataset	Train Size	Accuracy
Mittal *et al.* [19]	DFDC	-	84.40%
Wodajo *et al.* [26]	DFDC	112K	91.50%
Bondi *et al.* [3]	DFDC	-	92.20%
Ours (Random cut-out)	DFDC	48K	**98.24%**

We present a comparison of results our model achieved on the DFDC dataset in table 6. We show that while being trained on smaller training set, our model still achieves exceptional performance scores as compared to other relevant works proposed in the past.

5 CONCLUSION AND FUTURE WORK

Detecting deepfake media is crucial as well as challenging. Besides other challenges of deepfake media detection, for example, poor generalization capability of the detection models, deepfake media is also adversarial in nature and continues to evolve rapidly. In this study we presented an early fusion based hybrid transformer network for deepfake media detection. Our model achieved comparable results to most of the state-of-the-art deepfake detection techniques [22, 28] . After this, we plan to train and evaluate our model on other prominent deepfake detection datasets, such as, Celeb-DF [17], ForgeryNet [14] and others. We also plan to analyze the generalization capability of our model on unseen data in future studies, while trying to visually interpret our model to know what kind of features it utilizes more while making decisions and how it differentiates between different categories of deepfake media, i.e., face swapping, face re-enactment etc.

In future we will focus our work mainly on improving the generalization capability of the deepfake detection models, and further work on improving the adversarial robustness of the detection models.

6 ACKNOWLEDGMENTS

This research was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centres for Research-based Innovation scheme, project number 309339.

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. https://arxiv.org/abs/1809.00888

[2] Belhassen Bayar and Matthew C. Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security* (2016).

[3] L. Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. 2020. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection. *Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS)* (2020), 1–6.

[4] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://arxiv.org/abs/1610.02357

[5] Umur Aybars Ciftci, ˙Ilke Demir, and Lijun Yin. 2020. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE. https://arxiv.org/abs/1901.02212

[6] Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. 2017. Recasting Residual-based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. Available: https://arxiv.org/abs/2006.07397.

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv: Computer Vision and Pattern Recognition* (2020).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2021).

[11] Jessica J. Fridrich and Jan Kodovský. 2012. Rich Models for Steganalysis of Digital Images. *IEEE Transactions on Information Forensics and Security* 7 (2012), 868–882.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems, (NIPS)*. Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[13] David Güera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *Proceedings of the 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. https://ieeexplore.ieee.org/document/8639163

[14] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 4358–4367.

[15] Tero Karras, Samuli Laine, and Timo Aila. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4396–4405.

[16] Sohail Ahmed Khan and Hang Dai. 2021. Video Transformer for Deepfake Detection with Incremental Learning. In *Proceedings of the 29th ACM International Conference on Multimedia.*

[17] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3204–3213.

[18] Yisroel Mirsky and Wenke Lee. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys (CSUR)* (2021).

[19] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2823–2832. https://arxiv.org/abs/2003.06711

[20] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. 2019. Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. In *IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE. https://arxiv.org/abs/1906.06876

[21] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. https://ieeexplore.ieee.org/document/8682602

[22] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, L. Ma, Wei Feng, Yang Liu, and Jianjun Zhao. 2020. DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms. In *Proceedings of the 28th ACM International Conference on Multimedia.*

[23] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1–11.

[24] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 80–87. https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Sabir_Recurrent_Convolutional_Strategies_for_Face_Manipulation_Detection_in_Videos_CVPRW_2019_paper.html

[25] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv* abs/1905.11946 (2019).

[26] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. *ArXiv* abs/2102.11126 (2021).

[27] Xinsheng Xuan, Bo Peng, Wei Wang, and Jing Dong. 2019. On the generalization of GAN image forensics. In *In Sun Z., He R., Feng J., Shan S., Guo Z. (eds) Biometric Recognition. CCBR 2019*. Springer, Cham. https://doi.org/10.1007/978-3-030-31456-9_15

[28] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and S. Li. 2021. Face Forgery Detection by 3D Decomposition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2928–2938.