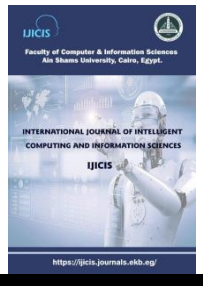
IJICIS, Vol.24, No.1, 55-68
DOI: 10.21608/ijicis.2024.272675.1324

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



DEEPFAKE VIDEO DETECTION USING VISION TRANSFORMER

Shereen A. Hussein*

Department of Computer Science,
Faculty of Computers and Artificial Intelligence, Fayoum University,
Fayoum, Egypt
Sam26@fayoum.edu.eg

Seif N. Mohamed

Department of Computer Science,
Faculty of Computers and Artificial Intelligence, Fayoum University,
Fayoum, Egypt
Sn1279@fayoum.edu.eg

Abstract: Technology is always a double-edged sword, and with the astonishing advancements in technology, it is expected that the DeepFake problem will become more common and serious. DeepFake has recently caused a lot of trouble because its flaws outweigh its advantages. Since DeepFake has such a significant influence on individuals deception, instability of principles and falsification of evidence. Instead of just affecting people, it led to multiple incidents that affected the image of entire nations. In this paper, a model that has been built to mitigate the negative effects of deepFake and maintain an individual's reputation by detecting the alteration of people's photographs and videos. A model with integrated vision transformer architectures Deep-ViT and Cross-ViT is designed to process pre-extracted faces from FF++ dataset. The model distinguishes between the real and fake faces in two different perspectives, subclass detection on each manipulation method and overall detection of all types. The proposed model achieves an outstanding results and the highest accuracy in FaceSwap manipulation method with 98%.

Keywords: *DeepFake Classification, Feature Map, patch extraction & embedding, Attention Layer, Vision Transformer.*

Received 2024-02-25; Revised 2024-02-25; Accepted 2024-03-05

1. Introduction

DeepFakes are altered images or videos that are typically used to spread disinformation. Advanced techniques from the computer vision and deep learning domains are used to create synthetic yet extremely realistic images and videos When facial areas are combined, merged, superimposed, or replaced. The advancement of generative adversarial networks (GANs) [1] makes deepFake implementation possible. GANs are made up of two neural networks, which are series of algorithms that reveal relationships in data sets as a collection of photos of faces. Then, the two networks—one a "generator," the other a "discriminator"—are put in competition with one another. The discriminator tries to determine whether the images were produced

***Corresponding Author**: Shereen A. Hussein

Computer Science Department, Faculty of Computers and Artificial Intelligence, Fayoum University, Fayoum, Egypt

Email address: Sam26@fayoum.edu.eg

artificially, whereas the generator attempts to optimize an images of faces as output. The two networks improve one another's capabilities as they compete with one another. The end result is an output that gets better over time and becomes more and more difficult to distinguish fake from real.
DeepFakes are one of the most dangerous forms of misinformation, posing widespread and serious security and privacy dangers to important governmental institutions and common citizens worldwide [2]. Furthermore, adversarial entities use deepFake generation algorithms, which are continually developing, to spread illegal content in a variety of ways, such as ransomware, digital kidnapping, etc. [3]. Different deepFake generation techniques, such as FaceSwap which is used in accessing to authentication systems due to its high ability in exchanging faces in photos and video compositing. Several methods based on machine and deep learning have been developed to detect the fake images [4-10] As opposed to the conventional multi-head self-attention layer, the vision transformer layer gains from the re-attention process.

Author contributions
SH prepared the input image frames and applied the pre-processing including face localization, resizing and normalization. SN built the vision transformer model and performed the DeepFake classification. SH was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

2. Related Work

Lately, media forensics got an enormous attention, in part because of the growing DeepFakes concerns. Since the earliest DeepFakes databases of the first generation were released until the most recent DeepFakes databases of the second generation, numerous visual advancements have been made, creating fake videos practically undetectable to the human eye. After this threat emerged, several deepfake detection techniques have been presented. Convolutional neural networks (CNNs), which are widely used in deepFake detection techniques are used in frame features extraction, Long Short Term Memory (LSTM) for temporal analysis of sequences and Recurrent Neural Network (RNN) to fully use spatial and temporal differences across frames. Other distinct methods have appeared such as Face X-ray for More General Face Forgery Detection and DeepFakes Detection Using Estimated Heart Rate till using vision transformers.
Vision transformers presents extremely encouraging results in less computing time in DeepFakes detection due to its attention mechanism ability to rapidly learn of higher level of information such as hidden traces and intrinsic representations that are lost through the pooling layers of CNNs.

Guera, D., & Delp, E.J. [4] estimated the likelihood for a given video sequence being either DeepFakes or pristine. Multiple consecutive frames' features are analyzed by recurrent neural networks, such as Long Short-Term Memory (LSTM). The InceptionV3 with the fully-connected layer at the top of the network was eliminated to immediately generate a deep representation of each frame using the ImageNet pre-trained model, subsequent to the last pooling layers, 2048-dimensional feature vectors are employed as the sequential LSTM input.

Masi, L., et al. [5] presented a two-branch structure employing a bottleneck layer made of a Gaussian Laplacian (LoG), with one branch propagating the original information and the other suppressing the face content while amplification multi-band frequencies. Using LoG operator, Data from the frequency and color domains can be combined using the representation extractor based on connected dense layers. A cost function that eliminates the unrealistic facial samples from the feature space and reduces the natural faces diversity is

employed for more effective isolation of manipulated faces than more contemporary techniques that identify face changes using binary cross-entropy.

Li, X., et al. [6] suggested Spatial-Multiple Instance Learning (S-MIL) as a solution for the attack for partial faces in DeepFake video detection. The mismatch between faces is intended to be captured via a spatial-temporal instance, which can help DeepFake detect fakes more accurately. First, face is detected in the input videos' sampled frames. Afterwards, a CNN is fed the extracted faces to obtain features as instances. The spatial and temporal instances are retrieved using the corresponding encoding branches to generate spatial-temporal bags with different temporal kernel sizes. Together, these bags serve as a representation of a video. In order to determine the final fake scores for each bag and, by extension, for the entire video, S-MIL is eventually applied to these bags.

Bonettini, N., et al. [7] introduced the ensembling of various trained (CNN) models to detect the manipulation of faces in video sequences. EfficientNetB4 utilizing two distinct concepts Siamese training and attention layering. The Siamese training method improved classification results by focusing on the feature maps extraction with the necessary information of input frame. The attention layering method strengthened the network's training power of the model. This combination has leaded to promising results in detecting face manipulation.

Ismail, A., et al. [8] introduced You only look once a frequent detour Neural networks (YOLO-CRNNs), for detecting deepFake videos. Every frame in the video is analyzed by the YOLO face detector to identify facial areas, and EfficientNet-B5 is fine-tuned to extract the spatial features of these faces. Long-term bidirectional memory (Bi-LSTM) eventually gets these spatial features as a set of input sequences in order to extract temporal features.

Xia, Z., et al. [9] presented amodel is based on MesoNet use of classification, but is used with its pre-treatment model. Since just the face region of the DeepFake video is manipulated, Dlib is used to obtain the face image to mitigate the impact of the background region on subsequent detection. The DeepFake videos facial region is clearly smooth overall. The performance of detection model is enhanced by filtering out the low-frequency information from the facial image while retaining the high-frequency information with high texture discrimination through the use of a novel pre-processing technique.

Elhassan, A., et al. [10] introduced a deep learning method that utilizes mouth and teeth movements to detect fake videos. as a differentiating feature that are still very challenging to handle when making fake videos. The work extended the earlier work of the application of multi-transfer learning approaches including DenseNet121, DenseNet169, vgg16, vgg19, EfficientNetB0, EfficientNetB7, MobileNet, ResNet50, InceptionV3, and Xception to improve the DeepFake videos detection and classification capabilities using the extracted features from the teeth and mouth frames as a biological signal.

Wodajo, D., & Atnafu, S. [30] Presented a model for the detection of DeepFakes that is composed from Vision Transformer (ViT). and Neural Network (CNN). While the ViT gets the learned features as input, the CNN extracts the image's learnable features. After that, it classifies them according to an attention mechanism and decides whether or not a certain video is real.

Passos, L. A., et al. [31] highlighted the most important studies conducted in recent years on deepFake detection using deep learning methods. By training and validating the models on different deepFake datasets, the researches noticed a decline in performance. NeuralTextures manipulation of the FaceForensics++

dataset provided lower accuracy in most related studies which encourages the exploration of more intricate forgery aspects provided through various manipulation methods as well as overcoming obstacles towards generating more precise techniques not only in deepFakes existance detection but in specifying the forged parts in the manipulated images.

Khan, S. A., & Dang-Nguyen, D. T. [32] presented a comparative analysis of different deep learning models (transformers, CNNs) for deepFake detection on different benchmarks. This analysis evaluated the model's performance and examined the generalization capabilities to help in developing more accurate and reliable deepFake detection systems. The inferred results that transformer models perform better than that of the CNN models.

As the prior review, the research community is being interested in developing a reliable and effective methods for DeepFake detection task. However, the widely used detection methods rely on CNN-based structures and localized characteristics. it is noticeable that these models don't perform well on unseen data due to the model generalization challenge. This paper focuses on improving the facial forgery detection using vision transformers.

3. Proposed Model

An integrated vision transformer architectures are applied to the pre-extracted faces. Deep-ViT and Cross-ViT are undergo supervised training to perform the DeepFake detection process whether the face has been manipulated or not. The overall architecture of the proposed model is presented in Fig. 1.

3.1.Data Acquisition

Face Forensics (FF++) dataset [27] as in Fig. 2 is used for this paper. It comprises of 1000 original video clips that have been altered using DeepFakes, Face2Face, FaceSwap and NeuralTextures automated face manipulation methods. The data is obtained from 977 YouTube videos, all of which include a trackable, primarily frontal face without occlusions that allows automated tampering techniques for producing a convincing forgery. CelebDF-V2 [33] dataset consists of 5639 fake videos generated using Encoder-Decoder models that swaps faces of individuals in target and source videos. It also contains 590 real videos are collected from YouTube, and contain interview videos of 59 celebrities having diverse ethnic backgrounds, genders, age groups.

3.2.Data Preprocessing

a. Face Localization and Extraction

Despite of the general image manipulation, the background area is preserved while DeepFake generating techniques are primarily restricted to facial areas. So, each video in the dataset is divided to 10 frames per second and the face areas were cropped from each frame using Dlib 19.24.1 that calibrate face landmarks [28].

b. Image Resizing

A resizing process is a need as the cropped face dimensions are variant from frame to another. 128 * 128 * 3 is the appropriate size for input in the model training process [17].

c. Image Normalization

The process of normalization involves the limitation of an image's pixel values into a predetermined range or distribution. The pixel values are scaled in the range [0, 1] with the intention of improving the quality and comparability of images by eliminating variances in pixel values across or within the same image.

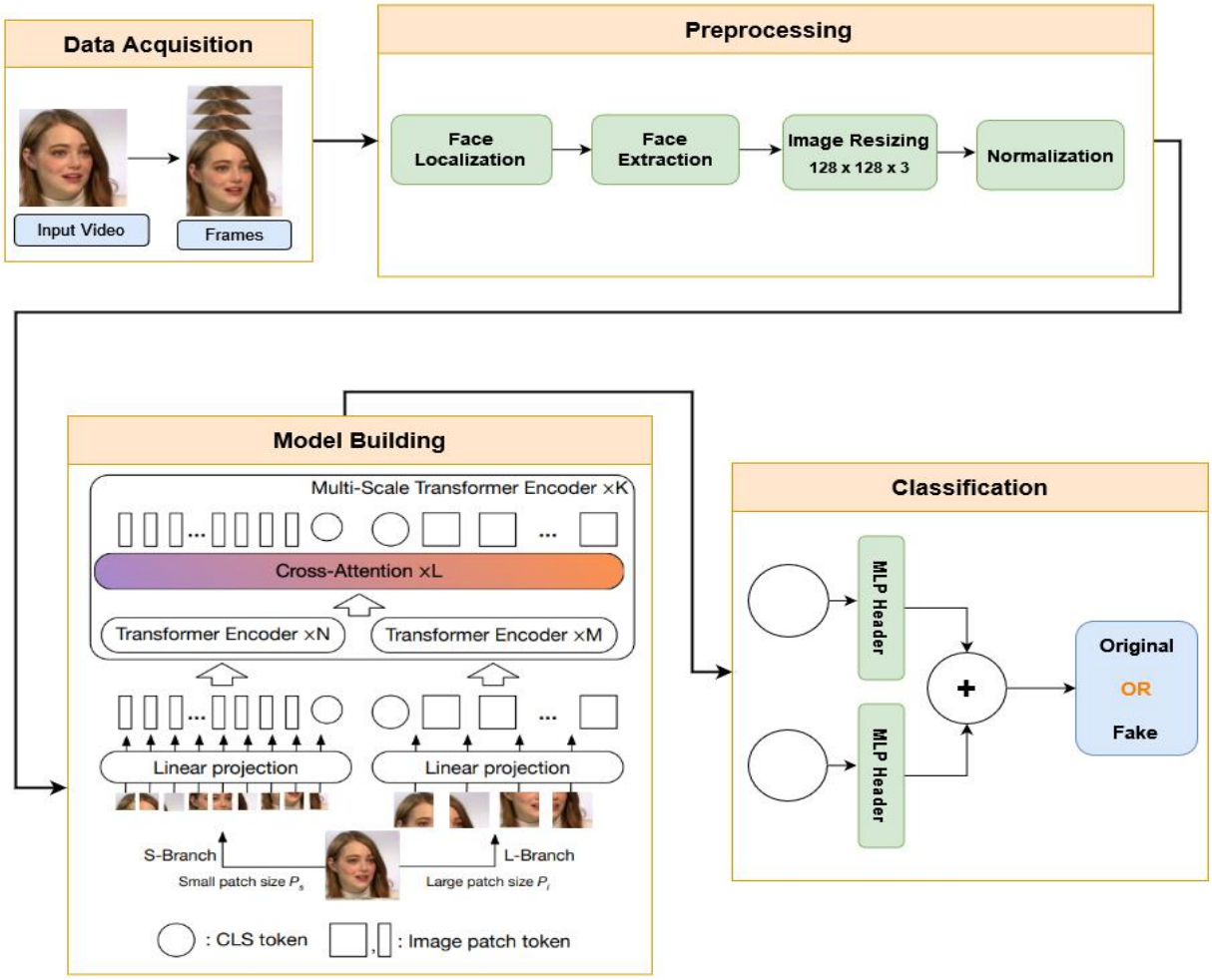


Fig.1 Proposed Model Architecture

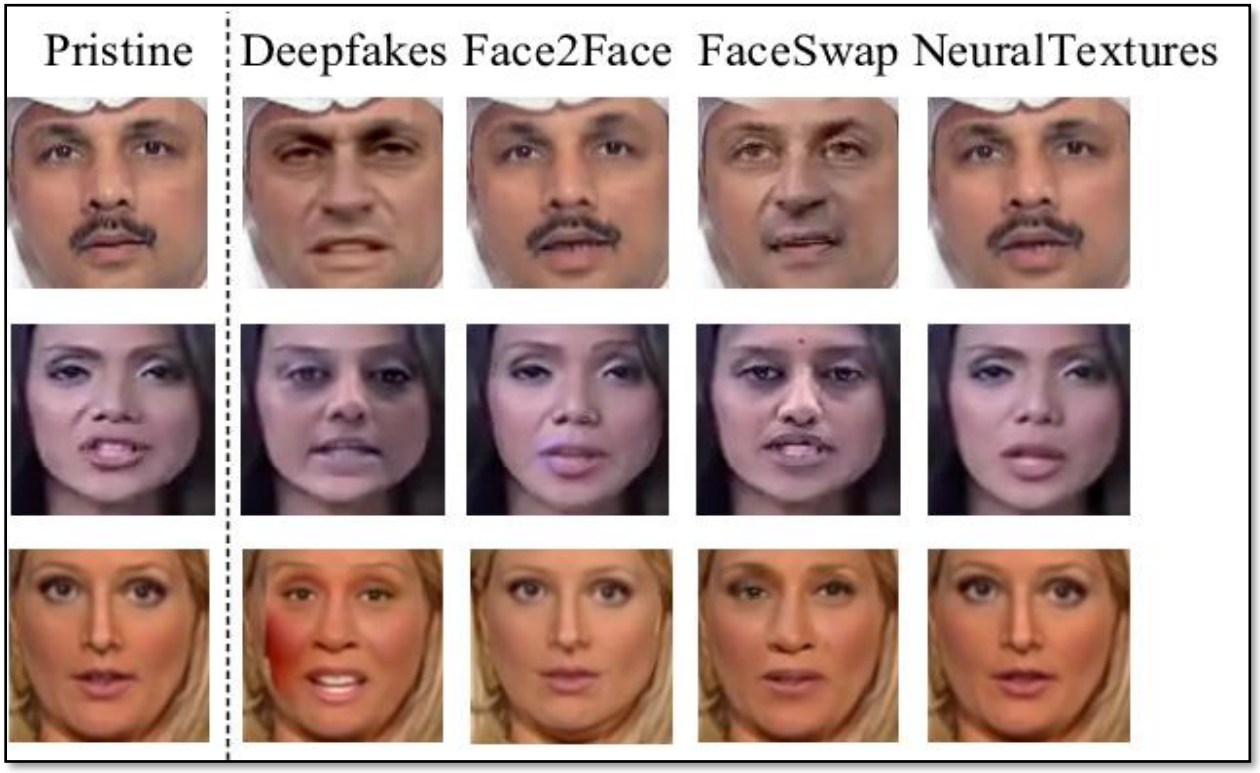


Fig.2 FF++ Dataset Sample [27]

3.3.Building Model

Originally, the Transformer architecture was intended for tasks including natural language processing (NLP), but has since been adapted for use with image data. ViT [22] represents an exciting new direction in the development of deep learning models for image processing, and may have important applications in areas such as computer vision.

In ViT, the input image is divided into fixed-size patches, and each patch is treated as a token, similar to how words are treated as tokens in NLP tasks. These patches are then passed through a series of self-attention layers, which allow the model to attend to different patches and capture long-range dependencies between them. In addition to the self-attention layers, ViT [23] also includes feedforward layers and layer normalization, which help to transform the input data and maintain stability during training as shown in Fig.3. The output of the final self-attention layer is then used as the image representation, which can be fed into a classification head or other downstream tasks.

One advantage of ViT over traditional convolutional neural networks (CNNs) is that it can handle variable-sized inputs without the need for pooling or cropping, which can lead to loss of information. Additionally, ViT has shown state-of-the-art performance on a variety of image classification benchmarks, including ImageNet, with fewer parameters than some competing models.

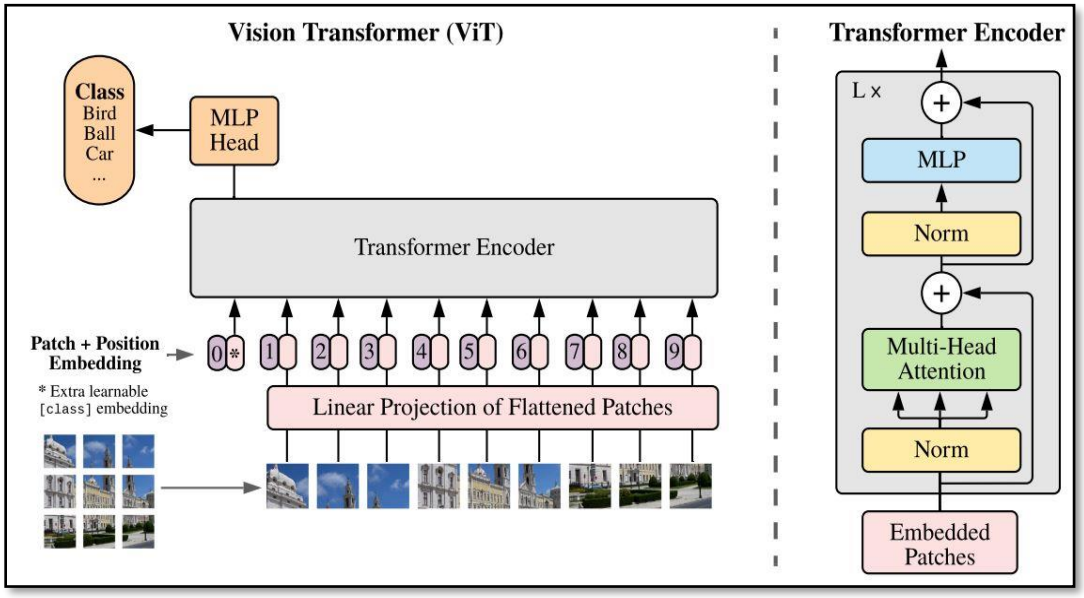


Fig.3 Vision Transformer Structure [23]

The proposed architecture consists of the two architectures of DeepViT [25] and CrossViT [26]. As shown in Fig. 4, DeepViT is a type of vision transformer that, in order to address the problem of attention collapse, swaps out the self-attention layer within the transformer block with a Re-attention module. This allows for the training of deeper ViTs. CrossViT is a specific type of vision transformer that extracts multi-scale feature representations for image classification using a dual-branch architecture.

Stronger visual features for classification of images are produced by combining image patches (or tokens in a transformer) of various sizes. Small and large patch tokens with two distinct branches of various computational complexities are processed, and these tokens are repeatedly fused to complement one another. An effective cross-attention module is used to produce fusion, and in it, each transformer branch creates a non-patch token as an agent of information exchange with the other branch through attention. This enables the attention map to be generated in fusion in linear time as opposed to quadratic time otherwise.

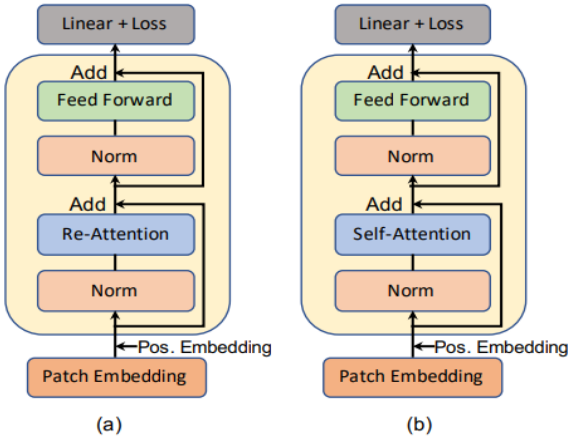


Fig.4 Re-attention vs. Self-attention approach [17]

a. Patch Extraction & Embedding

Patch extraction is the process of dividing an image into smaller, non-overlapping regions called patches. Before analyzing images using transformer models, Patches are needed to be extracted from images in 2 different scale (large branch and small branch), in small branch, patch size was 4 x 4 and projection it to 192 dimensions, and in large branch, patch size is 8 x 8 and projection in to 384 dimensions. Then to maintain positional information of each patch, patch embeddings are integrated with position embeddings, then concat each branch with 'CLS token'.

b. Transformer Encoder with re-attention layer

As demonstrated, the feature maps in the upper layers of deep vision transformer models frequently resemble one another. This indicates that the self-attention mechanism is unable to acquire useful concepts for learning representations [24]. In other words, the attention collapse problem that classic multi-head self-attention layers experience makes it difficult to scale up the vision transformer model and degrades model performance. In contrast to prior visual transformer frameworks, which comprise each transformer encoder of a feed-forward multilayer perceptron and a multi-head self-attention layer, the transformer encoder proposed in this paper composed of both a feed-forward multilayer perceptron and a re-attention mechanism.

c. Cross Attention Layer

The core concept of the proposed cross-attention layer is illustrated by Cross-Attention Fusion, where the fusion involves the CLS token from one branch and patch tokens from the other branch. specifically. To more effectively and efficiently fuse multi-scale features, each branch uses the CLS token as an agent to first exchange information among the patch tokens from the other branch and then back project it to its own branch [23]. As a result of the CLS token's prior learning of abstract information across all patch tokens in its own branch, interacting with the tokens at the other branch permits the incorporation of information at another scale. Following the fusion with other branch tokens, the CLS token engages with its own patch tokens once more at the subsequent transformer encoder. It can now transfer the knowledge it has gained from the other branch to its own patch tokens, improving each patch token's representation. The cross-attention module for the large branch (L-branch) is described in the sections that follow, and the procedure is carried out for the small branch (S-branch) by simply changing the indexes.

3.4. DeepFake Classification

After performing the aforementioned processes, the small branch token and the large branch token were projected in a one dimension then a sigmoid function is applied to obtain a value v between 0 and 1. If the number is more than or equal to .5 ($v >= .5$), the frame is real; else, it is fake ($v < .5$).

4. Discussion & Experimental Results

This paper demonstrates that the highest accuracy is achieved with FF++ dataset using VIT model. All experiments are implemented using Google Cloud , specifically a Google Cloud VM instance with 200 GB SSD storage, 16 GB RAM, and a 4-core CPU running on Windows server. A Cloud TPU v2-8 device, connected in the (us-central1-f) zone is integrated to further computational power and acceleration enhancements. Accessing the TPU through Google Cloud's TPU Research Cloud program, has allowed leveraging its advanced processing capabilities, contributing to the success of the model learning endeavors. The performance of the proposed model is evaluated by these measures' precision (1), recall (2), F1-score (3), accuracy (4), area under curve (5), and loss (6) in the testing phase equations shown in Tables 1 , 2 and 3. Moreover, Loss training and validation are calculated during different epochs as shown in Figs. 5, and 6. Also, the proposed model performance is compared with other models that work on the same dataset using two ways in classification as shown in Tables 4, 5 and 6.

Table 1: Performance Equations Summary

Assessments	Equation	Equ. No	Assessments	Equation	Equ. No
Precision (P)	$\frac{TP}{TP+FP}$	(1)	Accuracy (Acc)	$\frac{TP+TN}{TP+TN+FP+FN}$	(4)
Recall (R)	$\frac{TP}{TP+FN}$	(2)	AUC	$\frac{1+(\frac{TP}{TP+FN})-(\frac{FP}{TN+FP})}{2}$	(5)
F1-Score	$2*\frac{P*R}{P+R}$	(3)	Loss	$-\frac{1}{N}*\sum_{i=1}^{N}\mathrm{Yi}*\log(\mathrm{P(Yi)})+(1-\mathrm{Yi})*\log(1-\mathrm{P(Yi)})$	(6)

Where:
True Positive (TP) the model predicts the positive class correctly. True Negative (TN) model correctly classifies the negative class. In a false positive (FP), the model predicts the positive class incorrectly. In false negative (FN), the model predicts the negative class incorrectly.
Also, Yi represents the actual class and log(P(Yi) is the probability of that class in total N data values.

Table 2 and Table 3: Performance Measure Values

Type	Class 0 (Fake)			Class 1 (Real)		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
DeepFake	0.97	0.92	0.95	0.93	0.97	0.95
FaceSwap	0.97	0.92	0.95	0.93	0.97	0.95
Neural Textures	0.93	0.89	0.91	0.89	0.93	0.91
Face2Face	0.93	0.73	0.82	0.73	0.93	0.82
Overall	0.92	0.72	0.81	0.77	0.94	0.85

Type	Acc	AUC	Loss
DeepFake	0.95	0.99	0.15
FaceSwap	0.95	0.99	0.14
Neural Textures	0.82	0.90	0.44
Face2Face	0.91	0.97	0.23
Overall types	0.83	0.92	0.37

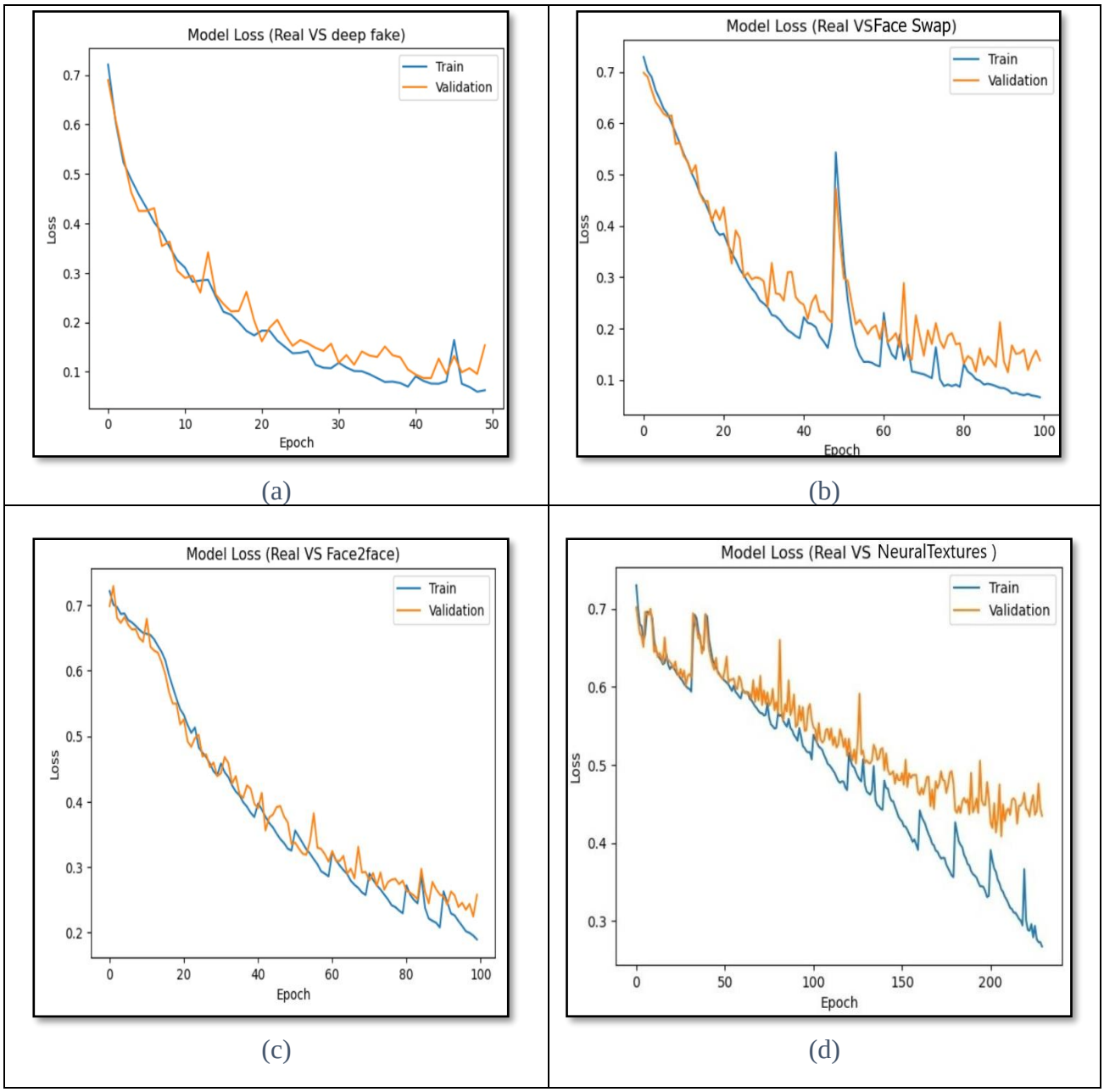


Fig.5 Loss of training and validation for each different types (a) DeepFake, (b) FaceSwap, (c) Face2Face, and (d) Neural Textures

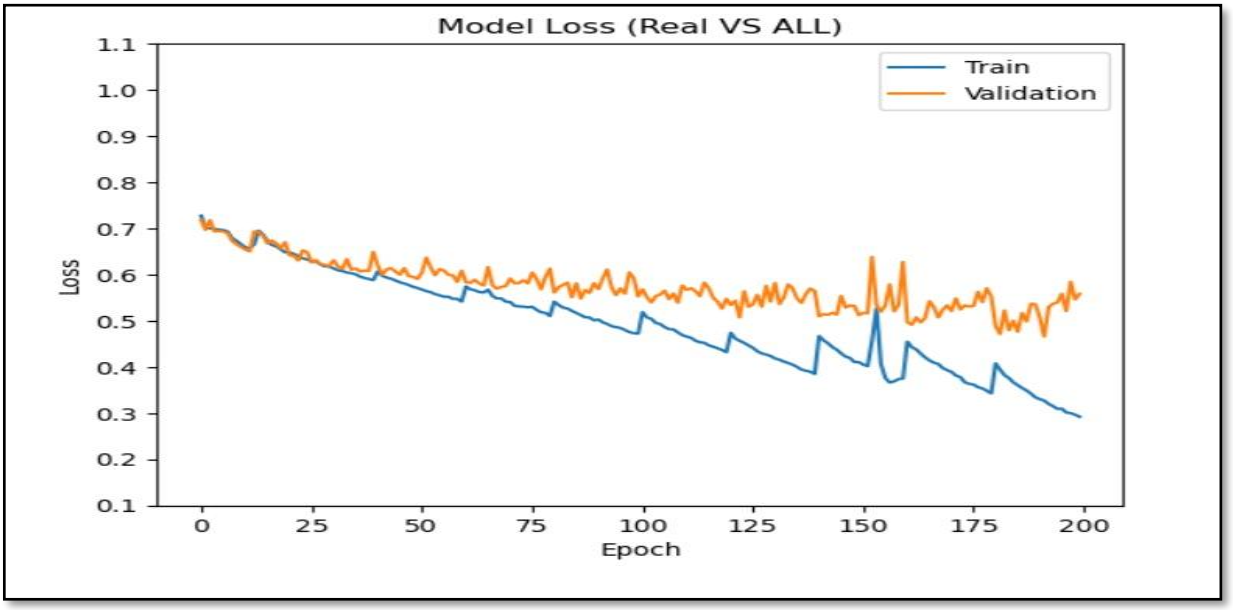


Fig.6 Loss of training and validation for all FF++ Dataset

5. Conclusion

DeepFake technology is no longer commensurate to encourage and believe in its use as its impact on people trust erosion in the realism of media contents. one of its neagative effects, ability to cause distress to targeted people, heighten disinformation and increase hate speech, and even could stimulate political tension, inflame the public. Nowadays, it is a paramount essential to detect the content manipulation with its various types due to the increasing of approaches of DeepFake creation and the quick spread in various media platforms. This paper presents a deepFake detection framework combines CrossVIT and DeepVIT, utilizing vision transformer's unique characteristics to model both local image features and global pixel relationships at the same time in contrast of other deepfake detection techniques that use CNNs as their foundation. This paper can efficiently capture varying scales of alterations because of its multi-stream design. The proposed model classification has been applied in different experiments and its performance in both intra-dataset (trained and evaluated on same dataset) and inter-dataset (trained on one dataset and evaluated on the remaining datasets excluding the training dataset) settings are compared with other models in Tables 4, 5 and 6. First, the overall classification model has achieved 92.4% and 83.1% AUC on FF++ and Celeb-DF (V2) datasets respectively. Secondly, the model is trained with samples of all types of manipulated videos of FF++ deepFake detection datasets and evaluated 88.9% AUC on Celeb-DF (V2) dataset. As in FF++, there are four types of manipulated videos such as deepFake, faceswap, face2face and neural texture. finally, the model is trained on three out of these four types of the manipulated videos and tested on the remaining one. It achieved 98.6 %, 98%, 97% and 90.3% respectively in subclasses classification.

a. Over-all classification

Table 4: AUC (%) Results on overall FF++ *and* Celeb-DF (V2) datasets

Models	Face Forensics++	Celeb-DF (V2)
VA-MLP [18]	66.4	55
Multi-task [19]	76.3	54.3
FWA [20]	80.1	56.9
Meso4 [21]	84.7	54.8
DSP-FWA [20]	93	64.6
TBRN [5]	93.2	73.4
Proposed Model	**92.4**	**83.1**

Table 5: AUC (%) Results on Celeb-DF (V2) when trained on FF++ dataset

Type Model	Celeb-DF (V2)
Xception [11]	73.7
CNN-AUG [12]	75.6
Patch-based [13]	69.6
Face X-ray [14]	79.5
CNN_GRU [15]	69.8
Lip Forensics [16]	82.4
DFDT [17]	88.3
Proposed Model	**88.9**

b. Subclasses Classification

Table 6: AUC (%) Results on each subset in FF++ dataset

Type Models	DeepFake	Face Swap	Face2face	Neural Textures
Xception [11]	93.9	51.2	86.8	79.7
CNN-AUG [12]	87.5	56.3	80.1	67.8
Patch-based [13]	94	60.5	87.3	84.8
Face X-ray [14]	99.5	93.2	94.5	92.5
CNN_GRU [15]	97.6	47.6	85.8	68.6
Lip Forensics [16]	99.7	90.1	99.7	99.1
DFDT [17]	99.8	93.1	99.6	99.2
Proposed Model	**98.6**	**98**	**97**	**90.3**

References

1. Mirsky, Y.; Lee, W. “The creation and detection of deepfakes: A survey”. ACM Computing Surveys (CSUR), 2021,
2. Vaccari, C.; Chadwick, A.”Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news”. Social Media+ Society, vol. 6, no. 1, 2020,
3. Maras, M.H.; Alexandrou, A. “Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos”. The International Journal of Evidence & Proof , vol. 23, no. 3, 2019.

4. Güera, D.; & Delp, E. J. “Deepfake video detection using recurrent neural networks”. In Proceedings of the 15th IEEE international conference on advanced video and signal based surveillance (AVSS) 2018.
5. Masi, I.; Killekar, A., Mascarenhas, R. M., Gurudatt, S. P., & AbdAlmageed, W. "Two-branch recurrent network for isolating deepfakes in videos." In the proceedings of the 16th European Conference Computer Vision–ECCV, Part VII, 2020.
6. Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; & Lu, Q. “Sharp multiple instance learning for deepfake video detection”. In Proceedings of the 28th ACM international conference on multimedia 2020.
7. Bonettini, N.; Cannas, E. D.; Mandelli, S.; Bondi, L.; Bestagini, P.; & Tubaro, S. "Video face manipulation detection through ensemble of cnns." In Proceedings of the 25th international conference on pattern recognition (ICPR), 2021.
8. Ismail, A.; Elpeltagy, M.; Zaki, M.; & ElDahshan, K. A. “Deepfake video detection: YOLO-Face convolution recurrent approach”. PeerJ Computer Science, 2021.
9. Xia, Z.; Qiao, T.; Xu, M.; Wu, X.; Han, L.; & Chen, Y. “Deepfake video detection based on MesoNet with preprocessing module”. Symmetry, vol.14, no. 5, 2022.
10. Elhassan, A.; Al-Fawa'reh, M.; Jafar, M. T.; Ababneh, M.; & Jafar, S. T. "DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning." SoftwareX 19, 2022.
11. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; & Nießner, M."FaceForensics++: Learning to detect manipulated facial images", IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
12. Wang, S. Y.; Wang, O.; Zhang, R.; Owens, A.; & Efros, A. A."CNN-generated images are surprisingly easy to spot… for now", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
13. Chai, L.; Bau, D.; Lim, S. N.; & Isola, P."What makes fake images detectable? understanding properties that generalize", Computer Vision – ECCV, 2020.
14. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; & Guo, B."Face X-ray for more general face forgery detection", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
15. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; & Natarajan, P. "Recurrent convolutional strategies for face manipulation detection in videos.", Interfaces (GUI), vol.3, no.1, 2019.
16. Haliassos, A.; Vougioukas, K.; Petridis, S.; & Pantic, M."Lips don’t lie: A generalisable and robust approach to face forgery detection", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2021.
17. Khormali, A.; & Yuan, J.-S."DFDT: an end-to-end deepfake detection framework using vision transformer." Applied Sciences, vol.12, no. 6, 2022.
18. Matern, F.; Riess, C.; Stamminger, M. “Exploiting visual artifacts to expose deepfakes and face manipulations”. In Proceedings of the IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.
19. Nguyen, H.H.; Fang, F.; Yamagishi, J.; Echizen, I. “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos”. In Proceedings of the IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2019.
20. Li, Y.; & Lyu, S. “Exposing DeepFake Videos by Detecting Face Warping Artifacts”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

21. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I.. “Mesonet: A compact facial video forgery detection network”. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 2018.
22. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. "Transformer in transformer." Advances in Neural Information Processing Systems, 2021.
23. Dosovitskiy, A.; Beyer, L..; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J. & Houlsby, N. “An image is worth 16x16 words: Transformers for image recognition at scale”. arXiv:2010.11929, 2020.
24. Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; & Gao, W. "Pre-trained image processing transformer." In the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.
25. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; & Feng, J. "Deepvit: Towards deeper vision transformer." arXiv:2103.11886, 2021.
26. Chen, C. F. R.; Fan, Q.; & Panda, R. “Crossvit: Cross-attention multi-scale vision transformer for image classification”. In Proceedings of the IEEE/CVF international conference on computer vision, 2021.
27. Papers with code - faceforensics++ dataset Available at: https://paperswithcode.com/dataset/faceforensics-1.
28. Dlib C++ library. Available at: http://dlib.net/.
29. Vision Transformer (VIT). Available at: https://huggingface.co/docs/transformers/model_doc/vit.
30. Wodajo, D., & Atnafu, S., “Deepfake video detection using convolutional vision transformer”. arXiv preprint arXiv:2102.11126, 2021.
31. Passos, L.A., Jodas, D., da Costa, K.A., Júnior, L.A.S., Colombo, D. & Papa, J.P., “A review of deep learning-based approaches for deepfake content detection”. arXiv preprint arXiv:2202.06095, 2022.
32. Khan, S. A., & Dang-Nguyen, D. T., “Deepfake Detection: A Comparative Analysis”. arXiv preprint arXiv:2308.03471. 2023.
33. 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, 2020.