ResViT: A Framework for Deepfake Videos Detection

Original Scientific Paper

Wasim Ahmad
Institute of Information Sciences,
Academia Sinica, Taiwan
Department of Computer Science,
National ChengChi University, Taiwan
was_last@iis.sinica.edu.tw

Imad Ali
Department of Computer Science,
University of Swat, KP, Pakistan
imadali@uswat.edu.pk

Sahibzada Adil Shahzad
Institute of Information Sciences,
Academia Sinica, Taiwan
Department of Computer Science,
National Chengchi University, Taiwan
adilshah275@iis.sinica.edu.tw

Ammarah Hashmi
Institute of Information Science,
Academia Sinica, Taiwan
Institute of Information Systems and Applications,
National Tsing Hua University, Taiwan
hashmiammarah0@gmail.com

Faisal Ghaffar
System Design Engineering Department,
University of Waterloo, Canada
faisal.ghaffar@uwaterloo.ca

Abstract *– Deepfake makes it quite easy to synthesize videos or images using deep learning techniques, which leads to substantial danger and worry for most of the world's renowned people. Spreading false news or synthesizing one's video or image can harm people and their lack of trust on social and electronic media. To efficiently identify deepfake images, we propose ResViT, which uses the ResNet model for feature extraction, while the vision transformer is used for classification. The ResViT architecture uses the feature extractor to extract features from the images of the videos, which are used to classify the input as fake or real. Moreover, the ResViT architectures focus equally on data pre-processing, as it improves performance. We conducted extensive experiments on the five mostly used datasets. Our analysis revealed that ResViT performed better than the baseline and achieved the prediction accuracy of 80.48%, 87.23%, 75.62%, 78.45%, and 84.55% on Celeb-DF, Celeb-DFv2, FaceForensics++, FF-Deepfake Detection, and DFDC2 datasets, respectively.*

Keywords*: deepfake, detection, GAN, vision transformer*

1. INTRODUCTION

Deepfake is creating and manipulating videos, audio, or images produced by deep learning methods and techniques that appear real [1]. Lip-sync [2], puppet-master [3], and face swap [4] are some of the techniques used for synthetic video, image, and speech generation. With such technological advancement, the creation of deepfake videos, audio, and images is rising [5, 6]. According to the report of DeepTrace [7], in September 2019, approximately fifteen thousand fake videos were found, which was about two times higher than the previous year. These included about 96% pornographic, while 99% were female celebrities whose faces were mapped on porn stars. Such deepfake videos may target famous personalities to denigrate a person, resulting in devastating damages. For example, deepfake videos can be used to destabilize the reputation of a political candidate by making the candidate appear to say or do things that never actually occurred.

Researchers are developing robust algorithms for differentiating real videos from fake ones to prevent this hazardous threat to society. For example, [8] and [9] tried to discover discrepancies in eye blinking for deepfake detection. To simulate eye blinking, [8] proposed a model that can be used to spawn the appearance of a face from a portrait. The same problem was also addressed by [9],

recommending a model that produces speaking videos with heads using facial expressions like eyes blinking. Brockschmidt *et al.* [10] proposed a facial forgery detection model for the detection of various spoofing methods, which helps detect reliably those detection methods that are invisible. FakeCatcher [11] is a deepfake detection technique that uses biological signals representing internal synthesizers and image generators. A convolution neural network model is proposed in [12] to identify the inconsistencies created during the creation of deepfakes.

However, these existing models for deepfake detections mainly focus on their architectures and ignore the importance of data pre-processing, which may improve the model performance [13]. Thus, training a model with proper data pre-processing techniques for detecting deepfake videos with higher accuracy. Moreover, these techniques lack generalizability for detecting deepfake. However, deep neural networks (DNNs) have shown superior performance in image classification compared to shallow layers [14, 15]; thus, carefully training a DNN model can get maximal deepfake artifacts for detecting deepfake videos with higher accuracy.

In this article, we propose, **ResViT**, which combines the **Res**Net model with the **V**ision **T**ransformer to identify deepfake videos efficiently. The ResViT has generalized architecture as it extracts all local and global features of videos' frames (images) via ResNet and classifies it as fake or real via the attention mechanism of the vision transformer. Also, the ResViT architectures focus equally on data pre-processing, as it improves performance. Moreover, we train the proposed ResViT model on a diverse set of face images using the largest dataset currently available to detect deepfakes created in different settings, environments, and orientations. To evaluate ResViT, we conducted extensive experiments on the five mostly used datasets of Celeb-DF, Celeb-DFv2, FaceForensics++, FF-Deepfake Detection, and DFDC2. Our analysis revealed that ResViT performed better than the baseline and achieved the prediction accuracy of 80.48%, 87.23%, 75.62%, 78.45%, and 84.55% on Celeb-DF, Celeb-DFv2, FaceForensics++, FF-Deepfake Detection, and DFDC2 datasets, respectively. The main contributions of this article are as follows:

1. We propose the ResViT framework, which combines the ResNet model with the vision transformer to identify deepfake images efficiently.
2. We propose not only to focus on the architectures of ResViT but also on data pre-processing, as it improves performance.
3. We propose we train the ResViT model on a diverse set of face images using the largest dataset currently available to detect deepfakes created in different settings, environments, and orientations.

The rest of the article is organized as follows: Section 2 presents the literature review, while Section 3 presents the proposed framework. The experiments and results are described in Section 4. Section 5 concludes this article.

2. LITERATURE REVIEW

Deepfake can be created by switching two different identities in the visual stream, i.e., image or video (sequence of images). FakeApp [16] is the first deepfake technique that uses two autoencoders (AE) networks. An AE is an encoder-decoder architecture, Feedforward Neural Network (FFNN), that is trained self-supervised to reconstruct the input stream. The encoder downsamples the input in FaceApp and converts it to a latent representation called latent face features. The decoder mirrors the encoder and works reverse to upsample the latent representation to reconstruct the face images [17]. Face synthesis and face swapping are techniques used for fake videos. Using the face synthesis technique, it's possible to create unseen realistic images from training examples [18]. Application of Image synthesis and, more specifically, face image synthesis are face frontalization, face aging, and pose-guided generation.

Face synthesis can be done by generative adversarial networks (GANs), where we create a generative model responsible for creating a realistic face image. GAN-based architectures, e.g., StyleGAN [19], produce more realistic images that resemble the original images. There is a technique called FaceSwapping, which is a generative adversarial network-based method to generate deepfake videos. Face swap is the process of swapping or inserting the facial identity of the source image into the target image. This fake generation is used to insert actors in different video clips [12]. Traditional computer vision techniques and GANs based approaches synthesize face swaps. FSGAN and RSGAN are also used to perform face-swapping tasks.

Similarly, Face expressions can also be exchanged among individuals. The Face2Face technique manipulates facial expressions and projects source images onto some target faces in almost real-time without delay [20]—Face2Face synthesis images under different lighting and environmental conditions. The deep learning techniques for deepfake video detection have three main categories [21]. The first set of methods focuses on the psychological and physical behavior of the videos. It includes head pose movement and tracking eye blinking. The second type focuses on GANs' fingerprints and biological signals. The last category is data-driven and focuses on visual artifacts. [22] also proposed a CNN model that leverages the image transformation or augmentation (i.e., rotation, scaling, and shearing). A novel approach based on deep learning to detect forged videos was proposed in [23]. They mainly focus on facial reenactments, face swapping, replay attacks, and computer-synthesized image spoofing.

Transformers architecture is usually used for language processing-related tasks, and an immense number of its applications are available in the literature on natural language processing tasks. On the contrary, its application remains limited in image processing, video processing, and computer vision research. [24], shows that CNN-based architecture can be replaced by its alternative,

so-called transformers, which perform better. It outperformed as compared to state-of-the-art CNNs while requiring fewer computational resources for model training. Deepfake is synthesized by autoencoder (encoder + decoder) and generative adversarial models [25].

Unlike these works, we utilized ResNet, a more general model, for feature extraction and integrated it with the vision transformer. We also focus on data processing, making it easy for the vision transformer in classification.

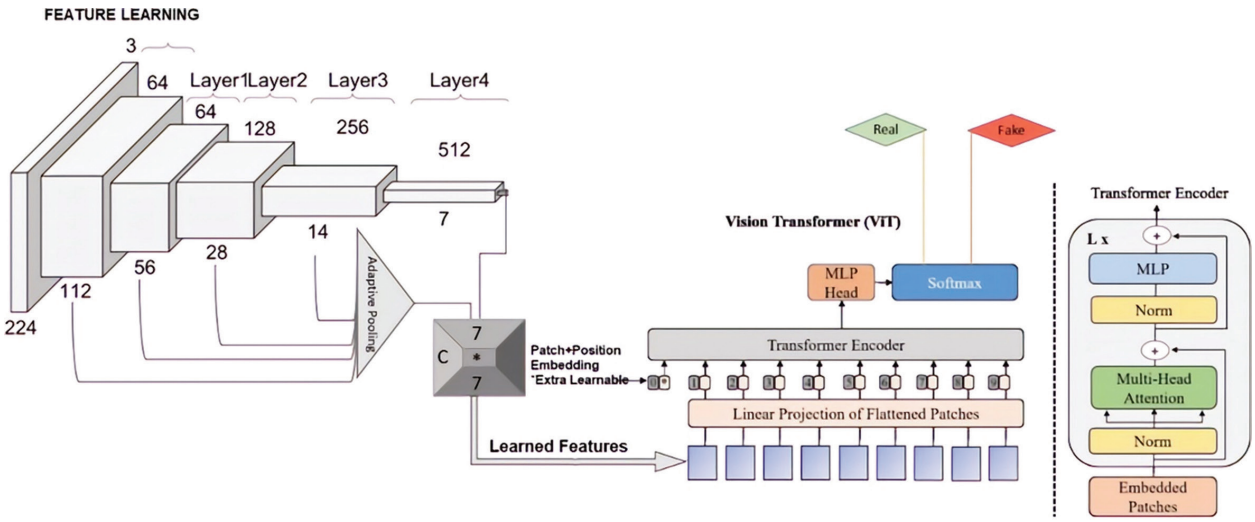


Fig. 1. The ResViT architecture.

Dataset	Generate Method	Total Videos	Actors	Train Images	Validation Images	Test Images
Celeb-DF	Deepfake	1203	13	10592	1245	1245
Celeb-DF-V2	Deepfake	6229	59	65936	7756	7756
FaceForensics++	Deepfake	2000	977	32499	3823	3823
FF-DFDC	Deepfake	1803	977	39727	4673	4673
DFDC2	Deepfake	2656	28	21088	2840	2840
Combined (FF)	Deepfake	2803	-	139879	16809	16809

Table 1. Datasets for deepfake detection

3. ResViT ARCHITECTURE

Fig. 1 shows the architecture of ResViT. The ResViT has feature learning and classification components. In the proposed ResViT, we use the ResNet 18 model to extract features from the images. The ResViT architecture works in a two-stage mechanism. Firstly, ResNet is used for extracting the features from the images. As shown in Fig. 1, we use the ResNet model to extract features from the images. We modify the ResNet model intermediate layers to get better image features. Each layer's output is the input of the next layer. We did not use the model's average pooling layer, flattening, and fully connected layers. Since we only need to extract the features, we do not use a fully connected layer. Finally, we return all four outputs and concatenate them. We reduce the dimensions to (Channels*7*7) and concatenate them. We apply different combinations of concatenating the outputs of the layer. Still, in some cases, we get a better model performance by concatenating the output of the first and last layer (x1 and x4) and the dimensions for that (3072*7*7).

Secondly, we used the vision transformer [24] for classification. Most natural language processing tasks use transformers, primarily for sequential tasks. After better performance on many tasks, the transformer is also thought to be used for computer vision tasks. The vision transformers follow the mechanism of the earliest transformer with some minor input signal adjustment. These are the main components of our model ResViT. After we get the features from the ResNet, we map all those features to the transformer. Transformers take the input image in patches. Therefore, we need to divide our image into patches. We split the feature map into seven patches that are not fixed, and one can use any patch size. The patches are then entrenched into a linear sequence with the dimension of 1*1024. We need to perform position embedding so that each patch can be placed after the other. Therefore, we need to divide our image into patches. We split the feature map into seven patches that are not fixed, and one can use any patch size. The patches are then entrenched into a linear sequence with the dimension of 1*1024. We need to perform position embedding so that each patch can be placed after the other. Therefore, the patches are further added up into position embedding.

The dimension for position embedding then becomes 2*1024 in this case. Compared to the original

transformer, the vision transformer uses only the encoder. So, the patch and position embeddings are forwarded to the vision transformer. The transformer encoder has two blocks: Multiheaded Self-Attention (MSA) and Multi-Layer Perceptron (MLP), whose head and the job are similar to conventional CNN, as shown in Fig. 1. The input dimension has 2048 channels while the out channels are two, which signifies the two classes (fake and real). It has almost 40 million learnable parameters, and to get the final output, the MLP head has applied SoftMax, which alleviates the weight values between 0 and 1. It consists of a Feed Forward Network and is followed by a norm layer to normalize the interior layer. There are eight heads in the transformer. ReLU nonlinearity and a couple more layers are part of MLP.

4. EXPERIMENTS

In this section, we present the experimental setup to implement the model. We deliver the results achieved by our model implementation and interpret the experimental results.

4.1. DATASET

Deep learning models learn from data; thus, the dataset must be carefully prepared for higher prediction accuracy. Therefore, we pre-process the data so that the model learns all the features correctly. We use a couple of libraries like BlazeFace and MTCNN for face extraction, known as the rapid processing of large amounts of images. The dimensions and format of the extracted images are 224*224 and JPEG, respectively. Some examples of face extraction from real and fake datasets can be seen in Fig. 2.

(a) Fake

(b) Real

Fig. 2. Extracted frames from videos.

We split the datasets into training, validation, and testing. The data is then augmented using the library called Albumentation, known for transforming a huge amount of data. We train our model first to detect whether the video is real or fake. To make our model perform better, we train our model with large data. We tried to split the dataset so that maximum data could be provided for training the model. It took a lot of time to train the model and finetune it for better performance. After training the model, the validation data is provided, which is unseen data, and the model is finetuned repeatedly. Finally, the testing data is provided to the model to evaluate and predict whether our model classifies the video in relative class(fake/real) or not. We used Celeb-DF [2], Celeb-DFv2 [3], Faceforensics++ [4] and Deepfake Detection Datasets to evaluate our model. We combine the FaceForensic Deepfake and Deepfake Detection datasets. The number of training, validation, and testing images for each dataset is shown in Table 1.

4.2. EVALUATION

Before feeding our dataset to ResViT, we normalized and augmented the dataset at each iteration of the training phase. We use the learning rate of 0.001 and weight decay of 0.000001 for ten epochs. Once the model is trained, 30 images are forwarded to the model for the classification process. We calculate the accuracy of our model by using the log loss function. We used a binary cross-entropy function for calculating the loss. The purpose of the log loss function is to calculate the probability distribution between 0 and 1. The real class is represented with the value of $0 > y < 0.5$, while the fake class is represented with the value of $0.5 >= y < 1$. For a fair comparison with the baseline, *CViT* [26], we trained our model with a batch size of 8 and 10 epochs under the same settings. The baseline model uses the VGG-16 architecture as a feature extractor and transformer as a classifier, using the DFDC dataset released by Facebook. We have limited resources; therefore, we did not use the DFDC dataset, which is approximately 470 Gigabytes.

We demonstrate our results by calculating the accuracies and losses for all datasets. We trained the baseline and the proposed models on all the datasets mentioned above and compared their results.

Fig. 3 shows the performance comparison of the ResViT on different datasets. The proposed ResViT performs better on each dataset and has achieved the prediction accuracy of 80.48%, 87.23%, 75.62%, 78.45%, and 84.55% on Celeb-DF, Celeb-DFv2, FaceForensics++, FF-Deepfake Detection, and DFDC2 datasets, respectively. The overall prediction accuracy of the ResViT on the combined datasets is 74.54.

Fig. 4 shows the performance comparison of the baseline and proposed ResViT under the same circumstances and resources. The figure shows that CViT has

achieved the prediction accuracy of 71.04%, 84.50%, 71.67%, 73.31%, and 73.4%2, on Celeb-DF, Celeb-DFv2, FaceForensics++, FF-Deepfake Detection, and DFDC2 datasets, respectively, while the ResViT achieved higher accuracies of 80.48%, 87.23%, 75.62%, 78.45%, and 84.55%, on the same datasets. The overall prediction accuracy of the CViT on the combined datasets is 68.37%, while the ResViT has 74.54% prediction accuracy. The ResViT prediction accuracy is 15.79% higher than the CViT on the combined dataset. This is because our proposed ResViT architecture utilizes the ResNet model for feature extraction, which has better generalization than the CNN model. Moreover, the proposed ResViT architecture focuses on pre-processing, which results in higher predictions than the baseline. Since we used the same settings as CViT, thus when ResViT results are better than the CViT, it automatically performs better than other baselines of the CViT.

Moreover, we present the training and validation losses and accuracy on three sample datasets for the proposed ResViT in Fig. 5 for the different number of epochs. Although we observe that the results improve with more epochs, we stick to the original settings of epochs of the CViT. We observe that the ResViT achieved better performance on all three sample datasets, as shown in Fig. 5 (a-e).

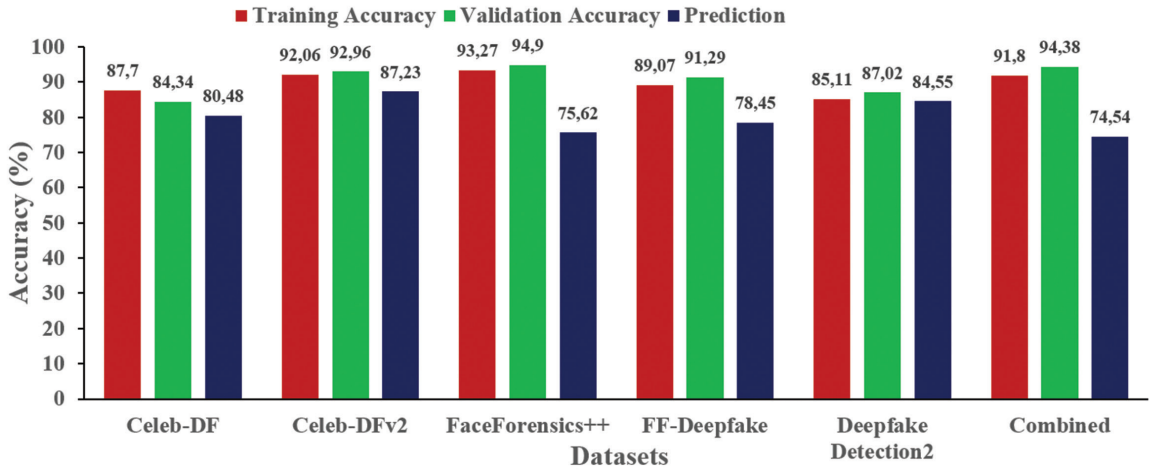


Fig. 3. ResViT performance on all datasets

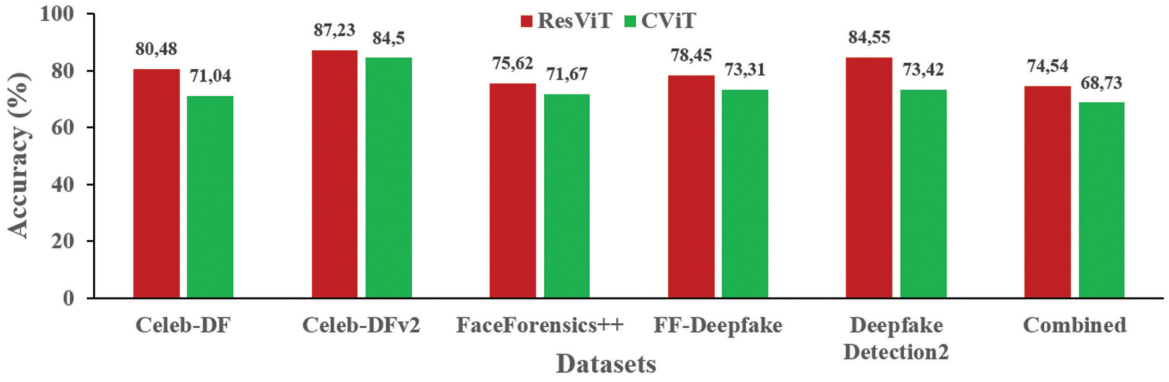
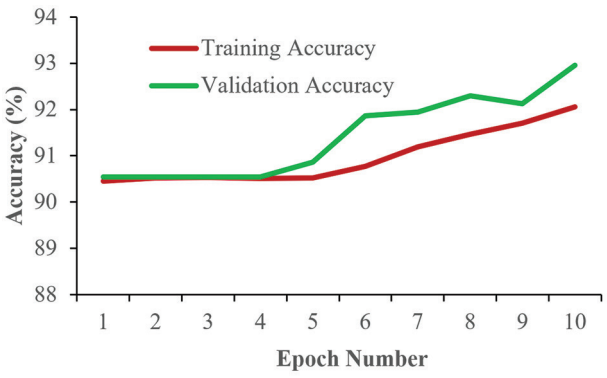
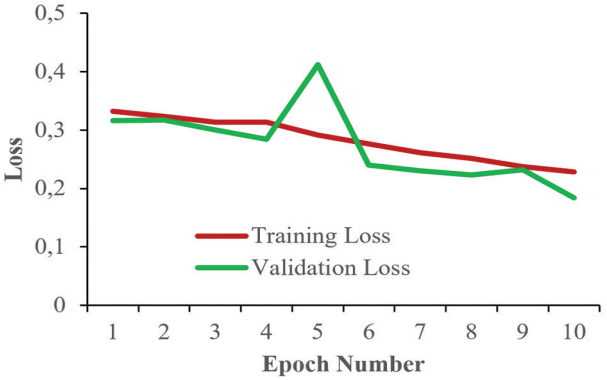


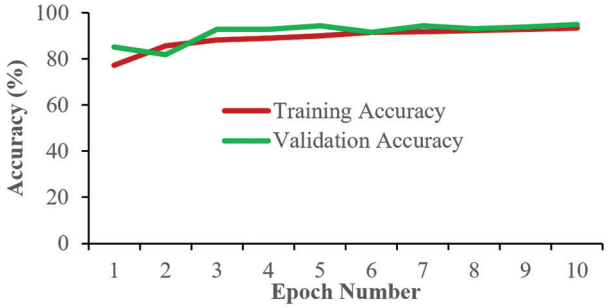
Fig. 4. ResViT and CViT prediction performance on all datasets.



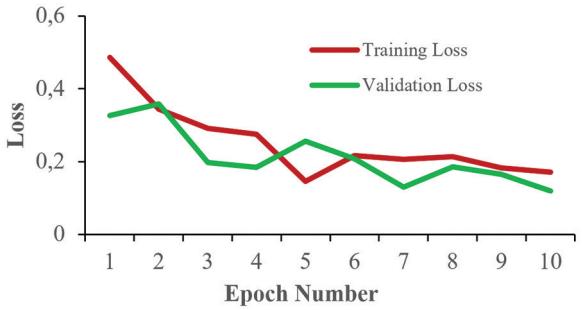
(a) Celeb-DFV2 training and validation accuracy



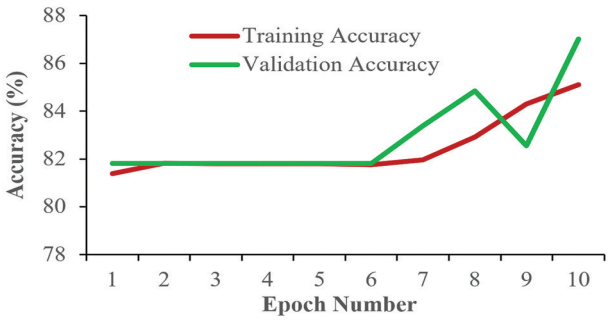
(b) Celeb-DFV2 training and validation loss



(c) FaceForensics++ training and validation accuracy



(d) FaceForensics++ training and validation loss



(e) DFDC2 training and validation accuracy

(f) DFDC2 training and validation loss

Fig. 5. ResViT validation losses and accuracy on three sample datasets for the different number of epochs.

5. CONCLUSION

In this article, we proposed ResViT, which combines the ResNet model with the Vision Transformer to identify deepfake videos efficiently. ResViT extracts all local and global features of videos via ResNet and classifies them as fake or real via the attention mechanism of the vision transformer. ResViT not only focuses on its architecture but also on pre-processing, which adds to higher prediction performance. We evaluated ResViT and baseline in the same settings with extensive experiments on the five mainly used datasets in deepfakes. We find that the proposed ResViT performs better than the baseline. We anticipated the better performance of ResViT, as the ResNet model has better generalization in feature extraction, and the pre-processing adds to prediction performance. Thus, such technology should be used to protect people, especially celebrities and politicians. In the future, we are determined to check the performance of the ResViT under massive datasets with more baseline models.

6. REFERENCES

[[1] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, "Semantic Image Synthesis With Spatially-Adaptive Normalization", Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019, pp. 2337-2346.

[2] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, C. Jawahar, "Towards Automatic Face-To-Face Translation", Proceedings of the 27th ACM International Conference on Multimedia, Nice France, 21-25 October 2019, pp. 1428-1436.

[3] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync From Audio", ACM Transactions on Graphics, Vol. 36, No. 4, 2017, pp. 1-13.

[4] Y. Nirkin, Y. Keller, T. Hassner, "FSGAN: Subject Agnostic Face Swapping and Reenactment", Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October - 2 November 2019, pp. 7184-7193.

[5] B. Chesney and D. Citron, "Deep Fakes: A Looming Challenge For Privacy, Democracy, And National Security", The California Law Review, Vol. 107, 2019, pp. 1753.

[6] E. Hsiang, "Deepfake: An Emerging New Media Object in the Age of Online Content", Boston University School of Law, 2020, Master Thesis.

[7] L. Zheng, Y. Zhang, V. L. Thing, "A Survey on Image Tampering and Its Detection in Real-World Photos", Journal of Visual Communication and Image Representation, Vol. 58, No. 1, 2019, pp. 380-399.

[8] H. X. Pham, Y. Wang, V. Pavlovic, "Generative Adversarial Talking Head: Bringing Portraits to Life With a Weakly Supervised Neural Network", arXiv:1803.07716, 2018.

[9] K. Vougioukas, S. Petridis, M. Pantic, "Realistic Speech-Driven Facial Animation with GANs", International Journal of Computer Vision, Vol. 128, No. 5, 2020, pp. 1398-1413.

[10] J. Brockschmidt, J. Shang, J. Wu, "On the Generality of Facial Forgery Detection", Proceedings of the 16th IEEE International Conference on Mobile Ad Hoc and Sensor Systems Workshops, Monterey, CA, USA, 4-7 November 2019, pp. 43-47.

[11] U. A. Ciftci, I. Demir, L. Yin, "FakeCatcher: Detection Of Synthetic Portrait Videos Using Biological Signals", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020

[12] Y. Li and S. Lyu, "Exposing Deepfake Videos By Detecting Face Warping Artifacts", arXiv:1811.00656, 2018.

[13] P. Charitidis, G. Kordopatis-Zilos, S. Papadopoulos, I. Kompatsiaris, "Investigating The Impact of Pre-processing And Prediction Aggregation on the Deepfake Detection Task", arXiv:2006.07084, 2020.

[14] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning For Image Recognition", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27-30 June 2016, pp. 770-778.

[15] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet Classification With Deep Convolutional Neural Networks", Advances In Neural Information Processing Systems, Tahoe, NV, USA, 3-6 December 2012, pp. 1-9.

[16] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi,... C. M. Nguyen, "Deep Learning For Deepfakes Creation And Detection: A Survey", Computer Vision and Image Understanding, Vol. 223, 2022, pp. 103525,

[17] M. A. Wani, F. A. Bhat, S. Afzal, A. I. Khan, "Advances in Deep Learning", First Edition, Springer Publisher, 2020.

[18] H. Huang, P. S. Yu, C. Wang, "An Introduction to Image Synthesis With Generative Adversarial Nets", arXiv:1803.04469, 2018.

[19] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, H. Li, "Protecting World Leaders Against Deep Fakes", Proceedings of CVPR Workshops, Long Beach, CA, USA, 15-21 June 2019, p. 38.

[20] Y. Mirsky and W. Lee, "The Creation and Detection Of Deepfakes: A Survey", ACM Computing Surveys, Vol. 54, No. 1, 2021, pp. 1-41.

[21] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, "MesoNet: A Compact Facial Video Forgery Detection Network", Proceedings of the IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 11-13 December 2018, pp. 1-7.

[22] H. H. Nguyen, J. Yamagishi, I. Echizen, "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12-17 May 2019, pp. 2307-2311.

[23] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, F. Zhu, "Deepfakes Detection with Automatic Face Weighting", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14-19 June 2020. pp. 668-669.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, S. Gelly, "An Image is Worth 16x16 Words: Transformers For Image Recognition at Scale", arXiv:2010.11929, 2020.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, Y. Bengio, "Generative Adversarial Nets", Advances in Neural Information Processing Systems, Montreal, Canada, 8-11 December 2014, pp. 1-9.

[26] D. Wodajo, S. Atnafu, "Deepfake Video Detection Using Convolutional Vision Transformer", arXiv:2102.11126, 2021.