International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 6, Nov- Dec 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

# Deepfakes and Detection of Synthetic Media using Efficient Net & Vision Transformers

Mithilesh M Nimbalkar

KeraleeyaSamajam's Model College, Dombivli East, Mumbai, Maharashtra, India

mithilesh.nimbalkar2001@gmail.com

## 1.ABSTRACT

In today's world access to freely available public databases and with fast-growing progress of deep learning techniques in a Particular GAN (Generative Adversarial Networks) has led these today generations to create very realistic fake content with the help of apps.This survey provides a deep review of detection techniques for synthetic media (Deepfakes)

**2. KEYWORDS:**Deepfakes,Synthetic Media, Deep Learning, Deepfake Detection,Efficient Net.

## 3. INTRODUCTION

The term "deepfake" belongs to a technology called as"deep learning," a form of Artificial Intelligence. Which Create a realistic-looking fake media by swapping faces in videos and digital content using AI deep learning .There are several ways to create fake images and videos, but the one of the most common way is to use deep neural networks with autoencoders that use face swapping techniques. First you have a target video to base your deepfakes on. Next, you'll need more video clips of the people you want to target. Video may be completely irrelevant.For example, your target may be a scene from a bollywood movie, and the video of that person you want in your movie is a randomly downloaded clip from YouTube.The autoencoder is a deep gaining knowledge of AI software tasked with reading the video clips is an expansion of angles and environmental situations, and then mapping that man or woman onto some other form of machine learning is brought to the mix, known as Generative Adversarial and improves any flaws in the deepfake inside more than one rounds, making it harder for GANs also are used as a famous approach for creation of deepfakes,counting on the of statistics to "examine" a way to increase new examples that mimic the real thing, with ach DEEPFAKE generation. There are two principal generative strategies to get sensible faces.Generative hostile Networks (GHNs) and Variational AutoEncoders (VAEs).

## 4. DEEPFAKE GENERATION

There are two principal generative strategies to get sensible faces.Generative hostile Networks (GHNs) and Variational AutoEncoders (VAEs).GAN uses two different networks.A discriminator that needs to be able to tell the whether the video is fake

or not,and a generator (network) that actually modifies the video in a sufficiently reliable way to fool the opponent.Highly reliable and realistic results have been achieved with his GANs, and numerous approaches such

as tarGAN and DiscoGAN have been gradually introduced.The best results in this area were obtained with StyleGAN-V2.The VAE-based solution instead uses a system consisting of two encoder-decoder pairs trained to decompose and reconstruct one of the two faces of the to be exchanged.Then you can switch the decoding part and reconstruct the target person's face with this.The best known applications of this technique were DeepFaceLab.

## 5. DEEPFAKE DETECTION

The trouble of deepfake detection has a significant interest now not simplest in thevisible domain. for example, the latest paintings analyzes deepfakes in tweetsfor locating and defeating false content material in social networks.In an attempt to cope with the problem of deepfakes detection in movies, numerous datasets were produced over the years. those datasets are grouped into 3 generations, the primary era inclusive of DF-TIMIT , UADFC and FaceForensics++ , the second technology datasets which include Google Deepfake Detection Dataset , celeb-DF , and in the end the third generation datasets, with the DFDC dataset and DeepForensics . In addition to thegenerations go, the bigger those datasets are, and the more frames they contain.specifically, on the DFDC dataset, that is the biggest and most entire,a couple of experiments have been done seeking to attain an effective methodfor deepfake detection. superb consequences were received with EfficientNet B7 ensemble technique in . different noteworthy methods consist of those carried outin , who attempted to become aware of spatio-temporal anomalies by using combining an EfficientNet with a Gated Recurrent Unit (GRU). a few efforts to capture spatiotemporal inconsistencies were made in the usage of 3DCNN networks and in ,which supplied a way that exploits optical go with the flow to hit upon video glitches. somemoreclassical methods have additionally been proposed to perform deepfake detection. In precise, the authors in proposed a method primarily based on ok-nearest buddies,while the work in exploited SVMs. Of note is the very current work of Giudiceet al. in which they presented an innovative method for figuring out so-known asGAN unique Frequencies (GSF) that constitute a completely unique fingerprint of different generative architectures. by using exploiting the Discrete Cosine remodel (DCT) they manipulate to pick out anomalous frequencies.greater these days, strategies based on imaginative and prescient Transformers were proposed.significantly, the technique supplied in acquired good outcomes via mixing.In this example, the transformer clips are mixed with the clips extracted from the Efficient Net B7 preskilled via global pooling after which surpassed to the Transformer Encoder.

## 6. METHOD

The proposed methods analyze the faces extracted from the source video todetermine whenever they have been manipulated. For this reason, faces are preextracted using a state-of-the-art face detector, MTCNN. Wepropose twomixed convolutional-transformer architectures that take as input a pre-extractedface and output the probability that the face has been manipulated. The twopresented architectures aretrained in a supervised way to

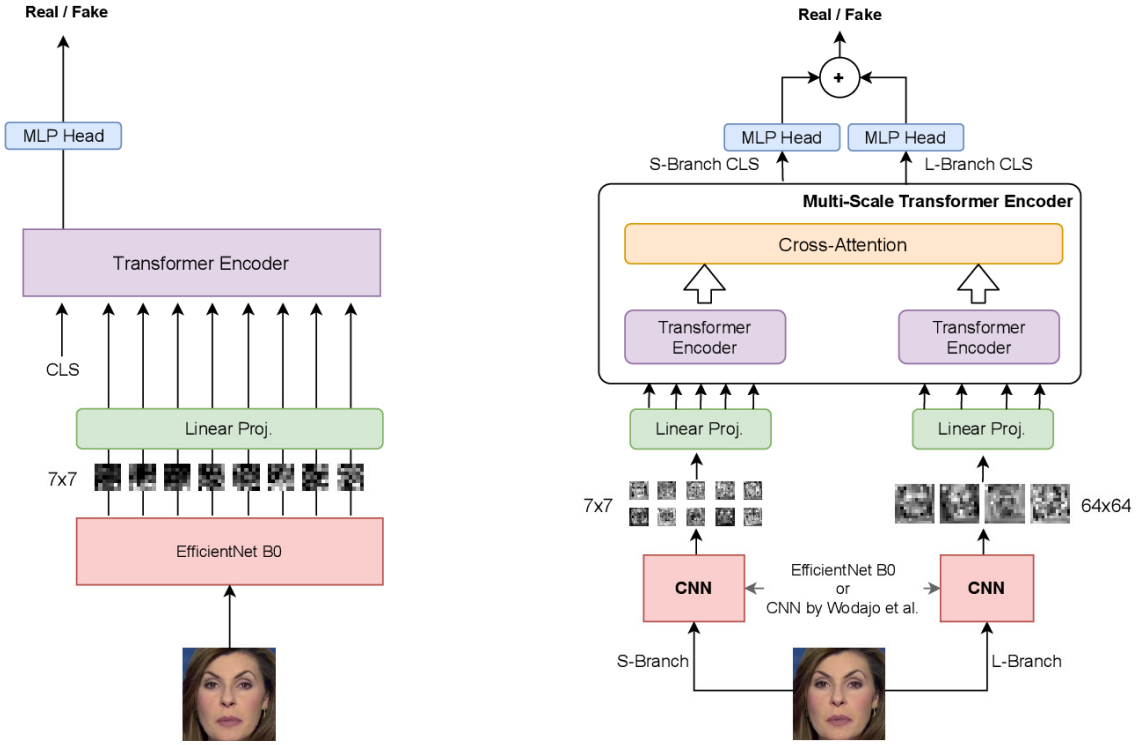
discern real fromfake examples. For this reason, we solve the detection task by framing it as abinary classification problem. Specifically, we propose the Efficient ViT and theConvolutional Cross ViT, better explained in the following paragraphs.The proposed models are trained on a face basis, and then they are used atinference time to draw a conclusion on the whole video shot by aggregating theinferred output both in time and across multiple faces.The Efficient ViT is composed of two blocks, a convolutionalnmodule for working as a feature extractor and Transformer Encoder, in a setupvery similar to the Vision Transformer (ViT). Considering the promisingresults of the EfficientNet, we use an EfficientNet B0,the smallest of the EfficientNet networks,as a convolutional extractor for processing the input faces.Specifically, the EfficientNet produces a visual feature for each chunk from theinput face. Each chunk is 7 × 7 pixels. After a linear projection, every featurefrom each spatial location is further processed by a Vision Transformer. The CLStoken is used
for producing the binary classification score. The architecture isillustrated in Figure 1a. The EfficientNet B0 feature extractor is initialized withthe pre-trained weights and fine-tuned to allow the last layers of the network toperform a more consistent and suitable extraction for this specific downstreamtask. The features extracted from the EfficientNet B0 convolutional network simplify the Transformer, because the CNN capabilities already embed vital low-level and localized in .The Convolutional go ViT proscribing the structure to the use most effective small patche. may not be the ideal desire,as artifacts delivered via deepfakes generation methods.because of this, we also introduce the Convolutional cross ViTarchitecture.The Convolutional go ViT builds upon both the green ViT and the multi-scale.extra in element, the Convolutional cross ViT makes use of two distinct branches: the S-branch.smallerpatches, and the L-branch, which fits on larger patches for having a wider.The visible tokens output by way of the Transformer Encoders from the 2 branches are comb.finally, the CLS tokens corresponding to the outputs from the 2 branches are use.these logits are summed, and a final sigmoid produces the final chances. A detailed overview of thisarchitecture is FortheConvolutional Cross ViT, we use two different CNN backbones. The former is the EfficientNet B0, which processes 7×7 image patches for the S-branch and 54×54 for the

L-branch. The latter is the CNN by Wodajoet al. , which handles 7 × 7 image patches for the S-branch and 64 × 64 for the L-branch.

## 7. EXPERIMENTS

We probed the presented architectures against some state-of-the-art methods on two widely-used datasets. In particular, we considered Convolutional ViT ,ViT with distillation , and Selim EfficientNet B7 , the winner of the Deep.




(a) Efficient ViT architecture.

(b) Convolutional Cross ViT architecture.

**Fig. 1:** The proposed architectures. Notice that for the Convolutional Cross ViT in (b), we experimented both with EfficientNet B0 and with the convolutional architecture by as feature extractors.

Fake Detection Challenge (DFDC). Note that the results for Convolutional ViT are not reported in the original paper, but are obtained by running the test code on the DFDC test suite using an available pre-trained model released by the authors.

## 8. DATASETS AND FACE EXTRACTION

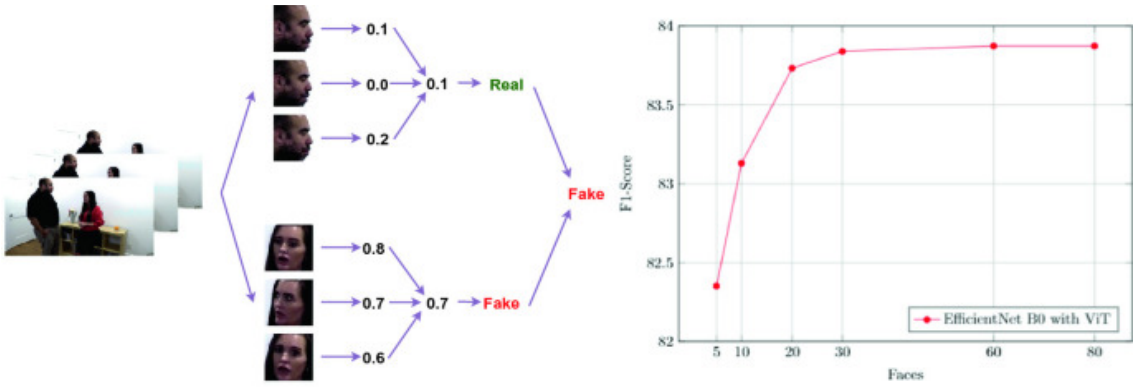
First, we ran some tests on FaceForensics++. The dataset is composed
of original and fake videos generated through different deepfake generations
techniques. For the evaluation, we took into account the videos generated in Deepfakes,Face2Face, FaceShifter, FaceSwap and NeuralTextures sub-datasets. We also used the DFDC test set containing 5000 videos. The model trained on the entire training set, which contains mock videos of all methods considered FaceForensics++ and DFDC dataset training videos were used for the calculation model accuracy rates

In order to compareour methods also on the DFDC test set, we tested Convolutional Vision
Transformer obtains the necessary AUC and F1-score on these videos
values for comparison.During training, we used MTCNN to extract faces from videos,and we have done data expansion as in Unlike them we he drew the faces so that they were always square and without padding. The acquired images are used during training, so the remaining part is ignored frames. We used the Albumutations library and used common transformations such as introducing blur, Gaussian noise, transposition, rotation and various isotropic changes in size during training.

## 9. INFERENCE

At derivation time, we set the true/false threshold to 0.55 as reported in.However,
instead of averaging them all, we proposed a slightly more sophisticated voting procedure rating on individual faces indistinct in the video. Specifically, we connected scores, grouping is by actor identifier. Face ID is available as output from the MTCNN face detector used. Score of the different actors are averaged over time to produce the probability of a face being present false. After that, the scores of the individual actors are combined using a hard vote. Especially if there is at least one actor's face that crosses the threshold, the entire video is classified as fake. The procedure is graphically explained seems statistically unnecessary inference time.CombiningEfficientNet and ViTs for Video Deepfake Detection.
in We claim that this .This approach is useful for better processing of videos in which there is only one actor's face was manipulated.
Furthermore, it is interesting to evaluate how the performance changes when a
different number of faces are taken into account at the time of derivation. To ensure that the tests are as light as possible and at the same time effective, we experimented on one of our networks see how the F1 score changes with the number of faces considered at the time of testing We have noticed that a plateau is reached when there are no more than 30 faces It is used, so using more than this number of faces




(a) Inference strategy with multiple faces in the same video.

(b) F1-score versus the number of extracted faces.

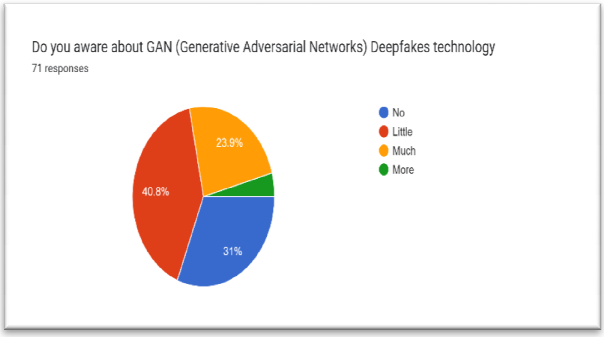
Table 1: Results on DFDC test dataset

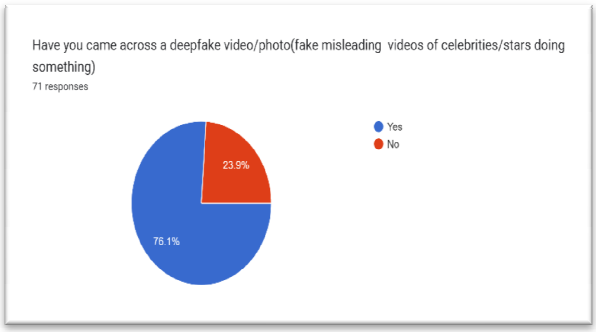
| Model | AUC | F1-score |
|---|---|---|
| ViT with distillation [Heo et al., 2021] | 0.978 | 91.9% |
| Selim EfficientNet B7 [Seferbekov, 2020] | 0.972 | 90.6% |
| Convolutional ViT | 0.843 | 77.0% |
| Efficient ViT (our) | 0.919 | 83.8% |
| Convolutional Cross ViT (our) | 0.925 | 84.5% |
| Efficient Cross ViT (our) | 0.951 | 88.0% |

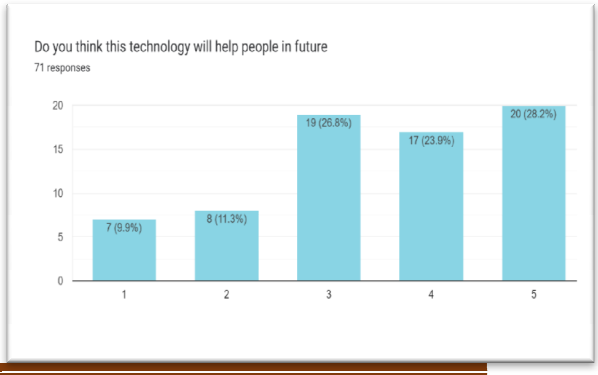
Table 2: Models accuracy on FaceForencics++

| Model | Mean | FaceSwap | DeepFakes | FaceShifter | NeuralTextures |
|---|---|---|---|---|---|
| Convolutional ViT [Wodajo and Atnafu, 2021] | 67% | 69% | **93%** | 46% | 60% |
| Efficient ViT (our) | 76% | 78% | 83% | 76% | 68% |
| Convolutional Cross ViT (our) | 76% | 81% | 83% | 73% | 67% |
| Efficient Cross ViT (our) | **80%** | **84%** | 87% | **80%** | **69%** |

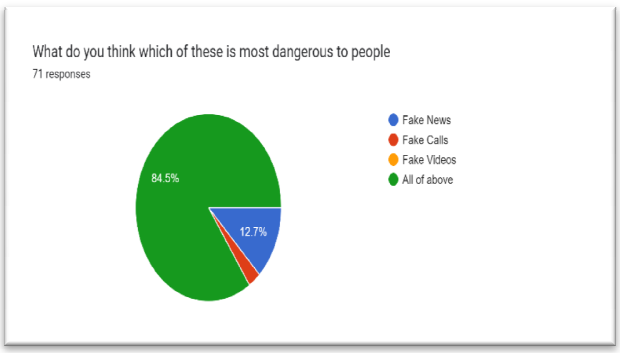
***FIGURES AND SURVEY RESULTS**

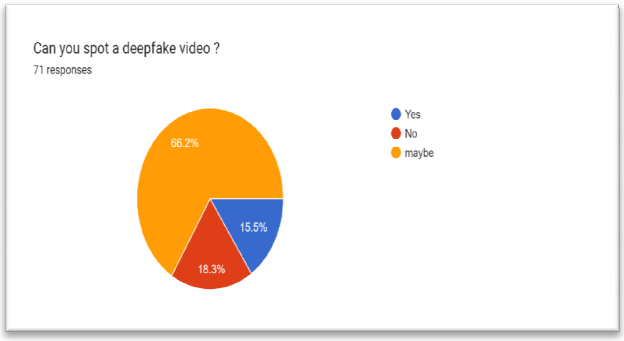


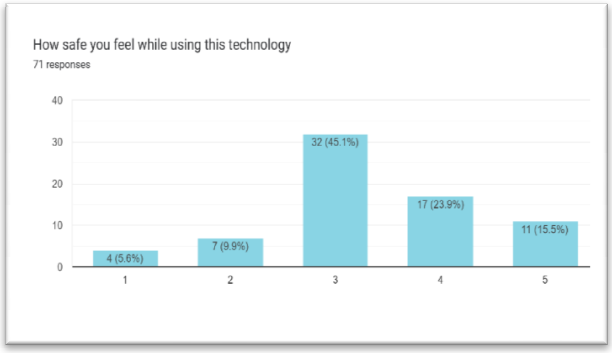



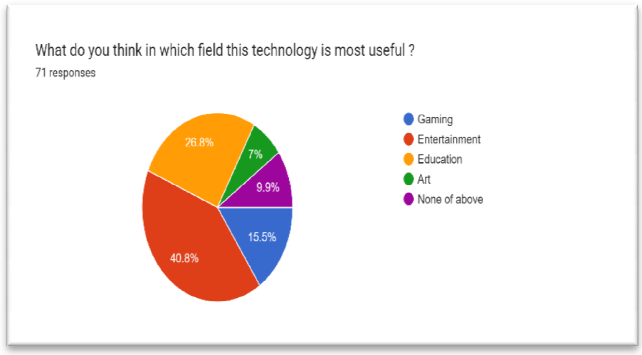






## DESCRIPTIVE STATISTICS

| Do you aware about GAN (Generative Adversarial Networks) Deepfakes technology | |
|---|---|
| | |
| Mean | 2.03030303 |
| Standard Error | 0.101204165 |
| Median | 2 |
| Mode | 2 |
| Standard Deviation | 0.822186521 |
| Sample Variance | 0.675990676 |
| Kurtosis | -0.686512952 |
| Skewness | 0.285561094 |
| Range | 3 |
| Minimum | 1 |
| Maximum | 4 |
| Sum | 134 |
| Count | 66 |

| Have you came across a deepfake video/photo(fake misleading videos of celebrities/stars doing something) | |
|---|---|
| Mean | 1.227272727 |
| Standard Error | 0.051979261 |
| Median | 1 |
| Mode | 1 |
| Standard Deviation | 0.422281515 |
| Sample Variance | 0.178321678 |
| Kurtosis | -0.233660131 |
| Skewness | 1.33204978 |
| Range | 1 |
| Minimum | 1 |
| Maximum | 2 |
| Sum | 81 |
| Count | 66 |

| Do you think this technology will help people in future | |
|---|---|
| Mean | 3.454545455 |
| Standard Error | 0.157459164 |
| Median | 4 |
| Mode | 3 |
| Standard Deviation | 1.279204298 |
| Sample Variance | 1.636363636 |
| Kurtosis | -0.719697124 |
| Skewness | -0.464155726 |
| Range | 4 |
| Minimum | 1 |
| Maximum | 5 |
| Sum | 228 |
| Count | 66 |

| What do you think which of these is most dangerous to people | |
|---|---|
| Mean | 3.575757576 |
| Standard Error | 0.126012184 |
| Median | 4 |
| Mode | 4 |
| Standard Deviation | 1.023727819 |
| Sample Variance | 1.048018648 |
| Kurtosis | 2.483185182 |
| Skewness | -2.074117564 |
| Range | 3 |
| Minimum | 1 |
| Maximum | 4 |
| Sum | 236 |
| Count | 66 |

| Can you spot a deep fake video | |
|---|---|
| Mean | 2.515151515 |
| Standard Error | 0.092220533 |
| Median | 3 |
| Mode | 3 |
| Standard Deviation | 0.749203151 |
| Sample Variance | 0.561305361 |
| Kurtosis | -0.145777091 |
| Skewness | -1.184952216 |
| Range | 2 |
| Minimum | 1 |
| Maximum | 3 |
| Sum | 166 |
| Count | 66 |

## CONCLUSION

In this research, we have demonstrated the effectiveness of mixed convolutional transformer networks in the task of Deepfake detection. especially, we used pre-trained convolutional networks which includes the widely used EfficientNet B0 to extract visible capabilities and we relied on Vision Transformers to retrieve information global description for the subsequent task. We have shown that it is possible to obtain state-of-the-art results without the need for distillation techniques from models based on convolutional or ensemble networks. Using a patch the EfficientNet-based extractor proved to be particularly efficient even when simple using the smallest network in this category. EfficientNet also led to better results than the generic convolutional network

trained from scratch used in Wodajo et al .We then proposed a mixed architecture, Convolutional Cross ViT, that works at two different scales to capture local and global details. Tests passeddemonstrated the importance of multilevel analysis using these models determination

of image manipulation. We also paid special attention to the inference phase. Especially us introduced a simple but effective voting scheme for explicitly resolving multiples faces in the video. Scores from multiple actor faces are first averaged over time, and only then a hard vote is used to decide if at least one face has been tampered with. This inference mechanism produced slightly better and more stable results than global average pooling of scores made by previous methods.

## Bibliography

1. Amato, G., Bolettieri, P., Falchi, F., Gennaro, C., Messina, N., Vadicamo, L., Vairo, C.: Visione at video browser showdown 2021. In: International Conference on Multimedia Modeling. pp. 473–478. Springer (2021) 2. Amato, G., Ciampi, L., Falchi, F., Gennaro, C., Messina, N.: Learning pedestrian detection from virtual worlds. In: International Conference on Image Analysis and Processing. pp. 302–312. Springer (2019) 3. Amerini, I., Galteri, L., Caldelli, R., Del Bimbo, A.: Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 4. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018) 5. Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. arXiv preprint arXiv:2103.14899 (2021) 6. Chesney, B., Citron, D.: Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 107, 1753 (2019) 7. Choi, Yunjey, e.a.: Stargan: Unified generative adversarial networks for multi-domain image- to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018) 8. Ciampi, L., Messina, N., Falchi, F., Gennaro, C., Amato, G.: Virtual to real adaptation of pedestrian detectors. Sensors 20(18), 5250 (2020) 9. Dolhansky, B.,

Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397

2. 10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 11. Dufour, N., Gully, A.: Contributing data to deep-fake detection research (2019), https://ai.googleblog.com/2019/09/ contributing-data-to-deepfake-detection.html 12. Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: Tweepfake: About detecting deepfake tweets. Plos one 16(5), e0251415 (2021) 13. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020) 14. Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting gandct anomalies. arXivpreprint arXiv:2101.09781 (2021) 15. Goodfellow,I.J.,Ozair, S., Courville, A., Bengio, Y.: . In: Advances in neural information processing systems 27 (2014) 16. Guarnera, L., Giudice, O., Battiato, S.: Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2020) 17. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., et al.: A survey on visual transformer. arXiv preprint arXiv:2012.12556 (2020) 18. Heo, Y.J., Choi, Y.J., Lee, Y.W., Kim, B.G.: Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353 (2021) 19. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2889–2898 (2020) 20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020) 21. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. arXiv preprint arXiv:2101.01169 (2021) 22. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. pp. 1857–1865. PMLR (2017) 23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 24. Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)

25. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020) 26. de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A.: Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749 (2020) 27. MacAvaney, S., Nardini, F.M., Perego, R., Tonellotto, N., Goharian, N., Frieder, O.: Efficient document re-ranking for transformers by precomputing term representations. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 49–58

3. [7] Shehzeen Hussain, PaarthNeekhara, Malhar Jere, FarinazKoushanfar, Julian McAuley “Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples” arXiv, November 2020
4. [8] Yuezun Li, SiweiLyu “Exposing DeepFake Videos By Detecting Face Warping Artifacts” arXiv, May 2019.