Combining EfficientNet and Vision Transformers for Video Deepfake Detection

Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi

Institute of Information Science and Technologies (ISTI), Italian National Research Council (CNR), Via G. Moruzzi 1, 56124 Pisa, Italy
`name.surname@isti.cnr.it`

Abstract. Deepfakes are the result of digital manipulation to forge realistic yet fake imagery. With the astonishing advances in deep generative models, fake images or videos are nowadays obtained using variational autoencoders (VAEs) or Generative Adversarial Networks (GANs). These technologies are becoming more accessible and accurate, resulting in fake videos that are very difficult to be detected. Traditionally, Convolutional Neural Networks (CNNs) have been used to perform video deepfake detection, with the best results obtained using methods based on EfficientNet B7. In this study, we focus on video deep fake detection on faces, given that most methods are becoming extremely accurate in the generation of realistic human faces. Specifically, we combine various types of Vision Transformers with a convolutional EfficientNet B0 used as a feature extractor, obtaining comparable results with some very recent methods that use Vision Transformers. Differently from the state-of-the-art approaches, we use neither distillation nor ensemble methods. Furthermore, we present a straightforward inference procedure based on a simple voting scheme for handling multiple faces in the same video shot. The best model achieved an AUC of 0.951 and an F1 score of 88.0%, very close to the state-of-the-art on the DeepFake Detection Challenge (DFDC). The code for reproducing our results is publicly available here: `https://tinyurl.com/cnn-vit-dfd`.

Keywords: Deep Fake Detection · Transformer Networks · Deep Learning

1 Introduction

With the recent advances in generative deep learning techniques, it is nowadays possible to forge highly-realistic and credible misleading videos. These methods have generated numerous fake news or revenge porn videos, becoming a severe problem in modern society [5]. These fake videos are known as *deepfakes*. Given the astonishing realism obtained by recent models in the generation of human faces, deepfakes are mainly obtained by transposing one person's face onto another's. The results are so realistic that it is almost like the person being replaced is actually present in the video, and the replaced actors are rigged to say things they never actually said [35].

The evolution of deepfakes generation techniques and their increasing accessibility forces the research community to find effective methods to distinguish a manipulated video from a real one. Nowadays, models based on Transformer architecture are gaining ground in the field of Computer Vision, showing excellent results in image processing [19], document retrieval [25], and efficient visual-textual matching [28,29]. Unlike Vision Transformers, CNNs still maintain an important architectural prior, the spatial locality, which is very important for discovering image patch abnormalities and maintaining good data efficiency. CNNs, in fact, have a long-established success on many tasks, ranging from image classification [12,37] and object detection [32,1,7] to abstract visual reasoning [26,27].

In this paper, we use the power of convolutional and transformer models to tackle the problem of video deepfake detection. Specifically, we analyze different solutions based on the combination of convolutional networks — particularly the EfficientNet B0 — with different types of Vision Transformers [9]. We compare the results with the current state-of-the-art, keeping into consideration both accuracy and network complexity. Our proposed models are frame-based, as many others in literature. Nevertheless, we also propose a method to handle multiple sequential frames at inference time. Specifically, we propose a simple yet effective voting mechanism that handles multiple face instances across multiple frames to judge the genuineness of the video shot. We show that this methodology could lead to better and more stable results.

2 Related Works

2.1 Deepfake Generation

There are mainly two generative approaches to obtain realistic faces: Generative Adversarial Networks (GANs) [14] and Variational AutoEncoders (VAEs) [21].

GANs employ two distinct networks. The discriminator, the one that must be able to identify when a video is fake or not, and the generator, the network that actually modifies the video in a sufficiently credible way to deceive its counterpart. With GANs, very credible and realistic results have been obtained, and over time, numerous approaches have been introduced such as StarGAN [6] and DiscoGAN [20]; the best results in this field have been obtained with StyleGAN-V2 [18].

VAE-based solutions, instead, make use of a system consisting of two encoder-decoder pairs, each of which is trained to deconstruct and reconstruct one of the two faces to be exchanged. Subsequently, the decoding part is switched, and this allows the reconstruction of the target person's face. The best-known uses of this technique were DeepFaceLab [31], DFaker[1], and DeepFaketf[2].

[1] https://github.com/dfaker/df

[2] https://github.com/StromWine/DeepFake_tf

2.2 Deepfake Detection

The problem of deepfake detection has a widespread interest not only in the visual domain. For example, the recent work in [11] analyzes deepfakes in tweets for finding and defeating false content in social networks.

In an attempt to address the problem of deepfakes detection in videos, numerous datasets have been produced over the years. These datasets are grouped into three generations, the first generation consisting of DF-TIMIT [22], UADFC [38] and FaceForensics++ [33], the second generation datasets such as Google Deepfake Detection Dataset [10], Celeb-DF [23], and finally the third generation datasets, with the DFDC dataset [8] and DeepForensics [17]. The further the generations go, the larger these datasets are, and the more frames they contain.

In particular, on the DFDC dataset, which is the largest and most complete, multiple experiments were carried out trying to obtain an effective method for deepfake detection. Very good results were obtained with EfficientNet B7 ensemble technique in [34]. Other noteworthy methods include those conducted in [30], who attempted to identify spatio-temporal anomalies by combining an EfficientNet with a Gated Recurrent Unit (GRU). Some efforts to capture spatio-temporal inconsistencies were made in [24] using 3DCNN networks and in [2], which presented a method that exploits optical flow to detect video glitches. Some more classical methods have also been proposed to perform deepfake detection. In particular, the authors in [15] proposed a method based on K-nearest neighbors, while the work in [38] exploited SVMs. Of note is the very recent work of Giudice et al. [13] in which they presented an innovative method for identifying so-called GAN Specific Frequencies (GSF) that represent a unique fingerprint of different generative architectures. By exploiting the Discrete Cosine Transform (DCT) they manage to identify anomalous frequencies.

More recently, methods based on Vision Transformers have been proposed. Notably, the method presented in [36] obtained good results by combining Transformers with a convolutional network, used to extract patches from faces detected in videos.

State of the art was then recently improved by performing distillation from the EfficientNet B7 pre-trained on the DFDC dataset to a Vision Transformer [16]. In this case, the Vision Transformer patches are combined with patches extracted from the EfficientNet B7 pre-trained via global pooling and then passed to the Transformer Encoder. A distillation token is then added to the Transformer network to transfer the knowledge acquired by the EfficientNet B7.

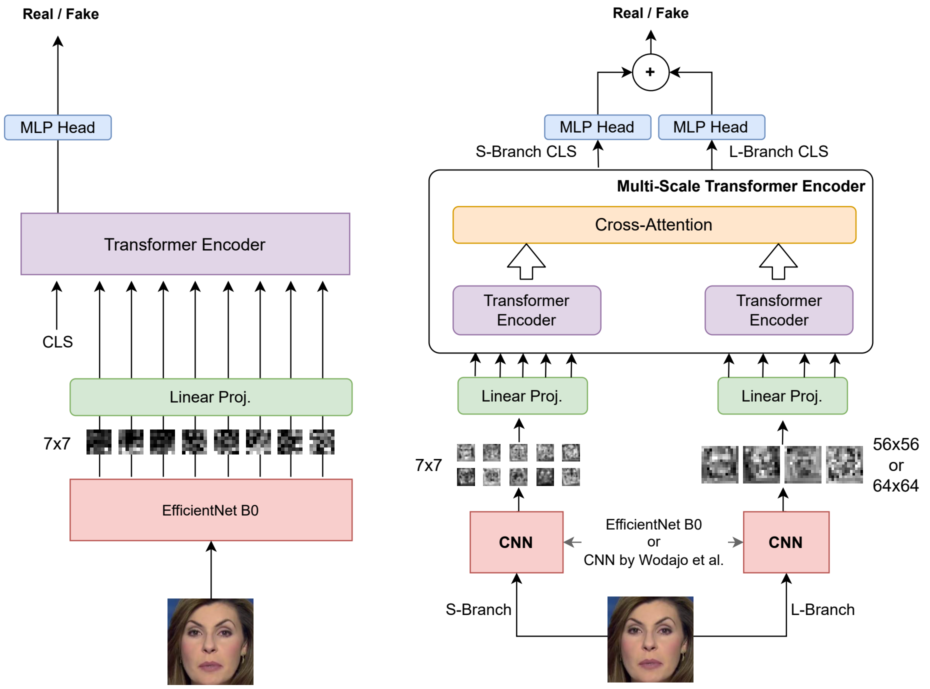
3 Method

The proposed methods analyze the faces extracted from the source video to determine whenever they have been manipulated. For this reason, faces are pre-extracted using a state-of-the-art face detector, MTCNN [39]. We propose two mixed convolutional-transformer architectures that take as input a pre-extracted face and output the probability that the face has been manipulated. The two presented architectures are trained in a supervised way to discern real from fake examples. For this reason, we solve the detection task by framing it as a binary classification problem. Specifically, we propose the *Efficient ViT* and the *Convolutional Cross ViT*, better explained in the following paragraphs.

The proposed models are trained on a face basis, and then they are used at inference time to draw a conclusion on the whole video shot by aggregating the inferred output both in time and across multiple faces, as explained in Section 4.3.

The Efficient ViT The Efficient ViT is composed of two blocks, a convolutional module for working as a feature extractor and a Transformer Encoder, in a setup very similar to the Vision Transformer (ViT) [9]. Considering the promising results 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 the input face. Each chunk is 7 × 7 pixels. After a linear projection, every feature from each spatial location is further processed by a Vision Transformer. The CLS token is used for producing the binary classification score. The architecture is illustrated in Figure 1a. The EfficientNet B0 feature extractor is initialized with the pre-trained weights and fine-tuned to allow the last layers of the network to perform a more consistent and suitable extraction for this specific downstream task. The features extracted from the EfficientNet B0 convolutional network simplify the training of the Vision Transformer, as the CNN features already embed important low-level and localized information from the image.

The Convolutional Cross ViT Limiting the architecture to the use only small patches as in the Efficient ViT may not be the ideal choice, as artifacts introduced by deepfakes generation methods may arise both locally and globally. For this reason, we also introduce the Convolutional Cross ViT architecture. The Convolutional Cross ViT builds upon both the Efficient ViT and the multi-scale Transformer architecture by [4]. More in detail, the Convolutional Cross ViT uses two distinct branches: the *S-branch*, which deals with smaller patches, and the *L-branch*, which works on larger patches for having a wider receptive field. The visual tokens output by the Transformer Encoders from the two branches are combined through cross attention, allowing direct interaction between the two paths. Finally, the CLS tokens corresponding to the outputs from the two branches are used to produce two separate logits. These logits are summed, and a final sigmoid produces the final probabilities. A detailed overview of this architecture is shown in Fig. 1b. For the Convolutional 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 56 × 56 for the L-branch. The latter is the CNN by Wodajo et al. [36], which handles 7 × 7 image patches for the S-branch and 64 × 64 for the L-branch.



(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 [36] as feature extractors.

4 Experiments

We probed the presented architectures against some state-of-the-art methods on two widely-used datasets. In particular, we considered Convolutional ViT [36], ViT with distillation [16], and Selim EfficientNet B7 [34], the winner of the Deep Fake Detection Challenge (DFDC). Notice that the results for Convolutional ViT [36] are not reported in the original paper, but they are obtained executing the test code on DFDC test set using the available pre-trained model released by the authors.

4.1 Datasets and Face Extraction

We initially conducted some tests on FaceForensics++. The dataset is composed of original and fake videos generated through different deepfake generation techniques. For evaluating, we considered the videos generated in the 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 includes fake videos of all considered methods of FaceForensics++ and the training videos of DFDC dataset, was used to calculate the accuracy measures of the model, reported separately. In order to compare our methods also on the DFDC test set, we tested the Convolutional Vision Transformer [36] on these videos obtaining the necessary AUC and F1-score values for comparison.

During training, we extracted the faces from the videos using an MTCNN [39], and we performed data augmentation like in [34]. Differently from them, we extracted the faces so that they were always squared and without padding. We used the Albumentations library [3], and we applied common transformations such as the introduction of blur, Gaussian noise, transposition, rotation, and various isotropic resizes during training.

4.2 Training

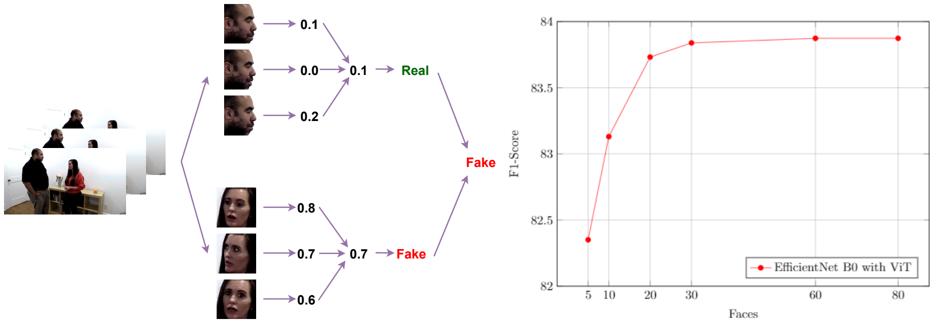
We trained the networks on 220,444 faces extracted from the videos of DFDC training set and FaceForensics++ training videos, and we used 8070 faces for validation from DFDC dataset. The training set was constructed trying to maintain a good balance between the real class composed of 116,950 images and fakes with 103,494 images.

We used pre-trained EfficientNet B0 and Wodajo CNN feature extractors. However, we observed better results when fine-tuning them, so we did not freeze the extraction layers. We used the standard binary cross-entropy loss as our objective during training. We optimized our network end-to-end, using an SGD optimizer with a learning rate of 0.01.

4.3 Inference

At inference time, we set a real/fake threshold at 0.55 as done in [16]. However, we proposed a slightly more elaborated voting procedure instead of averaging all ratings on individual faces indistinctly within the video. Specifically, we merged the scores, grouping them by the identifier of the actors. The face identifier is available as an output from the employed MTCNN face detector. The scores from different actors are averaged over time to produce a probability of the face being fake. Then, the per-actor scores are merged using hard voting. In particular, if there is at least one actor face passing the threshold, the whole video is classified as fake. The procedure is graphically explained in Fig. 2a. We claim that this approach is helpful to handle videos in which only one of the actors' faces has been manipulated.

In addition, it is interesting to evaluate how the performance changes when a varying number of faces are considered at inference time. To ensure that the tests are as light yet effective as possible, we experimented on one of our networks to see how the F1-score varies with the number of faces considered at testing time (Fig. 2b). We noticed that a plateau is reached when no more than 30 faces are used, so employing more than this number of faces seems statistically useless at inference time.



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

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

Fig. 2: Inference.

Table 1: Results on DFDC test dataset

Model	AUC	F1-score	# params
ViT with distillation [16]	0.978	91.9%	373M
Selim EfficientNet B7 [34]†	0.972	90.6%	462M
Convolutional ViT [36]	0.843	77.0%	89M
Efficient ViT (our)	0.919	83.8%	109M
Conv. Cross ViT Wodajo CNN (our)	0.925	84.5%	142M
Conv. Cross ViT Eff.Net B0 - Avg (our)	0.947	85.6%	101M
Conv. Cross ViT Eff.Net B0 - Voting (our)	0.951	88.0%	101M

† Uses an ensemble of 6 networks.

4.4 Results

Table 1 shows that all models developed with EfficientNet achieve considerably higher AUC and F1-scores than the Convolutional ViT presented in [36], providing initial evidence that this specific network structure may be more suitable for this type of task. It can also be noticed that the models based on Cross Vision Transformer obtain the best results, confirming the theory that joined local and global image processing brings to better anomaly detection.

The models with Cross Vision Transformer show a particularly marked improvement when using the EfficientNet B0 as a patch extractor. Although the AUC and F1-score remain slightly below other state-of-the-art methods (in the first two rows of Table 1), these results were obtained using neither distillation nor ensemble techniques that complicate both training and inference. In fact, we can notice how the Cross Vision Transformer with the EfficientNet extractor can reach a competitive performance using less than 1/3 of the parameters of the top methods.

Furthermore, in the last two rows of Table 1 we can notice how our voting procedure used at inference time can slightly improve the results with respect to

Table 2: Models accuracy on FaceForensics++

Model	Mean	FaceSwap	DeepFakes	FaceShifter	NeuralTextures
Convolutional ViT [36]	67%	69%	**93%**	46%	60%
Efficient ViT (our)	76%	78%	83%	76%	68%
Conv. Cross ViT Wodajo CNN (our)	76%	81%	83%	73%	67%
Conv. Cross ViT EfficientNet B0 (our)	**80%**	**84%**	87%	**80%**	**69%**

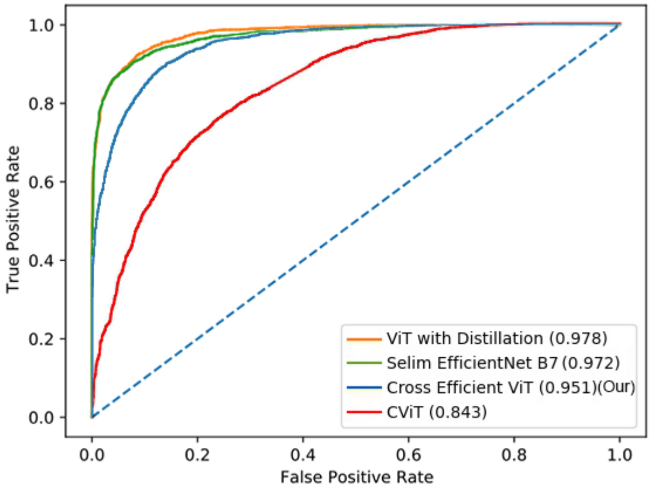


Fig. 3: ROC Curves comparison between our best model and others on DFDC test set.

a plain average of the scores from all the faces indistinctly, as done by the other methods. In Fig. 3 we report a detailed ROC plot for the architectures on the DFDC dataset.

In order to compare the developed models also on another dataset, we carried out some tests also on FaceForensics++. As shown in Table 2, our models outperform the original Convolutional ViT [36] on all sub-datasets of FaceForensics++, excluding DeepFakes. This is probably because the network could better generalize on very specific types of deepfakes. It is worth noting how the results obtained in terms of accuracy on the various sub-datasets confirm the assumption already made in [36]: some deepfakes techniques such as NeuralTextures produce videos that are more difficult to find, thus resulting in lower accuracy values than other sub-datasets. However, the average of all our three models is higher than the average obtained by the Convolutional ViT. The Convolutional Cross ViT achieves the best result with the EfficientNet B0 backbone, obtaining a mean accuracy of 80%.

5 Conclusions

In this research, we demonstrated the effectiveness of mixed convolutional-transformer networks in the Deepfake detection task. Specifically, we used pre-trained convolutional networks, such as the widely used EfficientNet B0, to extract visual features, and we relied on Vision Transformers to obtain an informative global description for the downstream task. We showed that it is possible to obtain remarkable results, very close to the state-of-the-art, without necessarily resorting to distillation techniques or ensemble networks. The use of a patch extractor based on EfficientNet proved to be particularly effective even by simply 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 [36]. We then proposed a mixed architecture, the Convolutional Cross ViT, that works at two different scales to capture local and global details. The tests carried out with these models demonstrated the importance of multi-scale analysis for determining the manipulation of an image.

We also paid particular attention to the inference phase. In particular, we presented a simple yet effective voting scheme for explicitly dealing with multiple faces in a video. The scores from multiple actor faces are first averaged over time, and only then hard voting is used to decide if at least one face was manipulated. This inference mechanism yielded slightly better and stable results than the global average pooling of the scores performed by previous methods.

References

1. 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)
2. 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)
3. Buslaev, A., Parinov, A., Khvedchenya, E., Iglovikov, V.I., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. ArXiv e-prints (2018)
4. 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. pp. 357–366 (2021)
5. Chesney, B., Citron, D.: Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. **107**, 1753 (2019)
6. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018). https://doi.org/10.1109/CVPR.2018.00916
7. Ciampi, L., Messina, N., Falchi, F., Gennaro, C., Amato, G.: Virtual to real adaptation of pedestrian detectors. Sensors **20**(18), 5250 (2020)
8. 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 (2020)

9. 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)
10. Dufour, N., Gully, A.: Contributing data to deep-fake detection research (2019), `https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html`
11. Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: Tweepfake: About detecting deepfake tweets. Plos one **16**(5), e0251415 (2021)
12. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 (2021)
13. Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting gan dct anomalies. Journal of Imaging **7**(8), 128 (2021)
14. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. In: Advances in neural information processing systems 27 (2014)
15. 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)
16. 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)
17. 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)
18. 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)
19. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (CSUR) (2021)
20. 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)
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014)
22. Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
23. 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)
24. de Lima, O., Franklin, S., Basu, S., Karwoski, B., George, A.: Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749 (2020)
25. 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 (2020)
26. Messina, N., Amato, G., Carrara, F., Falchi, F., Gennaro, C.: Testing deep neural networks on the same-different task. In: 2019 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–6. IEEE (2019)

27. Messina, N., Amato, G., Carrara, F., Gennaro, C., Falchi, F.: Solving the same-different task with convolutional neural networks. Pattern Recognition Letters **143**, 75–80 (2021)
28. Messina, N., Amato, G., Esuli, A., Falchi, F., Gennaro, C., Marchand-Maillet, S.: Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **17**(4), 1–23 (2021)
29. Messina, N., Falchi, F., Esuli, A., Amato, G.: Transformer reasoning network for image-text matching and retrieval. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5222–5229. IEEE (2021)
30. Montserrat, D.M., Hao, H., Yarlagadda, S.K., Baireddy, S., Shao, R., Horváth, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., et al.: Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 668–669 (2020)
31. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J., et al.: Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535 (2020)
32. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
33. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
34. Seferbekov, S.: Dfdc 1st place solution (2020), "https://github.com/selimsef/dfdc_deepfake_challenge"
35. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion **64**, 131–148 (2020)
36. Wodajo, D., Atnafu, S.: Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126 (2021)
37. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
38. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265. IEEE (2019)
39. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)