Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection

Davide Alessandro Coccomini
davidealessandro.coccomini@isti.cnr.it
Inst. of Information Science and
Technologies "A. Faedo"
Consiglio Nazionale delle Ricerche
Pisa, Italy

Roberto Caldelli*
roberto.caldelli@unifi.it
National Inter-University Consortium
for Telecommunications (CNIT)
Florence, Italy

Fabrizio Falchi
fabrizio.falchi@isti.cnr.it
Inst. of Information Science and
Technologies "A. Faedo"
Consiglio Nazionale delle Ricerche
Pisa, Italy

Claudio Gennaro
claudio.gennaro@isti.cnr.it
Inst. of Information Science and
Technologies "A. Faedo"
Consiglio Nazionale delle Ricerche
Pisa, Italy

Giuseppe Amato
giuseppe.amato@isti.cnr.it
Inst. of Information Science and
Technologies "A. Faedo"
Consiglio Nazionale delle Ricerche
Pisa, Italy

ABSTRACT

Deepfake Generation Techniques are evolving at a rapid pace, making it possible to create realistic manipulated images and videos and endangering the serenity of modern society. The continual emergence of new and varied techniques brings with it a further problem to be faced, namely the ability of deepfake detection models to update themselves promptly in order to be able to identify manipulations carried out using even the most recent methods. This is an extremely complex problem to solve, as training a model requires large amounts of data, which are difficult to obtain if the deepfake generation method is too recent. Moreover, continuously retraining a network would be unfeasible. In this paper, we ask ourselves if, among the various deep learning techniques, there is one that is able to generalise the concept of deepfake to such an extent that it does not remain tied to one or more specific deepfake generation methods used in the training set. We compared a Vision Transformer with an EfficientNetV2 on a cross-forgery context based on the ForgeryNet dataset. From our experiments, It emerges that EfficientNetV2 has a greater tendency to specialize often obtaining better results on training methods while Vision Transformers exhibit a superior generalization ability that makes them more competent even on images generated with new methodologies.

CCS CONCEPTS

• **Applied computing** → **Computer forensics**; • **Computing methodologies** → *Computer vision.*

*Also with Universitas Mercatorum, Rome, Italy.

KEYWORDS

Deep Fake Detection, Transformer Networks, Deep Learning

1 INTRODUCTION

The advancement of modern Deep Learning techniques is allowing society to evolve in many ways, helping in the achievement of increasingly stunning results in virtually every field. However, this progress also hides pitfalls with possible uses of Deep Learning that can be detrimental to the well-being of people. One of the most worrying emerging phenomena is undoubtedly that of deep fakes. These are images or videos manipulated by means of advanced Deep Learning techniques, to make the subjects filmed say or do things that they would never have said or done. The result of these techniques can then be used to destroy someone's reputation or to provoke conflict or manipulate the reality of events to one's advantage.

Distinguishing an image or video manipulated by these techniques from a genuine one has therefore become the goal of many researchers who have developed innovative methodologies, often also based on deep learning, to carry out what is called deepfake detection. In general, these techniques try to find any artefacts or anomalies that may be introduced during the manipulation process.

Machine Learning algorithms, however, need data to be trained, often in large quantities, and a number of datasets have sprung up trying to be as complete as possible in representing the various results that can be obtained with deepfake generation systems. In fact, there are numerous and varied techniques to manipulate multimedia content and ideally we would like to obtain a deepfake detector capable of identifying them regardless of the technique used for manipulation. Even more so, it would be ideal to have a

system capable of identifying deepfakes generated by novel methods, examples of which are not present in the training dataset. In other words, we would like a deepfake detector capable of learning the general concept of deepfake and not simply being trained to recognise specific anomalies introduced by one or more specific deepfake generation methodologies. In this research, we therefore attempted to find out which of the main Deep Learning techniques was most capable of generalising the concept of deepfake and therefore proved robust in identifying images manipulated by methods it had never been trained on. The comparison was made between the two categories of neural networks used in this field, Vision Transformer and Convolutional Neural Networks. Our experiments showed that the former were less inclined to specialise in a specific method and were able to obtain consistent results even on deepfakes generated with novel methodologies. On the other hand, the Convolutional Neural Networks appear able to reach better performances in terms of accuracy on the training methods.

2 RELATED WORKS

2.1 Deepfake Generation

Deepfake Generation techniques are the set of methods used to manipulate a human face in order to make it appear different or to replace its identity in a realistic manner. There are two main categories of approaches, those based on Variational AutoEncoders (VAEs) [21] and those based on Generative Adversarial Networks (GANs) [13]. Methods based on VAEs use encoder-decoder pairs to decompose and recompose two distinct faces. By then swapping the decoders, it is possible to obtain one of the two faces from the other face, yielding quite credible results.

The GAN-based methods use two different networks. The first one, network called discriminator, is trained to classify if an image is fake or not, and a second network called generator that instead must succeed, starting from a noisy image, to generate one sufficiently credible to deceive its counterpart. GANs have been particularly effective in the field of deepfake generation, with excellent results achieved with methodologies based on networks such as DiscoGAN [20], StarGAN [5] and StyleGAN-V2 [18].

Regardless of the technique used to carry out the manipulation, the various deepfake generation approaches are distinguished according to the specific way in which the image is modified. Among these the following stand out:

- *Face Transfer*: it transfers both identity-aware and identity-agnostic content (e.g. expression and pose) from a source face to the target face;
- *Face Swap*: it transfers the identity of the source face to the target face while preserving identity-agnostic content;
- *Face Stacked Manipulation (FSM)*: set of methodologies some that transfer both the identity and the attributes of the target on the source while others that alter the attributes of the swapped target after the transfer of the identity;
- *Face Reenactment*: it preserves the identity of the source subject but manipulates the intrinsic attributes such as mouth or expression;
- *Face Editing*: it edits external attributes such as age, gender or ethnicity.
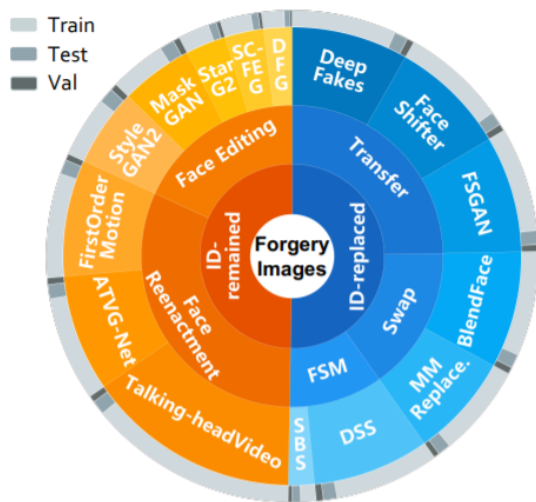
2.2 Deepfake Detection

With this wide variety of deepfake generation methodologies and their increasing effectiveness, it has therefore become extremely important to develop systems to distinguish a manipulated image from a real one. This is a problem that also affects other areas such as text, with recent work such as [11] analyzing deepfakes in tweets to identify fake content in social networks. Anyway, as these techniques are often applied on videos, many video deepfake detectors emerged. Some recent works proposed the exploitation of temporal information to recognise inconsistencies. For example [2] try to catch motion dissimilarities in the temporal structure of a video sequence by exploiting optical flow fields. However the majority of approaches are frame-based, classifying the video frame by frame. In order to train effective deep learning models for deepfake detection, a number of datasets have been created over the years, including the first DF-TIMIT [23], UADFC [35] and FaceForensics++ [29], Celeb-DF [26], Google Deepfake Detection Dataset [10] and the more recent DFDC [9], Deepforensics [16] and ForgeryNet [14]. The latter dataset is the most complete, largest and includes the greater variety of existing deepfake generation methods, since it is still recently published there are not many papers based on it. Much research has been carried out exploiting the DFDC dataset as it is one of the most complete and challenging in circulation and a specific type of convolutional neural network has emerged as particularly effective in fulfilling the task, the EfficientNet. The latter is the basis of many solutions that have obtained state-of-the-art results on the cited dataset such as the winning solution of the deepfake detection challenge [30]. With the advent of Vision Transformers and their successes in the field of Computer Vision, some interesting deepfake detection solutions have emerged. For example, the method presented in [34] which obtained good results by combining Transformers with convolutional networks used to extract patches from faces. Also interesting is the work done to exploit a pretrained EfficientNet B7 with a Vision Transformers by training it through distillation presented in [15]. A recent work on merging different types of Vision Transformers such as the Cross Vision Transformer [3] and EfficientNet B0 is presented in [6]. EfficientNet has recently been further improved with the introduction of EfficientNetV2 [32], a version of EfficientNet that is more optimised for smaller models, faster training and better ImageNet [7] accuracy than its predecessor and some Vision Transformers.

3 APPROACH

In this section, we analyze in detail the dataset and the models used to carry out our experiments.

3.1 Dataset

To be able to validate the ability of a neural network to detect deepfakes generated by methods other than those used for the construction of the training set, it is necessary to use a dataset containing a multitude of deepfake generation methods and keeps track of them. For this reason the dataset selected to carry out the experiments is ForgeryNet [14], one of the widest deepfake datasets available. ForgeryNet consists of 2.9M images and 220k video clips. For our experiments we will only use the set of images for which we have, for each of them, an associated label identifying if it has

Figure 1: Visualization of the various Deepfake generation techniques within the dataset grouped by category [14]

.

been manipulated or not and the method adopted to perform such manipulation. The fake images are generated through the use of 15 different manipulation approaches [4, 8, 12, 17, 19, 24, 25, 27, 28, 31] with more than 36 mix-perturbations on more than 4300 distinct subjects. Examples of applied perturbations are optical distortion, multiplicative noise, radom compression, blur and many others shown in more detail in the ForgeryNet paper. This variety therefore allows for the most comprehensive comparison possible between the two categories of neural networks. The methodologies used can be grouped into two macro-categories, *ID-remained* and *ID-replaced*, as shown in Figure 1. In the first case the identity of the subject in the image is not replaced but only manipulations are carried out on his face. In the second category, on the other hand, the identity is replaced by transferring a different face from the one actually present in the image. In turn, these categories are divided into 5 sub-categories: Face-Reenactment and Face Editing, belonging to the ID-remained category and Face Transfer, Face Swap and FSM, belonging instead to the ID-replaced category. All these approaches represent a large part of the main methods of deepfake generation known to date. The images in ForgeryNet also include people in a variety of contexts. To make the task of the two networks simpler, as in many other deepfake detection methods, a face extraction phase is carried out through the use of a state-of-the-art face detector, MTCNN [36]. The considered models are trained and tested on a face basis and we performed data augmentation like in [30]. Differently from them, we extracted the faces so that they were always squared and without padding. We used the Albumentations library [1], and during the training, we randomly applied common transformations. In particular, every time an image is passed to the network in training phase, this is resized randomly with three types of isotropic resize that differ for type of interpolation used (area, cubic or linear). After that, transformations are applied randomly, namely: image compression,

	Training Set		Validation Set	
	Real	Fake	Real	Fake
Subset 1 (0,1)	12.026	10.076	1,376	1,080
Subset 2 (0,7)	7,068	6,732	817	717
Subset 3 (0,10)	1.931	1,486	209	171

Table 1: Number of images per subset in the Single Methods experiment.

gaussian noise, horizontal flip, brightness or saturation distortion, grayscale conversion and shift, rotation or scaling.

3.2 Considered Architectures

In this research, we want to compare the cross-forgery generalisation capability of Convolutional Neural Networks, a category of neural networks widely used in this and many other Computer Vision tasks, with that of Vision Transformers [22], a more recent type of deep learning model that is proving to be particularly competitive. For the first category, an EfficientNetV2-M [33] was selected, a new version of the well-known EfficientNet that is more powerful and lighter. EfficientNets are widely used in deepfake detection and still form the basis of many state-of-the-art methods on the industry's leading datasets. The counterpart used is instead a ViT-Base, one of the first Vision Transformers presented and of similar dimensions with respect to the Convolutional network considered. Both networks are pretrained on ImageNet-21k and have been fine-tuned on sub-datasets extracted from ForgeryNet. The sub-datasets were constructed maintaining an almost perfect balance between fake and real images. In addition, only faces detected with a confidence level higher than 95% were considered in order to reduce the risk of false detection. The networks were trained freezing the weights of all blocks except for the last two which are specialised on the downstream task.
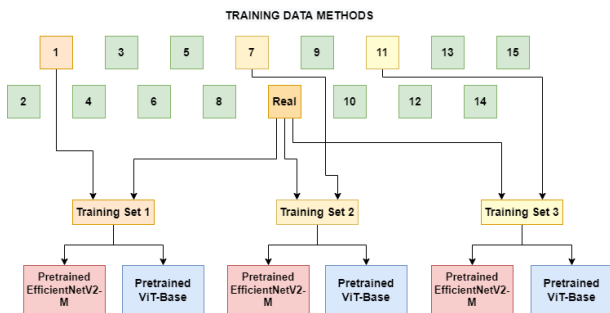
4 EXPERIMENTS

In this section, we describe all the experiments carried out in our research. The experiments are subdivided in two parts, the first in which we will use genuine images and images generated with a single method of deepfake generation at a time, and the second part in which instead we will consider more methods of deepfake generation, belonging to the same category, in the training phase. Since the labels of the ForgeryNet test set were not yet released at the time of the experiments, the validation set of this dataset, from now on called test set, was used to carry out all the tests while in the training phase, a portion equal to 10%, always the same for all the models, was selected from the sub-dataset considered. The latter will be called validation set from now on. The models were trained for a maximum of 50 epochs and a patience of 5 epochs on the validation set, using an SGD optimiser with a learning rate of 0.1 decreasing with a step size of 15 and a gamma of 0.1.

Figure 2: Line plots representing the accuracy values obtained by ViT-Base (blue) and EfficientnetV2-M (red) on the test set, trained on training sets consisting of real images and from left to right on images generated with the FaceShifter, Talking Head Video and StyleGAN2 methodology respectively. Observing the space between horizontal lines it is possible to notice how the variance of the Vision Transformer is lower than that of the CNN counterpart obtaining closer accuracies on different and also unseen methods. On the horizontal axis the 0 represent the real images and number from 1-15 the images generated with different generation methods.

4.1 Single Method Training

In this section, we describe the process used to investigate the ability of a model, trained on real images and images manipulated with a single deepfake generation method, to generalise the concept of deepfake to the point of recognizing images tampered with by other methods.

Figure 3: Training set construction for the Single-Method approach.

To perform this first comparison the two models were fine tuned on three subdatasets as shown in Figure 3. All of them contained unmanipulated images but also tampered images with a specific technique for each subdataset. The three techniques used are FaceShifter (1), Talking Head Video (7) and StyleGAN2 (11). These were selected as being quite different from each other so as to validate the effectiveness of the two networks on more varied manipulation approaches. As shown in Table 1, the sizes of the three datasets are quite varied but always well balanced between the two classes. In this experiment then, the models will only see anomalies introduced by one specific deepfakes generation method at a time. The models trained with the three sub-datasets were then tested with the images in the

	Training Set		Validation Set	
	Real	Fake	Real	Fake
Subset 1 (0,1,2,3)	59.237	59.215	6.596	6.566
Subset 2 (0,7,8,10)	10.408	11.232	1.200	1.205

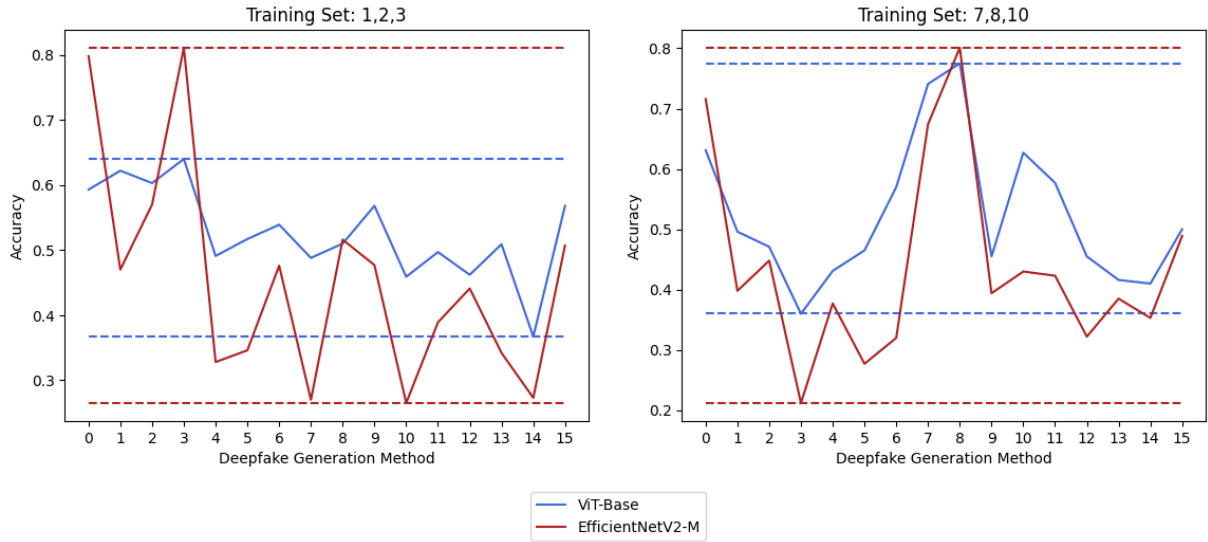**Table 2: Number of images per subset in the Multiple Methods experiment.**

test set, but also considering the generation methods not used in the training set.

As can be seen in the three plots in Figure 2 the Vision Transformers turn out to be more stable and less specialized. This model, even if reaches better results on the training methods, tends to have values of accuracy closer between the various methods of deepfake generation. On the other hand the EfficientNet often obtains higher accuracy than the Vision Transformer on the training methods but achieves much poorer accuracies on the others. For example, we can see a marked advantage of the EfficientNet over both method 7 and method 11 in the respective charts.

4.2 Multiple Methods Training

A further experiment was carried out by training on real images and on images manipulated with a group of methods belonging to the same category as shown in Figure 5. This choice derives from the fact that one of the two networks may be able to generalise even better in the presence of different generation methods, which therefore hopefully introduce a greater variety of artefacts.

For the first experiment, the training methods considered were those belonging to the Transfer category, i.e. FaceShifter (1), FS-GAN (2) and Deepfakes (3), as well as unmanipulated images. For

Figure 4: Line plots representing the accuracy values obtained by ViT-Base and EfficientNetV2-M on the test set, trained on a training set of real images and on images manipulated with methods belonging to the Transfer category (left) and those belonging to the Face Reenactment category (right). On the horizontal axis the 0 represent the real images and number from 1-15 the images generated with different generation methods.

the second experiment the methods belonging to the Face Reenactment category were used, namely Talking Head Video (7), ATVG-net (8) and First Order Motion (10). As shown in Table 2, both subsets are well balanced but differ in size. The former is in fact considerably larger, with about 120,000 traning images and representing the largest of the sets considered in the experiments, while the latter is smaller in size.

Figure 5: Training set construction for the Multiple-Methods approach.

Again, ViT-Base is found to have a significantly lower variance than EfficientNetV2-M in both experiments conducted, as can be seen from the horizontal lines in the charts in Figure 4. There is also a tendency for EfficientNet to focus on a subset of the methods presented in the training set. For example, in the case of the dataset consisting of images manipulated with Transfer techniques, the convolutional network obtains an accuracy value of 81.1% on method 3, while it remains particularly low on methods 1 and 2, with accuracy values of 47.0% and 57.0% respectively. On the other

hand, the Vision Tranformer obtains rather similar accuracy values on all three methods considered, although they are lower and are always around 62.0%.

The same behaviour can be observed in the second experiment. In this case, the EfficientNet reaches an accuracy of 80.1% on method 8 but drops drastically to 67.4% and 43.0% on the other two training methods, respectively 7 and 10. The Vision Transformer instead once again proves to be more stable with an accuracy of 74.1% and 77.5% on methods 7 and 8 and a less marked drop in performance on method 10 reaching 62.7%. In simple terms, in all plots the blue line representing Vision Transformers tends to remain higher than the red line representing EfficientNets.

To conclude, although the accuracies therefore tend to be rather low on the various novel methods, there is a tendency for Efficient-Nets to perform better on training methods than Vision Transformers, often achieving higher accuracies, but to generalise worse on novel methods.

4.3 Final Results

To numerically evaluate which of the two models is less likely to specialize in the deepfake generation methods used to build the training set, we chose to calculate the variance between the accuracies on the test set. More in detail, considering the list of 16 accuracies, one for each deepfake generation method plus the one obtained on the unmanipulated images, obtained from each network on the test images, the variance σ^2 was calculated as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

where $n = 16$ is the number of accuracies, x_i are the accuracy values and μ is their mean.

TRAINED		TESTED																																Variance	
		Real (0)		1		2		3		4		5		6		7		8		9		10		11		12		13		14		15			
		ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN	ViT	CNN
	0.1	63.1	73.7	69.8	66.9	45.9	42.5	30.4	20.3	50.1	36.9	51.6	25.6	39.8	41.6	48.0	38.3	46.1	45.8	50.0	43.9	39.9	29.6	53.1	48.0	36.4	35.0	37.3	44.7	41.7	36.7	41.4	56.4	0.009	0.018
	0.7	74.0	75.6	35.8	33.3	37.8	36.0	27.0	18.0	34.8	32.6	32.8	26.6	47.3	35.1	69.5	81.7	68.3	68.3	40.2	42.4	48.7	45.9	47.4	46.9	30.1	25.9	37.3	39.8	31.7	37.4	41.4	47.1	0.020	0.029
	0.11	57.8	55.5	48.1	59.1	55.5	62.6	38.8	35.8	42.7	55.5	60.9	42.5	52.6	55.3	60.4	63.6	62.3	73.3	56.1	71.2	48.1	50.1	59.4	75.4	51.0	60.1	47.8	62.1	43.9	60.4	62.9	73.6	0.005	0.011
	0.1.2.3	59.3	79.8	62.2	47.0	60.3	57.0	64.0	81.1	49.1	32.8	51.7	34.6	53.9	47.6	48.8	27.0	51.0	51.6	56.8	47.7	45.9	26.5	49.7	38.9	46.2	44.1	50.9	34.2	36.7	27.3	56.8	50.7	0.004	0.025
	0.7.8.10	63.1	71.6	49.6	39.8	47.1	44.8	36.0	21.2	43.1	37.7	46.5	27.7	57.0	32.0	74.1	67.4	77.5	80.1	45.5	39.4	62.7	43.0	57.7	42.3	45.5	32.2	41.6	38.5	41.0	35.3	50.0	48.9	0.013	0.024

Table 3: Table summarizing the accuracies obtained by the models on real test images (column 0) and on those manipulated with all deepfake generation methods considered (columns 1-15). The last column of the table contains the calculated variance values between the accuracies obtained by the models on the test set in the various deepfakes generation methods for each training sub-dataset.

A model with a lower variance will have obtained more similar accuracies between all deepfake generation methods regardless of whether they have been included in the training set and therefore will have learned the concept of deepfake more generally without specializing too much on the specific anomalies introduced by a method. From the data reported in the table 3 it can be seen that regardless of the methods used to build the training set, the variance associated with the Vision Transformers is always lower. On the other hand, in almost all cases the EfficientNet achieves higher levels of accuracy on training methods probably because it has learned to better recognize the specific anomalies introduced in these methods and all images containing different anomalies are considered non-deepfake. The case of the training on method 11 only, StyleGAN2, is interesting. In this case, both models have a marked decrease in performance on unmanipulated images. This probably derives from the fact that this specific method is particularly effective and introduces fewer anomalies than others present in the dataset, thus making the difference between a real image and a manipulated one more nuanced.

5 CONCLUSIONS

In this paper, we conducted a cross-forgery analysis to identify the most suitable deep learning architecture to tackle the deepfake detection task. The experiments carried out allowed us to have a first confirmation of the tendency of the Vision Transformers to better generalise the concept of deepfake, exhibiting less bias towards specific anomalies introduced by one or more deepfake generation techniques and thus making them more suitable to be applied in a real-world context. On the other hand, the convolutional networks and in particular, the EfficientNet, seem to be more prone to specialization, making them more applicable in contexts in which one wants to carry out deepfake detection, excluding the possibility that images manipulated with unpublished techniques can be introduced. Investigating the different ways of approaching the problem of the various deepfake detection solutions, not limiting ourselves exclusively to evaluating accuracy metrics on a subset of well-known and studied methods, is therefore fundamental to creating robust and long-lasting systems. With this research we have taken a step forward, highlighting in greater detail the behaviour of the main architectures used in the sector.

REFERENCES

[1] A. Buslaev, A. Parinov, E. Khvedchenya, V. I. Iglovikov, and A. A. Kalinin. 2018. Albumentations: fast and flexible image augmentations. *ArXiv e-prints* (2018). arXiv:1809.06839

[2] Roberto Caldelli, Leonardo Galteri, Irene Amerini, and Alberto Del Bimbo. 2021. Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters* 146 (2021), 31–37. https://doi.org/10.1016/j.patrec.2021.03.005

[3] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899* (2021).

[4] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7824–7833. https://doi.org/10.1109/CVPR.2019.00802

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8789–8797. https://doi.org/10.1109/CVPR.2018.00916

[6] Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. arXiv:2107.02612 [cs.CV]

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5153–5162. https://doi.org/10.1109/CVPR42600.2020.00520

[9] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* (2020).

[10] Nick Dufour and Andrew Gully. 2019. Contributing data to deep-fake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[11] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *Plos one* 16, 5 (2021), e0251415.

[12] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. 2019. Text-based Editing of Talking-head Video.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. In *Advances in neural information processing systems 27*. arXiv:1406.2661 [stat.ML]

[14] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. 2021. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4358–4367. https://doi.org/10.1109/CVPR46437.2021.00434

[15] Young-Jin Heo, Young-Ju Choi, Young-Woon Lee, and Byung-Gyu Kim. 2021. Deepfake Detection Scheme Based on Vision Transformer and Distillation. *arXiv preprint arXiv:2104.01353* (2021).

[16] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 2020. 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*. 2889–2898.

[17] Youngjoo Jo and Jongyoul Park. 2019. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 1745–1753. https://doi.org/10.1109/ICCV.2019.00183

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

[20] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*. PMLR, 1857–1865.

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

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

[23] Pavel Korshunov and Sébastien Marcel. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685* (2018).

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5548–5557. https://doi.org/10.1109/CVPR42600.2020.00559

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing High Fidelity Identity Swapping for Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5073–5082. https://doi.org/10.1109/CVPR42600.2020.00512

[26] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3207–3216.

[27] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 7183–7192. https://doi.org/10.1109/ICCV.2019.00728

[28] Ivan Petrov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, RP Luis, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: A simple, flexible and extensible face swapping framework. *ArXiv* abs/2005.05535 (2020).

[29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8110–11.

[30] Selim Seferbekov. 2020. *DFDC 1st place solution*. "https://github.com/selimsef/dfdc_deepfake_challenge"

[31] Aliaksandr Siarohin, Stephane Lathuiliere, S. Tulyakov, Elisa Ricci, and N. Sebe. 2019. First Order Motion Model for Image Animation. *ArXiv* abs/2003.00196 (2019).

[32] Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. arXiv:2104.00298 [cs.CV]

[33] Mingxing Tan and Quoc V. Le. 2021. EfficientNetV2: Smaller Models and Faster Training. (2021). https://doi.org/10.48550/ARXIV.2104.00298

[34] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv preprint arXiv:2102.11126* (2021).

[35] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8261–8265.

[36] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.