Preprints.org

Article

On the Generalization of Deep Learning Models in Video Deepfake Detection

Davide Alessandro Coccomini [*] , Roberto Caldelli , Fabrizio Falchi , Claudio Gennaro

Article

On the Generalization of Deep Learning Models in Video Deepfake Detection

Davide Alessandro Coccomini [1,*] ![ORCID], **Roberto Caldelli** [2,3] ![ORCID], **Fabrizio Falchi** [1] ![ORCID] and **Claudio Gennaro** [1] ![ORCID]

1   Istituto di Scienza e Tecnologie dell'Informazione, 56124 Pisa, Italy; name.surname@isti.cnr.it
2   CNIT, 50134 Florence, Italy; roberto.caldelli@unifi.it
3   Mercatorum University, 00186 Rome, Italy
*   Correspondence: davidealessandro.coccomini@isti.cnr.it

Abstract: The increasing use of deep learning techniques to manipulate images and videos, commonly referred to as "deepfakes," is making more and more challenging to differentiate between real and fake content. While various deepfake detection systems have been developed, they often struggle to detect deepfakes in real-world situations. In particular, these methods are often unable to effectively distinguish images or videos when these are modified using novel techniques which have not been used in the training set. In this study, we carry out an analysis of different deep learning architectures in an attempt to understand which is more capable of better generalizing the concept of deepfake. According to our results, it appears that Convolutional Neural Networks (CNNs) seem to be more capable of storing specific anomalies and thus excel in cases of datasets with a limited number of elements and manipulation methodologies. The Vision Transformer, conversely, is more effective when trained with more varied datasets, achieving more outstanding generalization capabilities than the other methods analysed. Finally, the Swin Transformer appears to be a good alternative for using an attention-based method in a more limited data regime. All the analyzed architectures seem to have a different way to look at deepfakes but since in a real-world environment, the generalization capability is essential, based on the carried out experiments the Vision Transformer seems to provide superior performances.

Keywords: deepfake detection; deep learning; computer vision; generalization

1. Introduction

Deep Learning has greatly impacted society, leading to impressive advancements in various fields. However, its use can also have negative consequences, like the creation of deepfakes. Deepfakes are manipulated images or videos that depict subjects in ways they never actually did or said, which can harm reputations or manipulate reality. Indeed, although deepfakes have numerous potential applications in the fields of entertainment, art, and education, they also pose significant security and ethical risks. For this reason, it is crucial to continue the development of robust deepfake detection methods to counteract such a threat. To tackle this problem, researchers have developed deepfake detection techniques, that are usually based on deep learning as well. These methods try to identify any anomalies introduced during the manipulation process, but they require large amounts of data for training. Furthermore, deepfakes are generated by resorting to different typologies of techniques and/or procedures (often even unknown) that emerge almost daily so it is not possible to follow each methodology and consequently to re-adapt the training phase. On this basis, to have more effective deepfake detectors, researchers aim for a system that can generalize the concept of deepfakes and identify them regardless of the manipulation technique used, even if it is novel and not present in the training data. In this research, a comparison was made among different deep learning architectures in order to validate their generalization capabilities specifically against deepfake videos. It is worthy to say that techniques used to manipulate videos do not necessarily introduce the same anomalies and features that can be embedded when tampering with still images. It is therefore interesting to

show how different models behave in this context. In particular, we compared three different kinds of network architectures: a convolutional network, such as *EfficientNet V2*, a standard *Vision Transformer* and also a *Swin Transformer* which is a specific type of transformer. Our experiments indicate that the Vision Transformer outperforms other models in terms of generalization ability. This stems in the availability of large quantity and diversity of data used during training which enable the model to abstract better the concept of deepfake. On the other hand, the Vision Transformer struggles to learn when data is limited, unlike the EfficientNet-V2 and Swin Transformer, which perform satisfactorily even under such constraints.

2. Related Works

2.1. Deepfake Generation

Deepfake generation techniques refer to the methods used to manipulate a human face, changing its appearance or identity in a realistic manner. There are two main categories of approaches: those based on Variational AutoEncoders (VAEs) [1] and those based on Generative Adversarial Networks (GANs) [2]. VAE-based methods use encoder-decoder pairs to decompose and recompose two distinct faces. By swapping the decoders, it is possible to transform one face into the other, resulting in a credible output. GAN-based methods, on the other hand, use two different networks: a discriminator, trained to classify whether an image is fake or real, and a generator that generates a fake face to fool the discriminator. This results in a feedback loop, where the generator is trained to improve its performance based on the output of the discriminator. GANs are typically more powerful than VAEs, but also more challenging to train. Some of the most popular GAN-based deepfake generation methods include Face2Face [3] and FaceSwap [4]. Recently, a number of other deepfake generation approaches have been proposed, leveraging the advancements in computer vision and deep learning. For example, [5] presents a method for synthesizing realistic talking heads from a single source video. [6] proposes StyleGAN, a highly-customizable deepfake generation method that allows for fine-grained control over the generated images.

2.2. Deepfake Detection

As deepfake generation methods become increasingly sophisticated, there is a growing need for systems that can distinguish between real and manipulated images. This is a problem not just in the field of images, but also in text, where recent work such as [7] has analyzed deepfakes in tweets to identify fake content on social networks. To tackle the challenge of deepfake detection in videos, many video-based deepfake detectors have been developed. Even if some approaches propose solutions which analyze also the temporal information of manipulated videos[8–11], the majority of methods are frame-based, classifying each video frame individually. Also, several competitions have been organized to stimulate the resolution of this task like [12] and [13]. To train effective deep learning models for deepfake detection, numerous datasets have been created over the years, including DF-TIMIT [14], UADFC [15], FaceForensics++ [16], Celeb-DF [17], Google Deepfake Detection Dataset [18], DFDC [12], Deepforensics [19], and ForgeryNet [20]. The latter dataset, which is the most complete, large, and diverse, has recently emerged as a popular choice for deepfake detection research. One type of convolutional neural network, EfficientNet, has emerged as particularly effective in solving the task, and is the basis of many state-of-the-art solutions, such as the winning solution of the deepfake detection challenge [21]. More recently, with the rise of Vision Transformers in Computer Vision, new deepfake detection solutions have been developed, such as the method in [22] which combines Transformers with convolutional networks to extract patches from faces, and the approach in [23] which uses a pretrained EfficientNet B7 with a Vision Transformer, trained through distillation. An innovative work on combining different types of Vision Transformers, such as the Cross Vision Transformer [24] and EfficientNet B0, is presented in [25]. EfficientNet has been further improved with the introduction of EfficientNetV2 [26], a version that is optimized for smaller models, faster

training, and better ImageNet performance. An evolution of this approach which combines together a TimeSformer and a Convolutional Neural Network is presented in [8] where also various deepfake detection problems like multi-identity and face size variation are treated.

3. The followed approach and the tested network architectures

To validate the neural network's ability to detect deepfakes generated by methods not used in its training set, a dataset containing a variety of deepfake generation methods and labels is needed. The chosen dataset for this purpose is ForgeryNet [20], which is one of the most comprehensive deepfake datasets available, containing 2.9 million images and 220,000 video clips. The fake images are manipulated using 15 different manipulations while the videos are manipulated using only 8 of them [27–36]. To each image and video, more than 36 mix-perturbations are randomly applied on more than 4300 distinct subjects. Examples of applied perturbations are optical distortion, multiplicative noise, random compression, blur and many others shown in more detail in the ForgeryNet paper [20]. Furthermore, the different manipulations applied can be grouped into two macro-categories, *ID-remained* and *ID-replaced*. The first category involves manipulations of the subject's face without changing their identity, while the second category involves replacing the subject's face with a different one. These two categories are further divided into four sub-categories: all the videos falling under the ID-remained category are manipulated with Face Reenactment methods, while the ID-replaced class is divided into Face Transfer, Face Swap, and Face Stacked Manipulation (FSM). These sub-categories collectively make up a significant portion of the deepfake generation techniques currently known. The ForgeryNet dataset includes people in various settings and situations.

The extracted frames are pre-processed, as done in many other deepfake detection methods [8–10,22,25] by introducing a face extraction step using the state-of-the-art face detector, MTCNN [37]. The models are trained and evaluated on a per-face basis and data augmentation was performed, similar to [8,21,25]. However, we extracted the faces to be squared and without padding. We exploited the Albumentations library [38] and applied common transformations randomly during training. Whenever an image is input to the network during training, it is randomly resized using three types of isotropic resize with different interpolation methods (area, cubic, or linear). Afterwards, random transformations such as image compression, gaussian noise, horizontal flip, brightness or saturation distortion, grayscale conversion, shift, rotation, or scaling, are applied.

The present paper is derived from another work [39] where we already conducted a similar cross-forgery analysis on the part of the dataset consisting of still images, in this case, we performed on videos and, in particular, we have made a specific comparison among different kinds of architectures. It is important to carry out this analysis on videos because the anomalies that are introduced in videos can also differ greatly from those that may result from the manipulation of a single image, and therefore, the behaviour of the various deep learning methods can also change greatly. In the ForgeryNet dataset, there is a label assigned to each video indicating whether it has been manipulated or not. Additionally, the label specifies the method employed to perform the manipulation. Among the methods used, FaceShifter and ATVG-Net manipulate all frames of the video, while the other methods partially manipulate the video frames, providing information on which frames have been manipulated and which ones are left unaltered.

To perform this comparative analysis on cross-forgery generalization capability we have considered three kinds of network architectures. A Convolutional Neural Networks (CNNs), a widely used type of neural network in computer vision, and a Vision Transformers (ViTs) [40], a newer, highly competitive deep learning model. As representative of the CNN category, we have taken an *EfficientNetV2-M* [41], which is a newer and more advanced version of the well-known EfficientNet. EfficientNets are widely used in deepfake detection and remain a cornerstone of many state-of-the-art methods on leading datasets. On the contrary, for the Vision Transformers, we have used the *ViT-Base*, a ViT of similar dimensions to the CNN which is one of the first versions introduced. However, a further third architecture, such as the Swin Transformer [42], has been taken into account; this has been

done because this type of Transformers is particularly interesting for our analysis in that although it is attention-based, the computation of attention takes place in a hierarchical manner emulating the convolutional layers of CNNs. Swin Transformer is an architecture for image classification that improves the traditional transformer approach by using hierarchical feature representations and a window-based attention mechanism. It divides the input image into patches and transforms them into low-dimensional feature vectors using a learnable projection. These vectors are then passed through a series of transformer blocks, organized into stages, to capture spatial and channel-wise dependencies. Finally, the output is passed through a classification head to produce the class probabilities. Swin Transformer achieves state-of-the-art performance while being computationally efficient and scalable to larger image sizes. The Swin Transformer selected for our experiments is the *Swin-Small*. All the models were pretrained on ImageNet-21k and fine-tuned on sub-datasets from ForgeryNet, which were constructed with a nearly equal balance of fake and real images as explained in the next section. To reduce false detections only faces with a confidence level above 95% were included. All networks were trained by keeping freeze a number of layers such that the trained parameters correspond to approximately 45M. In particular, only the last layers of the models considered were made trainable so that the number of parameters was always comparable between the various experiments, while the other layers' weights remained at the values based on the pretraining.

4. Experiments

4.1. General setup

The experiments conducted in our research are divided into two parts. In the first part, we used frames from pristine videos and manipulated frames from fake videos generated with one deepfake generation method at a time to compose a training set. Each obtained model is then tested against frames extracted from videos manipulated with the same generation methodology used at training time but also against other methods not seen during the training phase in order to investigate the generalization capacity of the different architectures. The classification task is always conducted frame-by-frame. In the second part, we used multiple deepfake generation methods grouped by category (ID-Replaced or ID-Remained) to construct the training set. Since the labels of the ForgeryNet test set were not available at the time of the experiments, we used the validation set, which we will refer to as the test set, for all evaluations. During training, a 10% portion, consistent for all models, was randomly selected from the training set and referred to as the validation set. The models were trained for up to 50 epochs with a patience of 5 epochs on the validation set, using an SGD optimizer with a learning rate of 0.1 that decreases with a step size of 15 and a gamma of 0.1.

4.2. Single Method Training

In this section, we outline the process used to examine a model's ability to recognize images manipulated by various deepfake generation methods, despite being trained on real images and images manipulated with only one deepfake method. In the first comparison, the three models under consideration, namely EfficientNetV2-M, ViT-Base and Swin-Small, were fine-tuned on each of the eight sub-datasets as illustrated in Figure 1. These sub-datasets consisted of both unaltered frames and frames manipulated using specific techniques, enumerated as FaceShifter(1), FS-GAN(2), DeepFakes(3), BlendFace(4), MMReplacement(5), DeepFakes-StarGAN-Stack(6), Talking-Head Video(7) and ATVG-Net(8).
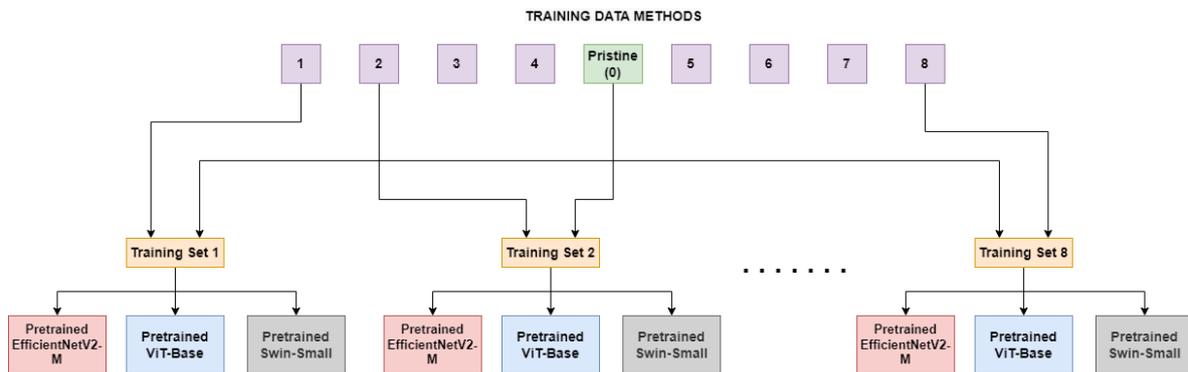
Figure 1. The *Single Method Training* setup: eight different training sets are constructed, each consists of frames manipulated with a deepfake generation method and pristine frames.

As displayed in Table 1, the sizes of the datasets vary quite largely. Pristine frames are most common within the dataset so, to ensure a good balance but without sacrificing too many of them, a subset equal to the number of fake frames of the method under training is randomly selected at each epoch. In this experiment, the models will only encounter, during the training, anomalies generated by one specific deepfake method at a time. The models trained on the sub-datasets were then tested on frames in the test set, including those manipulated by methods not used during training.

Table 1. Number of frames for Single Methods Training and Test setup.

Video Manipulation Methods	Training Frames	Test Frames
0 (*Pristine*)	118,255	47,369
1 (FaceShifter)	13,337	1,889
2 (FS-GAN)	48,122	8,732
3 (DeepFakes)	8,550	1,157
4 (BlendFace)	9,827	1,335
5 (MMReplacement)	270	115
6 (DeepFakes-StarGAN-Stack)	3,610	509
7 (Talking-Head Video)	26,338	2,199
8 (ATVG-Net)	37,449	5,383

4.3. Multiple Methods Training

A second experiment has been conducted by training the models on real frames and frames manipulated using a group of methods belonging to the same category (*ID-replaced* or *ID-remained*), as shown in Figure 2. This was done to determine if the networks can better generalize in the presence of diverse categories of manipulation methods, which may introduce a greater variety of artifacts.

In Table 2, the sizes, in terms of available frames, of the two different categories can be seen. As depicted in Figure 2, two models (for each network architecture) have been trained: the first one (*ID-Replaced*) is based on frames crafted by using methods belonging to the ID-Replaced category (methods from 1 to 6), while the second one (*ID-Remained*) is based on those ones coming from the ID-Remained category (methods 7 and 8). In both of them, also pristine images from unaltered videos are added to the training dataset.
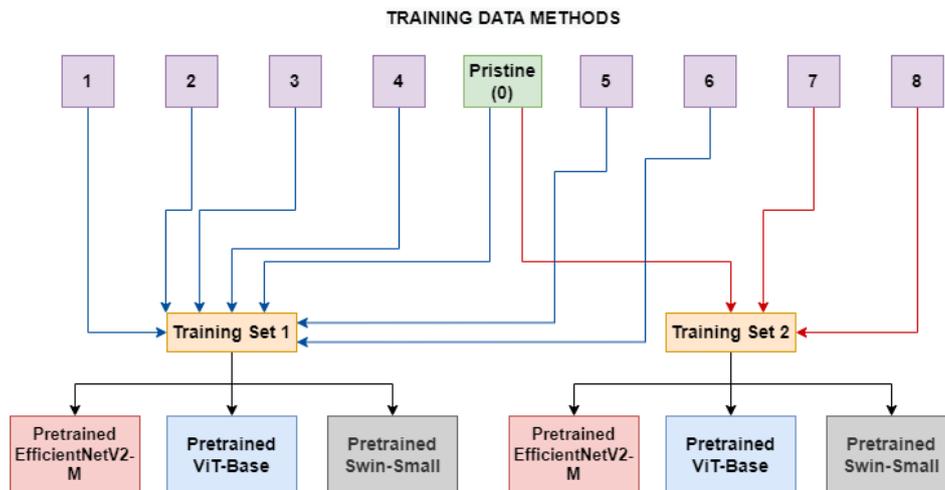
Figure 2. The *Multiple Method Training* setup: two different training sets are constructed, each consists of frames manipulated with deepfake generation methods related to the same category (blue lines for *ID-Replaced* and red lines for *ID-Remained*) and pristine frames.

Table 2. Number of frames for Multiple Methods Training and Test.

Video Manipulation Categories	Training Frames	Test Frames
0 (*Pristine*)	118,255	47,369
ID-Replaced	83,716	13,737
ID-Remained	63,787	7,582

5. Results

5.1. Single Method Training

Figure 3 shows the accuracies achieved by the three considered models trained in the *Single Method Training* setup, presented in the previous section, with respect to each of the methods comprised within the test set. Looking at the accuracies of the three models, it can be pointed out that the EfficientNetV2-M and the Swin-Small maintain results often above 80% in correspondence of test frames manipulated with the same methods used in the training set (as expected) and, at the same time, obtain a certain degree of generalization. In fact, the same models sometimes succeed in detecting frames manipulated with methods unseen during training, anyway reaching values of accuracy that are quite limited. The case of method number 5 (*MMReplacement*) is rather anomalous, though the detection percentage is often very high indeed; this behaviour is probably induced by the low number of available examples (see Table 1).

On the contrary, it can be easily noticed that, in all the cases, the ViT-Base is substantially unable to learn in the presence of relatively few training images. In fact, for instance, by training the model on methods 3, 4, 5 and 6 and then testing it on the test set, it is evident that the model is substantially underfitting and practically unusable compared to the two others taken into consideration. Interestingly, the Swin Transformer, although also based on the attention mechanism, is not particularly affected by this phenomenon and instead succeeds in obtaining competitive results on all contexts. This probably lies in the hierarchical nature that emulates the convolutional layers of traditional CNNs and thus allows it to exploit implicit inductive biases better. Good performances are anyway preserved, in any case, with respect to pristine frame detection. In this setup, the architecture based on a convolutional network seems to prove more capable of generalization.
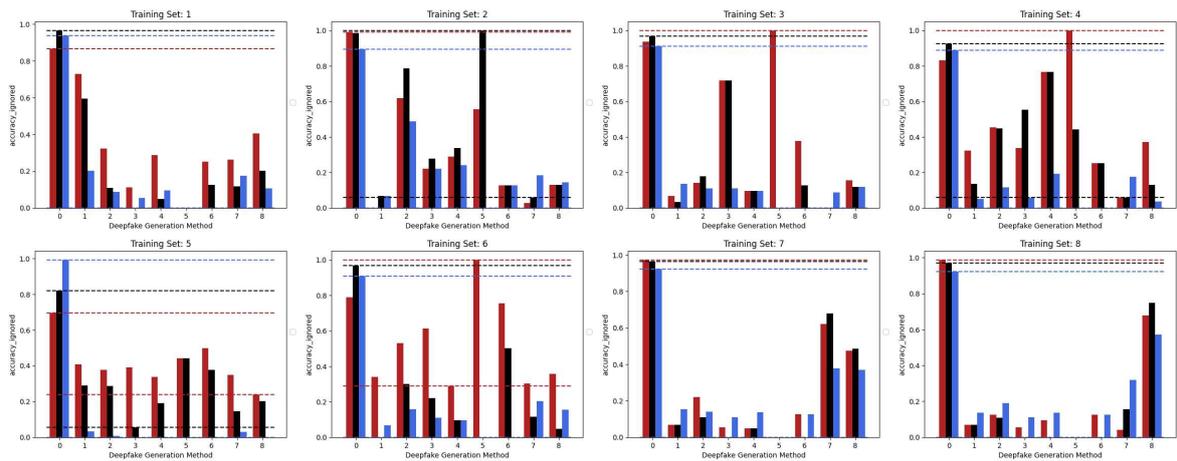
Figure 3. The performance in terms of accuracy achieved by each of the three considered models with respect to the 8 different training sets following the *Single Method Training* setup: EfficientNetV2-M (red), Swin-Small (black) and ViT-Base (blue).

The accuracies obtained from the three models are also shown in the confusion matrices in Figure 4 where all previously commented trends are reconfirmed again.
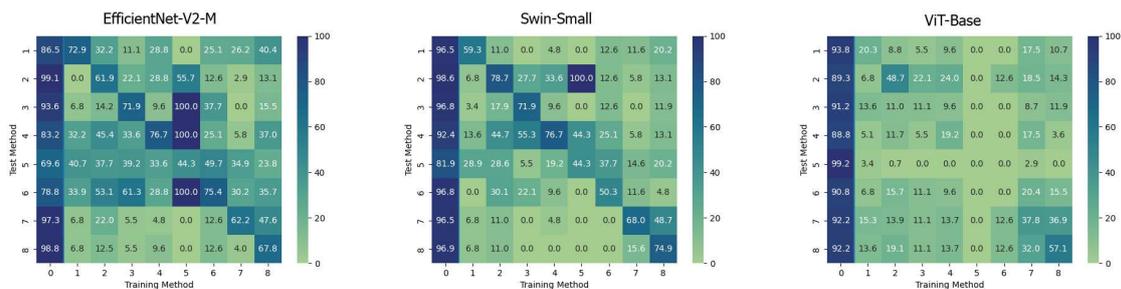

Figure 4. Confusion matrices of the frame-level accuracy values for the three models under consideration trained in the *Single Method Training* setup and tested on frames manipulated with all the available methods respectively.

5.2. Multiple Methods Training

The behaviour of the three networks is now analysed in the second considered setup, namely *Multiple Methods Training*, and corresponding results are shown in Figure 5. In this case, the datasets are composed of frames extracted from videos manipulated by not-only one method, so the models will have more difficulty to focus on specific artifacts and be forced to generalize. In this setup, the situation is significantly different from the previous one. Surprisingly, the classic Vision Transformer, which previously struggled to train effectively, is now the only model capable of generalizing well to frames that have been manipulated using techniques that were not present in the training data. This result probably stems from the fact that the training set consists of significantly more images than in the previous setup and it is strongly in line with what presented in [39].

In this case, the confusion matrices (see Figure 6) clearly show the greater generalization capacity of the Vision Transformer although at the expense of more false positives. In fact, the 'pristine' class is less accurately classified by this latter.
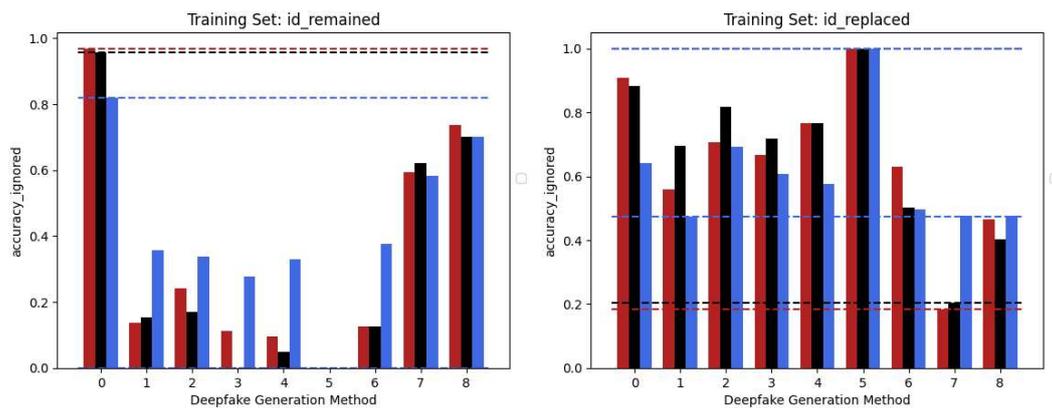
Figure 5. Accuracy performances achieved by each of the models considered in the two different training sets constructed following the *Multiple Methods Training* setup: EfficientNetV2-M (red), Swin-Small (black) and ViT-Base (blue). *ID-Replaced* methods (1-6), *ID-Remained* methods (7-8) and *Pristine* (0).
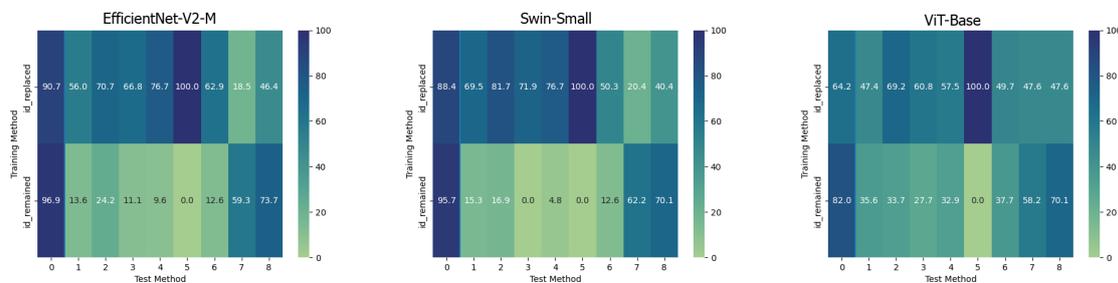

Figure 6. Confusion matrices of the frame-level accuracy of the three models trained in the *Multiple Method Training* setup and tested on frames manipulated with all available methods.

6. Conclusions

In this study, we investigated the generalization capabilities for detecting deepfake videos of various deep learning architectures by using two different setups. The first setup involved a limited training set constructed from both pristine and manipulated video frames adopting a specific method at a time and then testing versus all the different methods. In this setup, the EfficientNet-V2 convolutional network outperformed the Vision Transformer in learning from the less diverse and limited training data, while the Swin Transformer showed promising results. In the second setup, we considered a larger and more variegate training dataset that included frames coming from deepfake videos on which have been applied different manipulation methods but belonging to the same category (*ID-Replaced* or *ID-Remained*) and then performing cross-testing. Interestingly, the Vision Transformer demonstrated superior generalization capabilities and outperformed the convolutional network in detecting frames from videos manipulated with novel methods.

Our findings suggest that in real-world scenarios where large, diverse deepfake detection datasets are available and generalization is critical, the Vision Transformer may be the optimal choice for detecting deepfakes. However, in cases where the training data is limited, a convolutional network such as EfficientNet-V2 may be more suitable. The Swin Transformer provides a good balance between the two in terms of generalization and performance.

Overall, our study highlights the significance of considering the specific characteristics of the dataset and deep learning architecture when detecting deepfakes in order to be able to create a detector which may really be applied in the real world.

References

1. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**.
2. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. Advances in neural information processing systems 27, 2014, [arXiv:stat.ML/1406.2661].
3. Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2387–2395. doi:10.1109/CVPR.2016.262.
4. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast Face-Swap Using Convolutional Neural Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3697–3705. doi:10.1109/ICCV.2017.397.
5. Liu, M.Y.; Huang, X.; Yu, J.; Wang, T.C.; Mallya, A. Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications. *Proceedings of the IEEE* **2021**, *PP*, 1–24. doi:10.1109/JPROC.2021.3049196.
6. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks, 2018. doi:https://doi.org/10.48550/arxiv.1812.04948.
7. Fagni, T.; Falchi, F.; Gambini, M.; Martella, A.; Tesconi, M. TweepFake: About detecting deepfake tweets. *Plos one* **2021**, *16*, e0251415.
8. Coccomini, D.A.; Zilos, G.K.; Amato, G.; Caldelli, R.; Falchi, F.; Papadopoulos, S.; Gennaro, C. MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection, 2022. doi:10.48550/ARXIV.2211.10996.
9. Baxevanakis, S.; Kordopatis-Zilos, G.; Galopoulos, P.; Apostolidis, L.; Levacher, K.; Schlicht, I.B.; Teyssou, D.; Kompatsiaris, I.; Papadopoulos, S. The MeVer DeepFake Detection Service: Lessons Learnt from Developing and Deploying in the Wild. Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, 2022.
10. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring Temporal Coherence for More General Video Face Forgery Detection. ICCV, 2021, pp. 15024–15034. doi:10.1109/ICCV48922.2021.01477.
11. Caldelli, R.; Galteri, L.; Amerini, I.; Del Bimbo, A. Optical Flow based CNN for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters* **2021**, *146*, 31–37. doi:https://doi.org/10.1016/j.patrec.2021.03.005.
12. 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**.
13. Guarnera, L.; Giudice, O.; Guarnera, F.; Ortis, A.; Puglisi, G.; Paratore, A.; Bui, L.M.Q.; Fontani, M.; Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Messina, N.; Amato, G.; Perelli, G.; Concas, S.; Cuccu, C.; Orrù, G.; Marcialis, G.L.; Battiato, S. The Face Deepfake Detection Challenge. *Journal of Imaging* **2022**, *8*. doi:10.3390/jimaging8100263.
14. Korshunov, P.; Marcel, S. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685* **2018**.
15. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 8261–8265.
16. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1–11.
17. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207–3216.
18. Dufour, N.; Gully, A. Contributing data to deep-fake detection research, 2019.
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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2889–2898.
20. He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; Liu, Z. ForgeryNet: A Versatile Benchmark for Comprehensive Forgery Analysis. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4358–4367. doi:10.1109/CVPR46437.2021.00434.
21. Selim Seferbekov. *DFDC 1st place solution*. 2020. URL: https://github.com/selimsef/dfdc_deepfake_challenge.
22. Wodajo, D.; Atnafu, S. Deepfake Video Detection Using Convolutional Vision Transformer. *arXiv preprint arXiv:2102.11126* **2021**.

23.　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**.

24.　Chen, C.F.; Fan, Q.; Panda, R. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899* **2021**.

25.　Coccomini, D.; Messina, N.; Gennaro, C.; Falchi, F. Combining EfficientNet and Vision Transformers for Video Deepfake Detection, 2022, [arXiv:cs.CV/2107.02612].

26.　Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training, 2021, [arXiv:cs.CV/2104.00298].

27.　Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7824–7833. doi:10.1109/CVPR.2019.00802.

28.　Deng, Y.; Yang, J.; Chen, D.; Wen, F.; Tong, X. Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5153–5162. doi:10.1109/CVPR42600.2020.00520.

29.　Ohad Fried et al. *Text-based Editing of Talking-head Video*. June 2019.

30.　Jo, Y.; Park, J. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1745–1753. doi:10.1109/ICCV.2019.00183.

31.　Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8107–8116. doi:10.1109/CVPR42600.2020.00813.

32.　Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5548–5557. doi:10.1109/CVPR42600.2020.00559.

33.　Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Advancing High Fidelity Identity Swapping for Forgery Detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5073–5082. doi:10.1109/CVPR42600.2020.00512.

34.　Nirkin, Y.; Keller, Y.; Hassner, T. FSGAN: Subject Agnostic Face Swapping and Reenactment. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7183–7192. doi:10.1109/ICCV.2019.00728.

35.　Siarohin, A.; Lathuiliere, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First Order Motion Model for Image Animation. *ArXiv* **2019**, *abs/2003.00196*.

36.　Petrov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Luis, R.; Jiang, J.; Zhang, S.; Wu, P.; Zhou, B.; Zhang, W. DeepFaceLab: A simple, flexible and extensible face swapping framework. *ArXiv* **2020**, *abs/2005.05535*.

37.　Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* **2016**, *23*, 1499–1503.

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

39.　Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Amato, G. Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection. Proceedings of the 1st International Workshop on Multimedia AI against Disinformation; Association for Computing Machinery: New York, NY, USA, 2022; MAD '22, p. 52–58. doi:10.1145/3512732.3533582.

40.　Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; Unterthiner, T.; Zhai, X. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021.

41.　Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training **2021**. doi:10.48550/ARXIV.2104.00298.

42.　Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6559–6568.