LatentForensics: Towards frugal deepfake detection in the StyleGAN latent space

Matthieu Delmas
CentraleSupélec
IETR (UMR CNRS 6164)
Rennes, France
matthieu.delmas@centralesupelec.fr

Amine Kacete
IRT bcom
Cesson Sévigné, France

Simon Leglaive
CentraleSupélec
IETR (UMR CNRS 6164)
Rennes, France

Abstract—The classification of forged videos has been a challenge for the past few years. Deepfake classifiers can now reliably predict whether or not video frames have been tampered with. However, their performance is tied to both the dataset used for training and the analyst's computational power. We propose a deepfake detection method that operates in the latent space of a state-of-the-art generative adversarial network (GAN) trained on high-quality face images. The proposed method leverages the structure of the latent space of StyleGAN to learn a lightweight binary classification model. Experimental results on standard datasets reveal that the proposed approach outperforms other state-of-the-art deepfake classification methods, especially in contexts where the data available to train the models is rare, such as when a new manipulation method is introduced. To the best of our knowledge, this is the first study showing the interest of the latent space of StyleGAN for deepfake classification. Combined with other recent studies on the interpretation and manipulation of this latent space, we believe that the proposed approach can further help in developing frugal deepfake classification methods based on interpretable high-level properties of face images.

Index Terms—Deepfakes, Computer Vision, Latent Space, Frugality

I. INTRODUCTION

Forgery of videos, the creation of so-called deepfakes, has been on the rise for the past few years. Although it yields multiple benefits in different domains (*e.g.* special effects and data generation), its democratization comes with a few risks. It is now easier than ever for malevolent users to create fake media in order to discredit trusted sources, impersonate powerful political figures, or blackmail individuals. Fortunately, recent work in media forensics have yielded numerous ways to discriminate deepfakes from genuine videos. Most of those methods rely on the use of Convolutional Neural Networks (CNNs) [1] which are tailored to detect specific weaknesses or artifacts left by the forgery process [2] [3]. However, this means that as the forgery process improves, the cost of training such discriminators will increase. Working towards building models which require less resources such as training examples, computing power and time -also known as frugal AI- is a growing subject in the deep learning community, as sustainability becomes an ever-increasing priority.

A forensic method is proposed here to classify forged face images from genuine ones. Its training requires considerably less resources than current state-of-the-art models while still remaining competitive or even achieving better performance.

Recently, artificial image generation through Generative Adversarial Nets (GANs), notably of faces, has seen great progress with models such as StyleGAN [4] or CycleGAN [5] in both the quality of results and the understanding of the generation process. It is proposed to capitalize on those improvements by projecting suspect image data in the latent space of StyleGAN before performing the classification. The resulting pipeline, represented Figure 1, is quite simple but its strength resides in its combined efficiency and how easy it is to set up, especially compared to other state-of-the-art models.

Our contributions are as follows:

- A comparison of dimensionality reduction methods, including StyleGAN pseudo-inversions in the context of deepfake classification.
- An easy-to-train and effective deepfake classifier which works on images inverted in the latent space of StyleGAN.
- A demonstration in the context in which the proposed method is best-suited : when only limited data or computing power is available.

With the goal of reducing the dimensonality of the data in mind, a first study is made to compare the latent space of StyleGAN to a Principal Component Analysis (PCA) on the DeepfakeDetectionChallenge (DFDC) preview dataset [6]. Then, as there are numerous ways to obtain a StyleGAN latent code from an image, a benchmark of different StyleGAN inversion methods is made to find the most suitable in a deepfake classification context. Next, with the best possible tools, an end-to-end binary discrimination pipeline is set up and compared to state of the art models (MesoNet [2], XceptionNet [3] and EfficientNet [7]) on the CelebDF v2 [8] database. A particular importance is given to computing and data requirements to train the models. Finally, the approach is further pursued by studying how the proposed approach performs on small databases.

II. RELATED WORK

There are two main trains of thought when it comes to discerning a deepfake from a genuine video: artefact detection and subject identification. In the former method, models
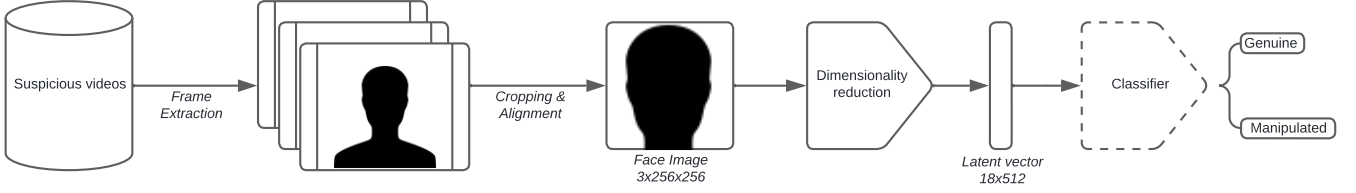
Fig. 1. The proposed pipeline, in dotted lines the trained model.

trained on images spot where manipulations have been made, these include MesoNet [2] or XCeptionNet [3]. Such models rely on inconsistencies in image patches, *i.e.* the presence of visual artifacts left by the forgery process. On the other hand, identification methods train a model to try to recognize if the portrayed person is the expected subject or an impersonator, instead of trying to spot if a certain frame has been tampered with.

A. Detection of Image Artifacts

CNNs have been achieving top performance in analyzing video frames for deepfake classification. MesoNet [2], for instance, is a lightweight convolutional network which relies on the density of details in image patches to detect whether or not an image have been tampered with. Such networks have been state of the art for a while, as highlighted by Rossler *et al.* in their FaceForensics++ (FF++) study [9]. More recently, the Deep Fake Detection Challenge [10] (DFDC) allowed multiple teams to compete in providing the best detection methods. Models based on FAb-NEt [11] came on top and achieved consistent performance. These neural networks are arguably one of the best solution to the deepfake detection problem, if one has enough computing power, training data, and if the deepfake manipulation method is known in advance. However, even though such methods hold great performance on given datasets and on raw images, success rates tend to drop when they are faced with unseen, compressed or altered data, as highlighted by the FF++ [9] and DFDC [10] studies. As such, basing the detection process on raw image analysis might not keep holding up in the future, in the adverserial race between deepfake detectors and creators. Furthermore and unfortunately, efforts have seldom been made to reduce the computational costs in training high-quality deepfake detectors.

B. Semantics-based Classification

To face this growing problem, methods have been put together to circumvent the need of having to rely on ever-subtler image artifacts, and to make the decision at a higher, more semantically robust level.

Agarwal *et al.* [12] and Cozzolino *et al.* [13] combine neural networks and statistical methods to extract a vector representing the identity of the speaker, and then compare it to one from genuine footage. However, while such methods tend to overfit less on particular manipulations, they need access to supplementary information compared to traditional CNN-based methods (*e.g.* footage which is known to be genuine, or temporal data).

In the present paper, the proposed method takes inspiration from these approaches. Instead of searching for manipulation artifacts, a good understanding of the deepfake creation process may allow for more efficient and robust detection protocols.

III. METHOD

This section first explains why it could be desirable to perform deepfake classification on a low dimensionality space, and then describes how such a dimensionality reduction might be performed, with an emphasis on StyleGAN inversions.

A. Overview

Basing the discrimination on a low-dimensional space in which to represent the suspicious frame may ease deepfake classification. Intuitively, after reducing the dimension of the input, details which could mislead the model (*e.g.* background, pose, lighting, compression noise) can be erased. The model could then focus on the most important information (*e.g.* "Are the subject's expressions coherent ?"). Furthermore, while CNNs achieve great performance in computer vision tasks, they require a lot of training data. A low-dimensionality model could be lighter, cheaper and easier-to-train if not based on CNN architectures. The curse of dimensionality makes the amount of data points required in finding a discrimination frontier exponentially more complex as the number of dimensions increases. As such, one could hope that if the deepfake data is projected in a space of lower dimensionality, fewer training points would be required to build a robust model.

A good projection space in this case is an efficient representation of the faces, which would make the distinction between the distribution of real faces and manipulated ones easy. The resulting decision barrier may even be more robust to slight changes in the data distribution (from compression, different context, etc.). However we need the projector to keep as much statistics relevant to the deepfake detection process (intuitively information such as the subject's identity and expressions for instance) as possible.

Formally, the projector is a function $P : \mathbb{R}^{l \times h \times w} \to \mathbb{R}^{d'}$, with $d = l \times h \times w$ the dimension of the original high-dimensional image and $d' \ll d$ the dimension of the lower-dimensional latent representation.
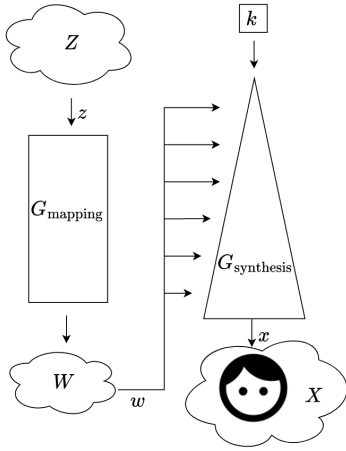
Fig. 2. The StyleGAN architecture summarised, k is a constant learned during training.

B. Dimensionality Reduction Algorithms

There are many ways to reduce data dimensionality: Principal Component Analysis (PCA) [14] can compress data while keeping as much variance as possible, Autoencoders [15] and Variational Autoencoders (VAEs) [16] are a (regularised) non-linear extension of this technique. Only the PCA method will be considered in the context of the proposed approach.

For images of faces in particular, the generative network StyleGAN [4], on top of producing images of excellent quality, has the advantage of having a unique architecture which generates data via an intermediate latent space. The structure of the generator network is summarized Figure 2. A normally sampled variable $z, z \sim \mathcal{N}(0,1)^{512}$ is first transformed to a latent code w through a mapping network $G_{\text{mapping}}(z) = w$. In turn, this variable is used at different stages of the proper image generation, done by another neural network $G_{\text{synthesis}}$.

The Latent space W defined as $w \in W, G_{\text{mapping}}(z) = w$ fits the criteria of a good support for the projected data: we already know there is a function from it to the face image space, and that it is of low dimensionality. In it, significant semantic directions have already been found [17], [18], [19], which suggest its structure is ideal to study or perform facial modifications which are easily interpretable by humans.

Multiple methods have been proposed by the community in the past years to obtain a suitable StyleGAN latent code from an image (real or generated), each with their own advantages and drawbacks.

To invert suspicious frames in the latent space of StyleGAN, the first proposed methods go through an optimization process, where the latent code is progressively modified so that the output of StyleGAN matches the desired image. Many solutions exist for this task, such as proposed originally by [4] but also by [20], [21] and [22]. Most of these solutions are based on gradient descent-algorithms to minimize a loss such as $L(w) = L_p(G_{\text{synthesis}}(w), x) + \alpha||G_{\text{synthesis}}(w) - x||_2^2$, with L_p a neural distance such as LPIPS [23]. One example of

optimization is described in Algorithm 1 with σ a measure of distance in the image space.

Algorithm 1 Optimization-based StyleGAN inversion

Require: $x \in \mathbb{R}^{1024 \times 1024 \times 3}, (\epsilon, \eta) \in \mathbb{R}^2$
  $G_{\text{synthesis}} : \mathbb{R}^{512} \to \mathbb{R}^{1024 \times 1024 \times 3}$
Ensure: $w^*, G_{\text{synthesis}}(w^*) = \hat{x}$
  $w \leftarrow w_0$
  while $\sigma(G_{\text{synthesis}}(w), x) > \epsilon$ **do**
    $L = L(G_{\text{synthesis}}(w), x)$
    $w \leftarrow w - \eta \nabla L$
  end while

The drawback of those methods is that unfortunately, the gradient descent optimization may take a few thousands steps before converging, along with problems that typically arise in optimization (*e.g.* presence of local minima, high instability). More recently, researchers developed encoder-based model which are trained to output the latent code from an image, either before performing later refinements, such as IDInvert [21], and E2Style [24] or directly, such as by Psp [25].

IV. EXPERIMENTS

A few experiments are described here to discuss the usefulness of using the latent space of StyleGAN to perform deepfake discrimination. It is first shown how this latent space is better suited than a standard Principal Components Analysis as a dimensionality reduction method in section IV-B. A comparison of different StyleGAN inversion methods is then proposed in section IV-C. A step to further reduce the data dimensionality is proposed section IV-D. Later, the full pipeline, as described Figure 1, is tested against state-of-the-art deepfake detection methods in section IV-E. Finally, the scenario of working with smaller databases is examined in section IV-F.

A. Data and Model Preparation

To conduct these experiments, three datasets have been chosen : the Deepfake Detection Challenge (DFDC) preview dataset [6], as well as CelebDF version 1 and 2 [8]. The former two datasets are of smaller scale than CelebDF v2. They were used for the experiments in sections IV-B, IV-C and IV-D for practical purposes, as well as in section IV-F to especially study the performance of the proposed method on smaller datasets. The biggest and most challenging dataset, CelebDF v2, was used for the main comparison in section IV-E.

For each dataset, about one in ten frames was uniformly sampled throughout the videos, this sampling was made to limit computing requirements, as well as to ensure a relatively diverse set of faces, not too correlated to one another. Each frame was processed according to the Flickr-Faces-HQ (FFHQ) [4] alignment protocol before having their dimensionality reduced, usually by inverting the images of the faces in the StyleGAN latent space. Training, validation and testing sets were separated at the video level, with 10% of the

Dataset	# Videos	# Frames
CelebDF v1	1203	23 902
DFDC preview	5214	178 668
CelebDF v2	6529	250 972

TABLE I
DESCRIPTION OF THE DATASETS USED

Dimensionality Reduction	Accuracy
PCA	0.78
StyleGAN Inversion	**0.88**

TABLE II
COMPARISON OF RANDOM FOREST CLASSIFIERS detection accuracy
ON LOW DIMENSIONALITY DATA

Inversion	Accuracy (%)	Throughput (s/frame)
IDInvert [21]	94.0	86.9
Psp [25]	96.31	0.0492
HFGI [30]	96.43	**0.0396**
E2Style [24]	**98.46**	0.0618

TABLE III
BINARY DEEPFAKE DETECTION ACCURACY AND SPEED DEPENDING ON
THE INVERSION METHOD.

dataset being reserved for testing. Table I shows the amount of examples in each dataset.

Aside from the Principal Components Analysis studied section IV-B, every dimensionality reduction method was taken pretrained on StyleGAN and were frozen since. This was a deliberate choice to keep the models agnostic relative to the deepfake detection method, and to ensure that the proposed method was still requiring few resources to train.

B. Dimensionality Reduction Comparison

A study is proposed here to confirm the efficacy of dimensionality reduction in the context of deepfake discrimination, and to benchmark the pseudo-inversion of StyleGAN against PCA.

To invert real world images of faces in the latent space of StyleGAN, the IDInvert [21] method is first considered. It is a well-known encoder-optimization hybrid approach which has been proposed as an early alternative to costly inversions based solely on optimization. An encoder network E is trained to produce an initial latent code of size 14×512 with domain knowledge (in this case images of faces) to improve on the inversion quality. The latent code outputted by the encoder $w_0 = E(x)$ is then optimized further to improve the results.

The network is trained to minimize a combination of distances calculated in a version of $W+$ space with 14 channels, the image space, and the feature space of the VGG network [26] $F(x)$. With $\lambda_{vgg}, \lambda_{dom}$ two weights, the resulting problem to optimize is described equation 1.

$$w^* = \operatorname*{argmin}_{w}(||x - G_{\text{synthesis}}(w)|| + \lambda_{vgg}||F(x) - F(S(w))||_2 + \lambda_{dom}||w - E(G_{\text{synthesis}}(w))||_2)$$

(1)

The IDInvert model was used pretrained on the FFHQ dataset, for a fair comparison, an Incremental PCA model [27] implemented with the scikit-learn package [28] was trained on the same dataset. This PCA model kept the 512 dimensions which explained the most variance. This number was chosen as the original StyleGAN model used a latent space of same dimensionality.

To test if there is a dimensionality reduction method better suited for discriminating deepfakes, different instances of a Random Forest classifier [29] are trained on the codes outputted by each projector. Those classifiers were composed of 1500 estimators and were trained to optimize the Gini criterion. They were implemented using scikit-learn , with other parameters left to default. The results are shown in Table II.

To classify deepfakes, the StyleGAN inversion method obtains the best performance by a considerable margin (10 percentage points).

This analysis shows that a few thousand dimensions, instead of the hundreds of thousands usually used, can contain enough information to discriminate genuine images from deepfakes to a satisfactory degree, even by models as simple as Random Forest classifiers. This motivates the progression towards a more frugal approach to deepfake classification.

C. StyleGAN Inversion Benchmark

A comparison of three of the most used StyleGAN inversion models is proposed here to determine which one is the most suitable for deepfake detection, on top of IDInvert already described in section IV-B.

- Psp [25] introduces eighteen independent modules "map2style", noted β, which transform the latent feature maps of a pre-trained ResNet R into style vectors from $W+$ such as $\forall i \in [\![1 \ ; \ 18]\!], w_i = \beta_i(R(x))$.
- E2Style [24] encodes an image to a first latent code $w_0 = E_0(x)$, which can be iteratively refined $w_t = E_t(x, G_{\text{synthesis}}(w_{t-1})) + w_{t-1}$. Each encoder E_t are trained separately in order to have the advantages of the hybrid approach without having to go through an optimization process.
- HFGI [30] is one of the most recent approach. A first latent code is encoded $w_0 = E_0(x), \hat{x}_0 = G_{\text{synthesis}}(w_0)$, a second network E_c then will transform the reconstruction error $\Delta = x - \hat{x}_0$ in another latent code. These two vectors are then merged by a module D which produces a final StyleGAN latent code $\hat{x} = G_{\text{synthesis}}(D(w_0, E_c(\Delta)))$.

For each inversion method, a simple five layer fully connected network is trained on the outputted latent codes. Both the binary discrimination accuracy of the trained models and the time it took to invert an image in the latent space are shown in table III. The inversion speed has been measured on a NVIDIA QUADRO RTX 3000 graphics card.

It is becoming clearer why dimensionality reduction can be relevant in deepfake discrimination, while earlier, optimization-based inversion methods took too long to be viable, more recent models allow for quick and efficient dimensionality reduction (as much as 2 000 times faster inversion), while improving further the quality of the results,
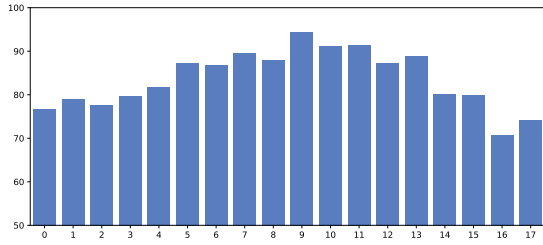
Fig. 3. Binary discrimination accuracy if only a single channel is selected

	100	50	25	10	05	01	MACS (M)
MesoNet	83.8	79.7	67.6	-	64.5	64.5	63
EfficientNet	88.0	85.0	76.8	66.7	64.4	62.4	540
XceptionNet	**93.7**	**89.7**	76.4	71.5	62.9	64.0	6010
MLP (Ours)	91.5	87.5	**83.4**	**77.7**	**70.0**	**65.0**	**2.79**

TABLE IV
COMPARISON OF BINARY DISCRIMINATION ACCURACY ACROSS VARIOUS DATA SUBSETS (% OF TRAINING EXAMPLES SEEN) MACS ARE IN MILLION OF OPERATIONS.

both visually, as shown in their respective papers, but also in deepfake discrimination.

D. Further reducing dimensionality

While the current research in StyleGAN inversion usually trends towards producing latent codes in the W+ space (dimensionality of 18×512), the original StyleGAN was supposed to work with latent codes of size 512 (the W space). Furthermore, experiments in style mixing, namely the ones performed in the original paper [4], demonstrated how the different steps in the generation of the data (which corresponds to different channels in W+ space) influenced the final results in drastically different ways. Early modification of the latent code resulted in modification of low level facial features, such as age, gender or global shapes, while later modifications influenced finer details, such as skin tone and hair color, without modifying the global shape of the face.

To see if further dimensionality reduction could be considered in the proposed method, multiple classifiers were compared. For each of the eighteen channels of the latent codes in W+ space, a fully connected neural network (five layers with ReLU activation functions) was trained on a different channel of latent codes from faces inverted in the W+ space. The data was pre-processed in the same manner as for the previous experiments. The previous database of latent codes of dimensionality 18×512, was splitted along channels into 18 different databases, on each of which a fully-connected neural network was trained. The results of those experiments are presented Figure 3.

Reassuringly, information contained in the earlier and final channels are not the most useful when it comes to deepfake detection. Models trained on the first five or latter four channels struggle to reach 80% accuracy. On the other hand, the tenth channel, controlling for mid-level features provides enough information to allow deepfake discriminators to reach more than 90% accuracy. This result is sensible, as in a finished deepfake, both low and high level features of the impersonated subject are kept (global shape of the face and hair color for example) while expressions and mid-level details are the ones being modified. This observation shows that the data dimensionality can be further reduced.

E. The Deepfake Discrimination Pipeline

Now that the utility of projecting images in the latent space of StyleGAN is confirmed, and that a few inversion methods are suitable for the task, a thorough comparison of the proposed method against state-of-the-art deepfake detectors is made. The benchmark focuses on the binary classification accuracy, the computational cost of the model, as well as the evolution of the performances depending on the number of training examples used. Three state-of-the-art models are trained from scratch on twenty epochs of the CelebDFv2 dataset: XceptionNet [3], MesoNet [2] and EfficientNet [7], as well as the proposed method (E2style dimensionality reduction and a Multilayer Perceptron with 5 fully connected layers and ReLU activation functions). XceptionNet was chosen for its well known performance in deepfake detection, EfficientNet and MesoNet because of their low computing requirements. The proposed method is trained on the tenth channel of the inverted codes in W+ space, while the CNNs are trained on images of size $256 \times 256 \times 3$.

To highlight the performances of each method on different amount of training data, discrimination accuracy has been measured on subsets of the training set. Table IV-E shows the binary discrimination accuracy on different subsets, as well as the amount of Multiply-Accumulate Operations (MACS) for each model that is trained. A column of the table displays the performance of a version of the model trained on a certain percentage of the training videos, from the full base (100%) to one percent of the training examples.

It is to be noted that the proposed method requires the inversion of the database in the latent space of StyleGAN beforehand, and as such, some extra computations must be done. However, unlike the models that are compared, the inversion models are used pre-trained on real images and would not necessitate to be re-trained every time a new deepfake manipulation method emerges. While state of the art methods achieve the best results on the full training set, their performances tend to fall when fewer training examples are available. The proposed method is competitive when all or half of the training data is available, outperforming MesoNet and EfficientNet while being behind XceptionNet by only around two percentage points. These performances are especially remarkable when the number of calculations done by the trained model are taken into account : the proposed method only computes around three million multiplications and accumulations, while XceptionNet computes around six billion of such operations. The proposed method also achieves

	DFDC-preview	CelebDF v1
EfficientNet	94.42	85.50
XceptionNet	95.95	83.55
MLP (Ours)	**98.46**	**89.35**

TABLE V

COMPARISON OF BINARY DISCRIMINATION ACCURACIES ON DIFFERENT
DATABASES

better performance as soon as around twenty-five percent of the examples are available, and keeps outperforming other state-of-the-art models when the data is even more scarce. As such, the proposed method is the better suited when a new deepfake manipulation method emerges and the amount of training data for such a manipulation is still rare. Furthermore, the lower number of operations computed by the final classifier is a necessary step towards deepfake detection democratization and ease of implementation.

F. Smaller databases

Every time a new deepfake manipulation method is proposed, state of the art models need to be trained again to provide satisfactory detection results. Creating new databases is long and expensive, as such, most of the well-used databases today were first proposed on a smaller scale (FaceForensics [31] before FaceForensics++ [9], CelebDF had a version one before being upgraded to v2, and the DFDC database [10] was first made available through a preview subset [6]). To simulate such new manipulation methods, the proposed method is tested on databases of smaller scale, CelebDF v1 and DFDC preview. The results are presented in table IV-F

When the amount of data is limited, the proposed method consistently outperforms state-of-the-art models, by as much as six percentage points. This makes it the solution of choice when dealing with a new deepfake manipulation method, or when it is unsure which kind of deepfake one is trying to detect. Especially since the proposed method mainly uses pre-trained manipulation-agnostic models (StyleGAN and E2Style), and that the models that do need to be trained for deepfake detection require much less computational power than traditional state of the art methods. Moreover, as shown in Table III, the other recent StyleGAN inversion models still achieved better accuracy than EfficientNet and XceptionNet on this dataset, further demonstrating that the power of the proposed method lies in the data format rather than the specific model. This allows the hope that as improvements in the domain are made (better inversion methods, or even better facial generation models), the proposed method would progress as well.

V. CONCLUSION

We believe the proposed method can be an efficient step towards better reactivity of deepfake detection methods in the arms race against content manipulation. Lower computing requirements and the necessity of fewer training examples for comparable or even better results than state-of-the-art models may help in having an adequate response quickly to upcoming new manipulation methods. Especially since generalization is still a burning issue in deepfake detection.

REFERENCES

[1] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Back-propagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[3] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[4] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[6] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[7] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[8] Pu Sun Yuezun Li, Xin Yang, Honggang Qi, and Siwei Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, 2020.

[9] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1–11.

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

[11] Olivia Wiles, A Koepke, and Andrew Zisserman, "Self-supervised learning of a facial attribute embedding from video," *arXiv preprint arXiv:1808.06882*, 2018.

[12] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim, "Detecting deep-fake videos from appearance and behavior," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6.

[13] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15108–15117.

[14] Karl Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[15] Mark A Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, vol. 37, no. 2, pp. 233–243, 1991.

[16] Diederik P Kingma and Max Welling, "Auto-encoding variational {Bayes}," in *Int. Conf. on Learning Representations*.

[17] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9243–9252.

[18] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt, "Pie: Portrait image embedding for semantic control," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[19] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris, "Ganspace: Discovering interpretable gan controls," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9841–9850, 2020.

[20] Rameen Abdal, Yipeng Qin, and Peter Wonka, "Image2stylegan: How to embed images into the stylegan latent space?," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4432–4441.

[21] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou, "In-domain gan inversion for real image editing," in *European conference on computer vision*. Springer, 2020, pp. 592–608.

[22] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang, "Gan inversion: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[23] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[24] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Lu Yuan, Gang Hua, and Nenghai Yu, "E2style: Improve the efficiency and effectiveness of stylegan inversion," *IEEE Transactions on Image Processing*, vol. 31, pp. 3267–3280, 2022.

[25] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296.

[26] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[27] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang, "Incremental learning for robust visual tracking," *International journal of computer vision*, vol. 77, no. 1, pp. 125–141, 2008.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[29] Leo Breiman, "Out-of-bag estimation," 1996.

[30] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen, "High-fidelity gan inversion for image attribute editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11379–11388.

[31] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.