UNTAG: LEARNING GENERIC FEATURES FOR UNSUPERVISED TYPE-AGNOSTIC DEEPFAKE DETECTION

Nesryne Mejri, Enjie Ghorbel, Djamila Aouada

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg

ABSTRACT

This paper introduces a novel framework for unsupervised type-agnostic deepfake detection called UNTAG. Existing methods are generally trained in a supervised manner at the classification level, focusing on detecting at most two types of forgeries; thus, limiting their generalization capability across different deepfake types. To handle that, we reformulate the deepfake detection problem as a one-class classification supported by a self-supervision mechanism. Our intuition is that by estimating the distribution of real data in a discriminative feature space, deepfakes can be detected as outliers regardless of their type. UNTAG involves two sequential steps. First, deep representations are learned based on a self-supervised pretext task focusing on manipulated regions. Second, a one-class classifier fitted on authentic image embeddings is used to detect deepfakes. The results reported on several datasets show the effectiveness of UNTAG and the relevance of the proposed new paradigm. The code is publicly available.

Index Terms— Type-agnostic Deepfake Detection, Unsupervised classification

1. INTRODUCTION

Deepfakes are realistic facial media that are either fully generated or partly altered using generative Neural Networks (NN). Over the last few years, remarkable advances in deepfake generation have been made, raising concerns about their misuse.

Given this threat, several deepfake detection methods have been introduced [1, 2, 3, 4]. Nevertheless, existing approaches remain hardly applicable to real-world scenarios given their lack of *inter-type* generalization. In fact, generalization can be addressed at two levels: (1) At the inter-type level, we mean robustness to unseen types of deepfakes. Possible types of deepfakes are face-swaps (FS), facial reenactments (FR), facial attribute manipulations (FAM), and fully synthetic faces (FSF); (2) At the *intra-type* level, we mean robustness to unseen forgery methods generating the same type of deepfakes. While intra-type generalization has been extensively studied [5], the topic of inter-type generalization remains less explored. Fig. 1 clarifies the distinction between intra-type and inter-type generalizations. It depicts the different types of deepfakes along with dataset examples incorporating them.

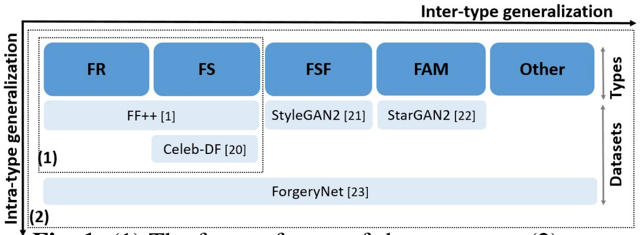

Fig. 1: (1) The focus of state-of-the-art versus (2) ours.

Earlier approaches formulate the problem of deepfake detection as an end-to-end supervised binary classification task [1, 6]. Unfortunately, such methods have shown poor *intra-type generalization* capabilities. Fully supervised NNs tend to overfit the training data inducing a drop in performance, as highlighted in [2, 7]. To overcome these limitations, some methods employed a non-contrastive self-supervision for extracting more generic features [8, 3, 9]. It consists in training a NN using an adequate data augmentation technique that mimics known artifacts. However, these methods are then fine-tuned using an annotated deepfake dataset; leading to poor inter-type generalization [4].

This paper addresses the under-explored research problem of *type-agnostic deepfake detection using unlabeled data*. As a solution, we propose to model the distribution of normal images/videos and detect deepfakes as anomalies. Such an approach also prevents the use of costly annotated data. To the best of our knowledge, unsupervised classification for deepfake detection has only been considered in [10] where a Variational Auto-Encoder (VAE) was used to learn the distribution of real data. However, while this approach can be conceptually employed for detecting any types of deepfakes, the authors do not explicitly consider more than two usual types (FS and FR). In Section 5, we show experimentally its limited inter-type generalization capabilities. Two facts might explain this. First, the generated features are not discriminative enough as the learning process is not implicitly guided to focus on specific artifact-sensitive regions. Sec-

This work is supported by the Luxembourg National Research Fund, under the BRIDGES2021/IS/16353350/FaKeDeTeR and UNFAKE, ref.16763798 projects, and by Post Luxembourg.

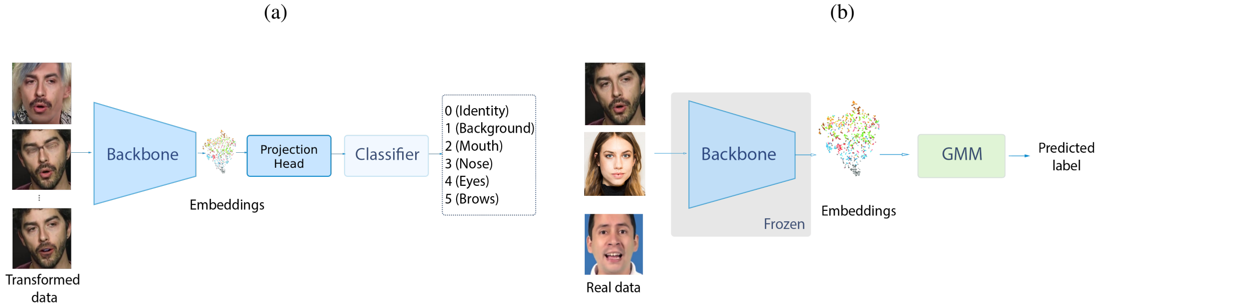

Fig. 2: Overview of UNTAG where (a) corresponds to a self-supervised step involving the prediction of spliced regions; and (b) corresponds to the estimation of a generative one-class classifier using the self-supervised features of real images.

ond, the Variational Auto-Encoder (VAE) assumes that the latent representations of real data follow a Gaussian distribution which might be too simplistic for modeling the complex distribution of real data. In this paper, we propose a novel Unsupervised Type-Agnostic deepfake detection (UNTAG) which leverages a non-contrastive self-supervision mechanism for learning generic yet discriminative features. A data augmentation technique termed R-Splicer is introduced for generating pseudo-labeled data. It augments real data by applying splicing and blending operations on regions of a given image. The selected regions are known to potentially incorporate artifacts for different types of deepfakes. Then, the augmented data are used to train a NN that detects the spliced regions. Our intuition is that by employing this self-supervision mechanism, the network will implicitly produce features that can target *artifact-sensitive* regions. Second, the feature learning step is followed by an unsupervised one-class generative classifier that estimates the probability density of real data; thus, considering only real data during training.

The paper is structured as follows: Section 2 formulates the problem. Section 3 proposes the new paradigm of type-agnostic deepfake detection using a one-class classifier. Our method, UNTAG, is detailed in Section 4. The experimental results are given in Section 5. Section 6 concludes this work.

2. PROBLEM FORMULATION

Let $\mathcal{D} = (\mathcal{I}, \mathcal{L})$ be a dataset composed of N images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ and their corresponding labels $\mathcal{L} = \{l_i\}_{i=1}^{N}$ with $l_i \in [\![0, 1]\!]$. $\mathcal{I}$ is defined by $\mathcal{I} = \mathcal{I}^R \cup \mathcal{I}^F$ where $\mathcal{I}^R$ and $\mathcal{I}^F$ are the subsets of real and fake images, respectively. For all $i \in [\![1, N]\!]$ and $\mathbf{I}_i \in \mathcal{I}$, the label $l_i = \mathbb{1}_{\mathbf{I}_i \in \mathcal{I}^F}$, with $\mathbb{1}$ being an indicator function. $\mathcal{I}^F$ is assumed to contain all types of deepfakes. The ultimate goal of deepfake detection is to find a function f such that,

$$\forall\ i \in [\![1, N]\!] \text{ and } \mathbf{I}_i \in \mathcal{I}, f(\mathbf{I}_i) = l_i. \tag{1}$$

Earlier methods mostly learn f in an end-to-end manner [1, 7], showing poor intra-type generalization capabilities. As a solution, self-supervised techniques usually decouple the learning process into two stages [3, 11], as described below

$$\forall\ i \in [\![1, N]\!] \text{ and } \mathbf{I}_i \in \tilde{\mathcal{I}},\ f_{\theta_2}(f_{\theta_1}(\mathbf{I}_i)) = l_i. \tag{2}$$

f_{θ_1} which aims at extracting rich representations is estimated by considering an auxiliary task. f_{θ_2} is then learned for discriminating between real and fake images based on the extracted representations. Typically, the estimation of f_{θ_1} involves only the set of real images $\mathcal{I}^R$. The latter is extended to a set of transformed images $\mathcal{I}^{Aug}$ associated with pseudo-labels $\mathcal{L}^{Aug}$, forming $\mathcal{D}^{Aug} = (\mathcal{I}^{Aug}, \mathcal{L}^{Aug})$, which is used to perform the auxiliary task. Hence, f_{θ_2} maps latent embeddings resulting from the auxiliary task to their corresponding labels. For the second phase, a subset denoted by $\tilde{\mathcal{D}} = (\tilde{\mathcal{I}}, \tilde{\mathcal{L}}) \subset \mathcal{D}$ is used in a supervised fashion. Note that $\tilde{\mathcal{I}} = \mathcal{I}^R \cup \mathcal{I}^{F'}$ and $I^{F'} \subset I^F$, since existing methods focus mostly on one to two types of deepfakes, e.g., face-swaps and facial reenactment. Although self-supervised mechanisms improve the *intra-type generalization* aspect, only considering a subset of fakes makes the inter-type generalization difficult.

3. A NEW PARADIGM FOR UNSUPERVISED TYPE-AGNOSTIC DEEPFAKE DETECTION

In this paper, we propose to address the problem of unsupervised type-agnostic deepfake detection. For that purpose, we propose to decouple the feature learning from the final classification as in Eq. (2). First, a self-supervised strategy tailored to the task of type-agnostic deepfake detection is leveraged for estimating f_{θ_1}. However, instead of learning a binary classifier during the second stage, the embeddings $f_{\theta_1}(\mathbf{I})$ generated from real samples $\mathbf{I} \in \mathcal{I}^R$ are assumed to follow a multivariate, Gaussian mixture distribution, such that $f_{\theta_1}(\mathbf{I}) \propto p(f_{\theta_1}(\mathbf{I})|l = 0)$ and l is the label of $\mathbf{I}$. The probability density $p(f_{\theta_1}(\mathbf{I})|l = 0)$ is defined as,

$$p(f_{\theta_1}(\mathbf{I})|l = 0) = \sum_{i=1}^{K} \Phi_i\, \mathcal{N}(f_{\theta_1}(\mathbf{I})|(\mu_i, \mathbf{\Sigma}_i), l = 0). \tag{3}$$

Note that $\sum_{i=1}^{K} \Phi_i = 1$, K is the number of Gaussian components and Φ_i is the weight of the component i. This assumption is in line with the concentration hypothesis [12] which suggests that the embeddings of real and fake data are respectively assumed to be concentrated and non concentrated in the

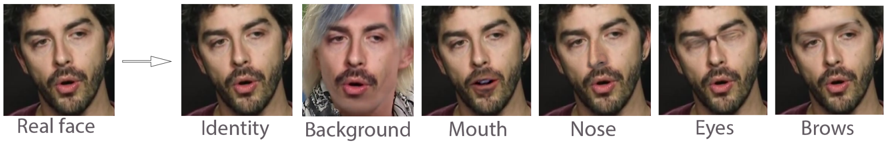

Fig. 3: The transformations generated by R-splicer given a real image

feature space. At this stage, the problem can be seen as a one-class classification, since only real images are taken into account for training. As real data is unlikely to be noise-free, we refer to this formulation as an unsupervised task. The function f_{θ_2} allows the discrimination between real and fake latent features and is computed as follows,

$$f_{\theta_2}(f_{\theta_1}(\mathbf{I}))) = 1 - \mathbb{1}_{[-L(\theta_2|f_{\theta_1}(\mathbf{I}))>\tau)]}, \tag{4}$$

where $L(\theta_2|f_{\theta_1}(\mathbf{I})) = -\mathrm{Log}(p(f_{\theta_1}(\mathbf{I})|l = 0)$ is the log-likelihood given the parameter $\theta_2 = (\mu_i, \mathbf{\Sigma}_i)_{i\in[\![1,K]\!]}$ and $\tau > 0$ is a predefined threshold.

4. UNTAG: UNSUPERVISED TYPE-AGNOSTIC DEEPFAKE DETECTION

Inspired by [13], we propose to estimate f_{θ_1} using an auxiliary task tailored for type-agnostic deepfake detection. Concretely, given a dataset $\mathcal{D}^{Aug} = (\mathcal{I}^{Aug}, \mathcal{L}^{Aug})$ of transformed images and their generated pseudo-labels, the pretext task learns in an end-to-end manner to classify which transformation was applied to an input image. Specifically, given an image $\mathbf{I}_m \in \mathcal{I}^{Aug}$ and its pseudo-label $l_m \in [\![0, k]\!]$, we estimate the correct pseudo-label such that

$$f_{\theta_3} \circ f_{\theta_1}(\mathbf{I}_m) = l_m. \tag{5}$$

$f_{\theta_1}(\mathbf{I}_m)$ denotes the features extracted by the backbone network and $f_{\theta_3}(f_{\theta_1}(\mathbf{I}_m))$ refers to the predicted pseudo-label. It is done by minimizing the following loss denoted by R_p,

$$R_p = \mathop{\mathbb{E}}_{\mathbf{I}_m\sim\pi_{\mathcal{X}}} \left[\, \mathbb{H}(l_m, L_{f_{\theta_3}\circ f_{\theta_1}}(l|\mathbf{I}_m) \,\right], \tag{6}$$

where $\pi_{\mathcal{X}}$ is the distribution of the augmented training data, $\mathbb{H}$ is the cross-entropy loss, and $L_{f_{\theta_3}\circ f_{\theta_1}}(l_m|\mathbf{I}_m) = L_{f_{\theta_3}}(l_m|f_{\theta_1}(\mathbf{I}_m))$ is the likelihood of label l_m given the image embeddings $f(\mathbf{I}_m)$. Overall, the key idea for learning discriminative features consists in applying suitable transformations to real images. Hence, we propose R-splicer as a data augmentation technique to generate $\mathcal{D}^{Aug}$.

R-Splicer. Augmenting real data by generating *pseudo-fake* images is a common practice in deepfake detection [14, 15, 16, 4, 8]. Such methods simulate characteristic face-swaps artifacts using simplistic operations [14, 15, 16, 4]. These augmentation strategies, coupled with self-supervision have significantly boosted the intra-type generalization of deepfake detectors. In particular, they mostly focus on creating synthetic blending or warping artifacts located in the boundaries of the facial area. As a result, these approaches struggle to achieve inter-type generalization as experimentally demonstrated in [4] on GAN-generated images. Hence, we argue that inter-type generalization can be enhanced by simulating artifacts not only in the facial boundaries but also in the background and in more localized facial regions. Our intuition is that each forgery type will likely introduce irregularities in different regions of images. In line with this assumption, R-Splicer applies splicing operations on a predefined set of facial and non-facial regions. In total, k ($k = 5$) regions are considered (background, mouth, nose, eyes, brows). The choice of regions is heuristically made by taking into account three elements: (1) areas in which artifacts are more likely to appear for different types of deepfakes; (2) areas with high-level semantics; and (3) simplicity of the splicing operation. A similar idea has been investigated in [8]. Nevertheless, this work differs from ours as the generated manipulations are used along with actual forged images to train a supervised binary classifier. Formally, a spliced image is defined as

$$\mathbf{I}_m = \mathbf{M}_i \odot \mathbf{I}_d + (\mathbf{J} - \mathbf{M}_i) \odot \mathbf{I}_r, \tag{7}$$

where $\mathbf{M}_i$ is a grayscale mask corresponding to the i^{th} region, $\mathbf{I}_r$ is the image to be spliced, $\mathbf{I}_d$ is the image donating its region of interest, $\mathbf{J}$ is the all-ones matrix, and $\odot$ is the element-wise multiplication. Therefore, using a set of n_r real images from $\mathcal{I}^R$, the dataset $\mathcal{D}^{Aug}$ is built by applying on each image all the predefined splicing operations denoted by $\mathcal{T} = \{\mathcal{T}_j\}_{j=0}^{k}$, where k is a heuristically chosen number of regions. Consequently, $\mathcal{I}^{Aug} = \bigcup_{i=0}^{n_r}(\bigcup_{j=0}^{k} \mathcal{T}_j(\mathbf{I}_i))$ with $\mathcal{T}_0$ being the identity transformation, i.e, $\mathcal{T}_0(\mathbf{I}) = \mathbf{I}$ for $\mathbf{I} \in \mathcal{I}^R$ and $\mathcal{T}_j$ for $j \neq 0$ a function that splices the j^{th} region and replaces it with the same region from another image. The pseudo-labels $\mathcal{L}^{Aug} = \bigcup_{i=0}^{n_r}(\bigcup_{j=0}^{k} j)$ are shown in Fig 2 a. For detecting deepfakes, a Gaussian Mixture Model (GMM) denoted by f_{θ_2} in Eq. (4) is used. After convergence, the pretext task network is frozen and is used to extract real image embeddings. These embeddings are used to estimate the parameters of a GMM. Then, at inference, the GMM discriminates between embeddings extracted from real images and non-authentic ones, as shown in Eq. (4).

5. EXPERIMENTS

5.1. Experimental protocol

Baselines. We compare UNTAG to five baselines: (1) A supervised detector called **DFD-HF** [17] (2) a self-supervised detector termed **DSP-FWA** [15]; (3) the unsupervised deepfake detection technique called **OC-FakeDect** [10] ; (4) a generic contrastive unsupervised approach entitled **SimCLR** [18]; and (5) a generic non-contrastive unsupervised method called **RotNet** [19]. Furthermore, for a fair comparison, besides training DFD-HF and DSP-FWA with their

Dataset	Supervised methods				Self-supervised methods				Unsupervised methods							
	DFD-HF [17]		DF-FH-OC		DSP-FWA [15]		DSP-FWA-OC		SimCLR [18]		RotNet [19]		OC-FakeDect [10]		UNTAG	
	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.
Celeb-DF [20]	43.12	50.70	25.40	50.00	49.47	49.50	52.61	52.60	43.06	56.22	72.05	69.75	74.10	69.95	**74.71**	**70.64**
FF++ [1]	51.21	50.75	31.06	50.00	53.65	53.36	72.00	71.79	51.44	59.72	75.28	70.71	54.16	54.27	**81.81**	**75.61**
StyleGAN2 [21]	50.66	52.35	59.87	37.36	63.57	63.10	50.93	50.62	37.97	56.46	59.26	60.87	49.84	65.82	**82.81**	**76.87**
StarGAN2 [22]	76.99	50.75	51.50	43.33	50.76	50.81	54.30	54.35	15.31	50.40	34.58	56.64	41.35	76.50	**91.14**	**87.30**
ForgeryNet [23]	43.10	50.32	37.66	49.95	51.65	51.40	57.20	57.13	54.18	57.23	51.82	53.84	63.81	60.32	**77.02**	**70.70**

Table 1: AUC and accuracy (Acc.) of UNTAG compared to the chosen baselines on different datasets.

Our pretext task	OCC	Celeb-DF [20]		FF++ [1]		StyleGAN [21]		StarGAN [22]		ForgeryNet [23]	
		AUC(%)	Acc.(%)	AUC(%)	Acc.(%)	AUC(%)	Acc.(%)	AUC(%)	Acc.(%)	AUC(%)	Acc.(%)
✓		62.51	61.17	54.53	53.75	45.64	50.28	18.17	50.10	55.29	57.73
	✓	27.43	53.55	54.85	43.21	69.75	73.99	65.46	70.42	24.82	51.20
✓	✓	**74.71**	**70.64**	**81.81**	**75.61**	**82.81**	**76.87**	**91.14**	**87.30**	**77.02**	**70.70**

Table 2: Results with and without the proposed auxiliary task and the GMM as one-class classifier (OCC).

original protocols, two variants **DFD-HF-OC** and **DSP-FWA-OC** are proposed, where the classification layers are discarded, and a GMM is fitted using real data embeddings.

Datasets. For our experiments, datasets with different types of deepfakes are considered: ForgeryNet [23] (FS, FR, FAM, FSF), FF++ [1] (FS, FR), Celeb-DF [20] (FS), StarGAN2 [22] (FAM) and StyleGAN2 [21] (FSF). ForgeryNet [23] is a recently introduced dataset. Compared to other datasets, it has the advantage to include all types of deepfakes. During testing, balanced sets of 2000 samples are built, where forged data is randomly sampled from forgery datasets, while real data is randomly sampled from the ForgeryNet validation set.

Implementation details. The regions are defined with Mediapipe landmarks[1]. R-Splicer generates from real data $20,406$ spliced images, which are used to finetune a ResNet-18 [24] in the auxiliary task. Data augmentation, such as random horizontal flipping and random grayscaling, was used. A GMM with 3 components (empirically fixed) is then fitted on real image embeddings.

5.2. Results

Comparison with the baselines. Table 1 reports the obtained results on the five considered datasets. UNTAG clearly outperforms state-of-the-art methods on all the datasets. Overall, unsupervised classification-based methods such as SimCLR [18], RotNet [19], OC-FakeDetect [10] and UNTAG are more effective for learning features that are robust to different types of forgeries. In contrast, methods that are learned in a supervised manner such as DFD-HF [17] and DSP-FWA [15] seem to be not suitable for type-agnostic deepfake detection. Despite the fact that DFD-HF [17] achieves an AUC of 91.63% using the original protocol of [17], changing the testing set impacts its performance. This suggests that the model overfits the high-frequency artifacts rather than learning type-agnostic features. Finally, the irrelevance of the features generated by DFD-HF [17] is confirmed when observing its unsupervised variant. In fact, the performance drops importantly when using DFD-HF-OC, contrary to DSP-FWA-OC, which learns from simulated warping artifacts. The results show that UNTAG also outperforms SimCLR [18], RotNet [19] and OC-FakeDect [10] regardless of the considered manipulation type. This success could be explained by the relevance of the proposed self-supervision task for deepfake detection. In fact, the self-supervision employed by RotNet [19] which is based on rotation predictions are less suitable for type-agnostic deepfake detection. Similarly, SimCLR [18] which is a contrastive approach achieves lower generalization performance than UNTAG.

Ablation Study. Table 2 reports the ablation study results. First, we consider the pretext task as a standalone classifier. To this end, the pretext task DNN is retrained as a binary classifier detecting spliced and non-spliced images and used for detecting deepfakes at inference. The results show that the network is only sensitive to face-swaps as in Celeb-DF [20], but performs poorly on GAN-generated images. Second, instead of fitting the GMM model with the real-image embeddings, we directly use the set of authentic images to estimate the GMM parameters. Results show that the GMM can distinguish between real and GAN-generated images, suggesting that these images have inherently different generation processes. Both experiments show that our pretext task and one-class classification are complementary and justify their use for type-agnostic deepfake detection.

6. CONCLUSION

This work has formulated the problem of deepfake detection as an unsupervised type-agnostic problem. A solution termed UNTAG using a one-class classifier and a self-supervision mechanism has been proposed. In particular, a novel auxiliary task tailored specifically for deepfake detection has been introduced. It aims at learning discriminative features by detecting manipulated regions with simple splicing-blending operations. Finally, a GMM is fitted on the learned representations of the real data. As a result, deepfakes can be detected as anomalies regardless of their types and without using any data annotation. Lastly, the newly formulated paradigm for type-agnostic deepfake detection is believed to be timely with a high potential to motivate the community.

[1] **Face**: 108, 68, 143, 213, 210, 208, 426, 430, 433, 372, 298, 337, **Brows**: 9, 68, 156, 124, 53, 52, 8, 282, 283, 353, 333, 298, **Eyes**: 8, 222, 224, 35, 230, 6, 450, 265, 445, 442, **Nose**: 193, 203, 164, 423, 417, **Mouth**: 164, 165, 212, 200, 432, 391

7. REFERENCES

[1] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proc. ICCV*, 2019, pp. 1–11.

[2] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, “Cnn-generated images are surprisingly easy to spot... for now,” in *Proc. CVPR*, 2020, pp. 8695–8704.

[3] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic, “Leveraging real talking faces via self-supervision for robust forgery detection,” *arXiv preprint arXiv:2201.07131*, 2022.

[4] Kaede Shiohara and Toshihiko Yamasaki, “Detecting deepfakes with self-blended images,” in *Proc. CVPR*, 2022, pp. 18720–18729.

[5] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu, “Countering malicious deepfakes: Survey, battleground, and horizon,” 2021.

[6] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez, “Deepfakes evolution: Analysis of facial regions and fake detection performance,” in *ICPR*. Springer, 2021, pp. 442–456.

[7] Nesryne Mejri, Konstantinos Papadopoulos, and Djamila Aouada, “Leveraging high-frequency components for deepfake detection,” in *IEEE Workshop on Multimedia Signal Processing*, 2021.

[8] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang, “Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection,” in *Proc. CVPR*, 2022, pp. 18710–18719.

[9] Sitong Liu, Zhichao Lian, Siqi Gu, and Liang Xiao, “Block shuffling learning for deepfake detection,” *arXiv preprint arXiv:2202.02819*, 2022.

[10] Hasam Khalid and Simon S Woo, “Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. in 2020 ieee,” in *CVPRW*, 2020, pp. 2794–2803.

[11] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li, “Deepfakeucl: Deepfake detection via unsupervised contrastive learning,” in *2021 IJCNN*. IEEE, 2021, pp. 1–8.

[12] Lukas Ruff, *Deep one-class learning: a deep learning approach to anomaly detection*, Technische Universitaet Berlin (Germany), 2021.

[13] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister, “Cutpaste: Self-supervised learning for anomaly detection and localization,” in *Proc. CVPR*, 2021, pp. 9664–9674.

[14] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, “Face x-ray for more general face forgery detection,” in *Proc. CVPR*, 2020, pp. 5001–5010.

[15] Yuezun Li and Siwei Lyu, “Exposing deepfake videos by detecting face warping artifacts,” *arXiv preprint arXiv:1811.00656*, 2018.

[16] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia, “Learning self-consistency for deepfake detection,” in *Proc. CVPR*, 2021, pp. 15023–15033.

[17] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu, “Generalizing face forgery detection with high-frequency features,” in *Proceedings of the IEEE/CVF CVPR*, 2021, pp. 16317–16326.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in *ICML*. PMLR, 2020, pp. 1597–1607.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, “Unsupervised representation learning by predicting image rotations,” *arXiv preprint arXiv:1803.07728*, 2018.

[20] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in *Proc. CVPR*, 2020, pp. 3207–3216.

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, “Analyzing and improving the image quality of stylegan,” in *Proc. CVPR*, 2020, pp. 8110–8119.

[22] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, “Stargan v2: Diverse image synthesis for multiple domains,” in *Proc. CVPR*, 2020, pp. 8188–8197.

[23] Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu, “Forgerynet: A versatile benchmark for comprehensive forgery analysis,” in *Proc. CVPR*, 2021, pp. 4360–4369.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, 2016, pp. 770–778.