Self-supervised Deepfake Detection by Discovering Artifact Discrepancies

Kai Hong[1,2], Xiaoyu Du[1,2]

[1]Nanjing University Of Science And Technology, Nanjing, 210014, China
[2]State Key Laboratory of Communication Content Cognition, Beijing, 100733, China

Abstract

Recent works demonstrate the significance of textures for the neural deepfake detection methods, yet the reason is still in explorations. In this paper, we claim that the artifact discrepancies caused by the face manipulation operations are the key difference between pristine videos and deepfakes. To imitate the discrepant situation from pristine videos, we propose an artifact-discrepant data generator to generate the negative samples by adjusting the artifacts in the facial regions with conventional processing tools. We then propose **D**eepfake **A**rtifact **D**iscrepancy **D**etector (DADD) method to discover the discrepancies. DADD adopts the multi-task architecture, associates each sub-task with a specific artifact set, and assembles all the sub-tasks for the final prediction. We term DADD as a self-supervised method since it never meets any deepfakes during the training process. The experimental results on the FaceForensics++ and Celeb-DF datasets demonstrate the effectiveness and generalizability of DADD.

Keywords

deepfake, self-supervised, artifact discrepancies

1. Introduction

Videos were a natural and convincing medium to spread information due to their abundant and strongly co-associated details, including appearances, actions, sounds, etc. This situation has changed due to the emergence of *Deepfakes*, the model-synthetic media in which the face or voice may be replaced with someone else's. The synthetic videos are resulting in negative impacts on individuals and society. Moreover, with the rapid development of generative techniques, the procedures making deepfakes have become substantially simple, while the products seem more realistic. This situation facilitates many domains, *i.e.,* the film industry, but potentially increases the probability of social issues. Therefore, the deepfake detection methods have garnered widespread attention.

Recent deepfake detection methods are mainly devised from two perspectives. The first one is used by the bio-inspired methods based on the observations and intuitive hypotheses over the datasets. Li *et al.* [1] focused on the abnormal eye blinking. Yang *et al.* [2] noted the inconsistency between the facial expressions and the corresponding head postures. Qi *et al.* [3] magnified the heart rhythm signal in videos and detected the disrupted heart rhythm signal. Li *et al.* [4] located the blending boundaries made by facial replacement methods to make the detection; the second perspective is to capture the forged features via the neural networks, including customized deep networks [5], classic neural networks [6], *et al.* . The neural methods achieve an extremely high performance [6]. But the dependence on the training datasets severely limits the model generalizability, which is very important in practical applications. For instance, the well-trained models may not work across datasets [7, 8], since the deepfakes are made by a variant of methods.

To retain the model effectiveness across the datasets, the traditional measures including data augmentation [9] and transfer learning [7] are introduced. However, these methods hardly reveal the inherent difference between pristine videos and deepfakes. To address this issue, self-supervised learning scheme is introduced to produce negative samples as the substitutes of true deepfakes to make the model learn specific features [10, 4]. The negative samples rely on the manual hypothesis of the differences between pristine videos and deepfakes, facilitating the construction of interpretable detection methods. Two typical works are FWA [10] and Face X-ray [4], where the former assumes that the artifacts are caused by the resizing and blurring operations on the facial regions, and the latter believes that deepfakes always have unseen blending boundaries. Their results demonstrate that recent neural networks mostly focus on generic visual artifacts rather than the videos themselves. Therefore, the negative samples generated with intuitive and empirical operations can facilitate the detection model and enhance the generalizability further.

In addition, many works point out that the videos and images have inherent signals like fingerprints, which are produced by the devices, the post-processing or the

* *Xiaoyu Du is the corresponding author.*

2021 International Workshop on Safety & Security of Deep Learning, August 21th, 2021, Virtual

kai_hong@njust.edu.cn (K. Hong); duxy@njust.edu.cn (X. Du)
0000-0003-3567-6396 (K. Hong); 0000-0002-4641-1994 (X. Du)

generative models [11, 12]. Inspired by these works, we make a bold hypothesis that the artifact discrepancies caused by the face manipulation operations are the key to detect deepfakes. Intuitively, all the frames in a pristine video have the same operation flow, thus they should have consistent fingerprints (*i.e.,* artifacts). In contrast, the replaced facial regions in deepfakes inevitably introduce discrepant artifacts. Focusing on the discrepant artifacts, we propose a self-supervised deepfake detection approach which comprises an **A**rtifact-**D**iscrepant **D**ata **G**enerator (ADDG) and a **D**eepfake **A**rtifact **D**iscrepancy **D**etector (DADD) to discover the discrepancy from the generated data. ADDG just uses the pristine video frames and perturbs the facial regions with the conventional processing tools, *e.g.,* blurring, scaling, rotation, replacement, etc. Although the perturbations do not change the frames in human sense, we believe that they have introduced the discrepancy in the artifact level. Thus the perturbed frames are taken as the negative samples (*i.e.,* substitutes of deepfakes) in our approach. DADD adopts the multi-task learning scheme, associates each sub-task with a type of generated data, and assembles all the sub-tasks for the final prediction. The prediction is constrained by $\ell_{2,1}$ norm[13, 14] which is a classic regularization for feature selection. The experimental results on the public datasets demonstrate that the model trained on the generated data can achieve a competitive performance, even it never sees the real deepfakes. This verifies the effectiveness and generalizability of our approach, and reveals that our hypothesis is a feasible perspective to detect the deepfakes.

The main contributions of our work are as follows:

- We hypothesize that the artifact discrepancies caused by the face manipulations are the key to detect deepfakes, thus propose a self-supervised deepfake detection approach to discover the discrepancy. The core is the Artifact-Discrepant Data Generator, which uses the pristine video frames only and perturbs the facial region with the conventional processing tools to generate the negative samples.
- To better address the artifact discrepancies, we propose Deepfake Artifact Discrepancy Detector, which adopts the multi-task learning scheme, associates each sub-task with a type of generated data, and makes the final prediction by integrating the sub-tasks. To guide the task feature selection, we adopt $\ell_{2,1}$ norm to constraint the learning process.
- Extensive experiments are conducted to demonstrate the effectiveness and generalizability of our proposed self-supervised approach, though it has never seen any real deepfakes through the training process.

2. Related Work

Bio-inspired methods. Some works have found that the actors' physiological characteristics in deepfakes are different from the real world. Li *et al.* [1] found that the actors in deepfakes have an abnormal blinking frequency and some even don't blink. Yang *et al.* [2] found that face orientation and head poses are related, but the correlation is destroyed in deepfakes. Due to the development of remote visual photoplethysmography (rppg) technology, the heart rate of actors in videos can be detected [15]. Based on this technology, Qi *et al.* [3] found the irregular heart rhythm of actors in deepfake. Similarly, Ciftci *et al.* [16] explored the biological signal difference between fake videos and real videos. However, the physiological signal artifacts reflected by different data sets are different, so specific data needs specific analysis.

Neural methods. Since deep neural networks can automatically extract images' deep features, many DNN-based detection methods have achieved satisfactory results. Zhou *et al.* [17] divided the image into different patches, and proposed a Two-stream network to detect the difference between patches. Afchar *et al.* [5] proposed a compact network structure MesoNet to detect fake videos. Nguyen *et al.* [18] proposed the use of capsule networks for deepfake detection. These methods indicate that a simple CNN network can indeed capture the relevant features of fake videos. In addition to these detection methods based on single-frame images, there are also methods based on multi-frame sequences. Guera *et al.* [19] extracted features from each frame by using CNN, then made decisions based on the feature sequence by using RNN. To capture the correlation of different frame features better, Sabir *et al.* [20] used a Bi-directional RNN. These neural methods can detect specific Deepfakes perfectly [6], but for unseen data, the detection performance will be greatly reduced [8].

Cross-data methods. Recently, the generalizability of detection methods has been emphasized. Xuan *et al.* [21] preprocessed training images to reduce obvious artifacts, forcing models to learn more intrinsic features. Cozzolino *et al.* [22] introduced an auto-encoder method that enabled real and fake images to be decoupled in latent space. Du *et al.* [7] believed that the detection model needs to focus on the forgery area, not the irrelevant ones, so they located the modified region and proposed an active learning method. Nirkin *et al.* [23] believed that the face and content of the fake image have inconsistent identity information, so they used face recognition method to detect deepfakes. However, these methods still require corresponding fake videos to complete the training, resulting in limited generalizability. Different amounts of data are bound to produce different results [24]. Another

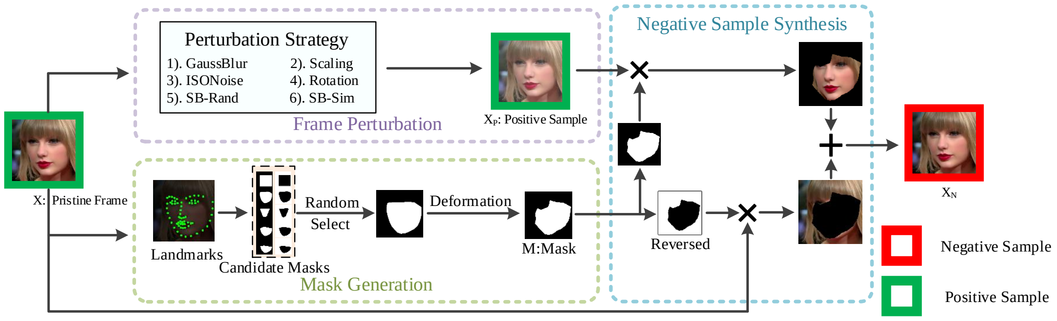

Figure 1: Overview of ADDG. Through the three modules, Frame Perturbation, Mask Generation, and Negative Sample Synthesis, the pristine frame X is converted to a negative sample X_n. The green boundary indicates that the frame should be treated as positive sample, while the red one indicates that the frame is negative, *i.e.,* it has discrepant artifacts.

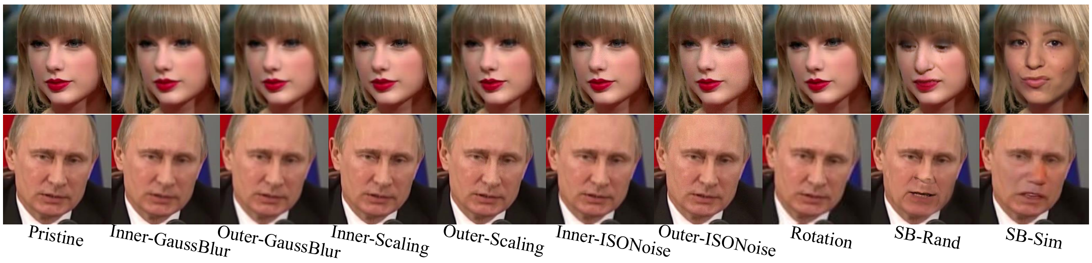

Figure 2: The perturbed examples of ADDG.

novel idea is not to use any fake image during training. FWA[10] expected to simulate face warping artifacts by adjusting the face area to different sizes and blurring it to produce similar texture artifacts. Face X-ray[4] generated images with boundary information during training dynamically. Zhao *et al.* [25] also used Face X-ray's method of generating training data and proposed a model for learning different patches' consistency. Therefore, discovering generation methods' common steps can facilitate the generalizability of the model.

3. Method

The images from different sources have different fingerprints which is caused by the devices, the post-processing operations and the generative models. The fusion of two different images leads to artifact discrepancies. This would be the key features of deepfakes, since deepfakes always have manipulated facial regions.Therefore, we propose Artifact-Discrepant Data Generator (ADDG). In order to better address the artifact discrepancy, we propose Deepfake Artifact Discrepancy Detector (DADD), which adopts the multi-task learning scheme to learn the features from each type of discrepancy data, respectively, and make the final prediction by incorporating the sub-tasks. Finally, considering that the different perturbations proposed have differing impacts, we introduce the $\ell_{2,1}$ regularization for feature selection.

3.1. Artifact-Discrepant Data Generator

As shown in Figure 1, ADDG takes in the pristine image X, and generate the negative sample (*i.e.,* artifact-discrepant sample) with three modules, frame perturbation, mask generation, and negative sample synthesis. Frame perturbation uses common image processing tools to change the fingerprint of the pristine frame like data augmentation, mask generation selects the perturbation area, and negative sample synthesis blends the pristine and perturbed frames to produce the discrepant artifacts. We will introduce the modules, respectively.

Frame Perturbation. We utilize one of the conventional image processing methods to change the fingerprint of the pristine frames, but ensure all the frames are still pristine. In this work, we use GaussBlur, Scaling, ISONoise, Rotation, SB-Rand, SB-Sim, as the

- **GaussBlur** is a commonly used data augmentation method in deepfake detection [21].
- **Scaling** refers to zooming out and then zooming in the image, which will change the texture.

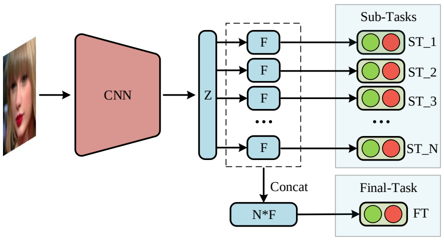

Figure 3: Overview of Deepfake Artifact Discrepancy Detector (DADD). Each sub-task is associated with a category of artifact discrepancy. The final task combines all the features from sub-tasks and do the final prediction.

- **ISONoise** is the inherent noise signal generated when the sensor captures photos, and we can obtain it by accessing the Albumentations library.
- **Rotation**'s purpose is to slightly adjust the face to produce artifacts in the form of a boundary.
- **SB-Rand** and **SB-Sim** refer to using the frames of somebody else's as the perturbation. '-Rand' indicates the frame is randomly selected, while '-Sim' indicates the frame has a similar face to the pristine frame, *i.e.,* the landmarks of the faces in these two frames are close. This operation is to introduce some more diverse texture information.

Note that all the operations in this module operate on the whole input. The generated results are pristine. Thus we denote it as X_P.

Mask Generation. This module decides the to-be-modified region in the frame. First, we locate the face landmarks in the pristine frame by using Dlib to guarantee the mask is associated with the given input. Then, by analyzing the usual operations of deepfakes, the modified area in the deepfake usually occurs on the face or the mouth, we empirically select some key points and preset five candidate masks and their reverse. The shape of the masks is presented in Figure 1. We randomly select a mask from the candidates. Because the key point detection may be inaccurate, and considering the generalization performance, we made a slight random deformation of the mask. Since the key is to the discrepancies between regions but not the perturbed facial region, we use the reversed region to perturb the corresponding background. The final mask is denoted as M, a matrix with the same shape of the input. We set two version of it.

The basic M is a 0-1 matrix, which has solid boundaries and may be easily recognized by the model. To generate hard samples, we generate M with soft boundary, where the values of M are smooth near the boundaries.

Negative Sample Synthesis. This module produces the negative samples by synthesizing the pristine and perturbed frames according the mask. Let X be the input pristine frame, X_P be the perturbed yet pristine frame, and X_N be the generated negative sample, the negative sample is generated by,

$$X_N = X_P \odot M + X \odot (1 - M), \tag{1}$$

where $\odot$ indicates the element-wise product.

Finally, we list our nine categories of negative samples. We term *Inner-* as the synthesis with common mask, that leaves the perturbation in the foreground. Similarly, we also term *Outer-* as the synthesis with reversed mask, that leaves the erturbation in the background. The categories without the previous prefixes only use the common mask, also. Specifically, the categories are Inner-GaussBlur, Outer-GaussBlur, Inner-Scaling, Outer-Scaling, Inner-ISONoise, Outer-ISONoise, Rotation, SB-Rand, and SB-Sim. Some generated examples are shown in Figure 2. Some samples showed no difference, this is due to the small degree of modification.

3.2. Deepfake Artifact Discrepancy Detector

The simplest yet effective measure to detect the artifact discrepancy is to train a model on the artifact-discrepant data. However, the fused dataset contains too many information, thus it is hard to force the model learn effective features. To address this problem, we propose Deepfake Artifact Discrepancy Detector (DADD), which adopts the multi-task learning scheme to learn the characteristics of each category, and then summarize the features to make the final prediction. The structure is shown in Figure 3.

In DADD, we first extract a common feature Z with a CNN (*e.g.,* Xception [26] in this work). Then, to suit the requirements of each sub-task, we project Z to private features of size F. The predictions of the sub-tasks are based on these private features. Then, we devise a final task based on the concatenation of all the private features.

To train DADD, we first train the sub-tasks in turn. When training sub-task ST_1, the data associated with ST_1 would be fed. We train the sub-tasks for k iterations and the train the final-tasks for t iterations, iteratively. Then the common and private features could both retain the significant features for the prediction. Eventually, in the test process, the prediction is the output of the final task.

3.3. Training

For all sub-tasks and final task, we adopt the cross entropy loss as the learning target. Let $\mathcal{L}_S$ be the sub-task loss and $\mathcal{L}_F$ be the final task loss, they are defined as,

$$\mathcal{L}_S = \mathcal{L}_F = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i), \tag{2}$$

where y_i indicates the ground truth, p_i indicates the output of the model, N indicates the number of the samples.

In addition, the purpose of DADD is to use the most suitable features. This is a feature selection task. Therefore, we introduce a feature selection regularization $\ell_{2,1}$ norm [27, 28, 29] to perform feature selection. Formally, $\ell_{2,1}$ regularization is,

$$\ell_{2,1}(W) = \|W\|_{2,1} = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{n}|W_{i,j}|^2}, \tag{3}$$

where W represents the parameter matrix, n represents the number of columns of the matrix, and d represents the number of rows of the matrix. The function of $\ell_{2,1}$ regularization is to sparse our parameter matrix's rows. In our task, each row's parameters represent the weights corresponding to the feature vectors extracted by each sub-task. We add $\ell_{2,1}$ regularization to the Final-Task training process, and the overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_F + \lambda \cdot \mathcal{L}_{2,1}, \tag{4}$$

where $\mathcal{L}_{2,1}$ indicates the regularization on the concatenated private features, and λ is a hyper-parameter. During training, we perform regular data augmentation on all types of data More detailed training procedure are listed in Algorithm 1.

Algorithm 1: Multi-Task learning Framework

```
Input: Training images $X$;
repeat
    for $i$ =0 to $k$ do
        for $n$ =1 to $N$ do
            Generate $X^{(n)}, y^{(n)}$;
            Minimize $\mathcal{L}_S(ST_n(X^{(n)}), y^{(n)})$
    for $i$ =0 to $t$ do
        Generate $X^{(1,...,N)}, y^{(1,...,N)}$;
        Minimize $\mathcal{L}(FT_n(X^{(1,...,N)}), y^{(1,,,,N)})$
until convergence;
```

4. Experiment

In this section, we conduct extensive experiments to demonstrate the effectiveness of our approach.

4.1. Experimental Setting

Datasets. To evaluate our approach, we leverage two dataset, FaceForensics++ [6] and Celeb-DF [8].

FaceForensics++ [6] comprises a set of pristine video (P) and four categories of fake videos, including DeepFakes (DF), Face2Face (FF), FaceSwap (FS) and NeuralTextures (NT) . Each category contains 1,000 videos. The dataset publisher give an official splitting list, that 720, 140, and 140 videos of each category are used for training, validation and test, respectively. In our experiments, we extract 20 frames per video. Then we adopt the training set of pristine videos only to train our model. We choose the parameters according to the validation set and evaluate the model on the test set.

Celeb-DF [8] is a challenging data set, which is mostly used for cross-dataset test. There are 38 real videos and 62 fake videos in this test set. We extract all frames from these videos. We select the model via the validation set of Faceforensics++ and evaluate our model Celeb-DF.

Note that the test data never appeared in the training datasets, especially the deepfakes. Moreover, Celeb-DF is an independent and hard dataset. Thus the test results can demonstrate the effectiveness of our generalizability across datasets.

Methods. To make fair comparison, we introduce two recent self-supervised deepfake detection methods, FWA [10] and Face X-ray [4], which also used real frames to generate training data during training dynamically. **FWA** believes that GaussBlur could construct warped faces, so they used different degrees of GaussBlur to construct negative samples. **Face X-ray** dynamically generates images with boundary information.

In our experiments, we use '-FWA' to denote the data generated by FWA, '-BI' to denote the data generated by Face X-ray, and '-ADDG' to denote the data generated by our proposed method. We also use 'Xcep-' to denote the Xception model, 'Xray-' to denote the X-ray model, and 'DADD-' to denote our method.

4.2. Performances

Table 1 demonstrates the results on DF and Celeb-DF. The results marked with references indicate that they are from the original. Two-stream [17] was trained on the SwapMe dataset [17]. Meso4 [5] was trained on an internal DeepFake dataset collected by the authors. HeadPose [2] was trained on the UADFV dataset [2]. For FWA, the dataset was collected from the Internet. The supervised methods perform badly when testing cross data. In contrast, the self-supervised methods in the second part of Table 1 mostly do well. That reveals the significance of the explorations on self-supervised methods.

Table 1
Comparison with baselines (AUC (%)). The first part is based on supervised methods, the second part is based on self-supervised methods

Method	FaceForensics++					Celeb-DF
	DF	FF	FS	NT	ALL	
Two-stream [17]	70.10	-	-	-	-	55.70
Meso4 [5]	84.70	-	-	-	-	53.60
HeadPose [2]	47.30	-	-	-	-	54.80
FWA [10]	79.20	-	-	-	-	53.80
Xcep-BI [4]	98.95	97.86	89.29	97.29	95.85	-
Xray-BI [4]	99.17	98.57	**98.21**	**98.13**	98.52	74.76
Xcep-FWA	94.09	91.89	62.55	85.78	83.58	53.76
Xcep-BI	99.52	94.76	95.95	90.64	95.22	76.36
DADD-ADDG ($\ell_{2,1}$)	**99.92**	**99.21**	97.72	97.90	**98.69**	**82.93**

Since FWA only considers the use of GaussBlur to simulate the warped face during the deepfake generation process, its generalizability is limited. As can be seen from Xcep-FWA, only DF and FF perform slightly higher. For Xcep-BI, the results are different because the specific settings of my experiment are different from the original paper. Xray-BI and our method DADD-ADDG ($\ell_{2,1}$) perform evenly on DF, FF, FS, and NT. DADD-ADDG ($\ell_{2,1}$) have an average improvement of 0.17% on FaceForensics++. But on the more difficult Celeb-DF, our method improves by 8.17%. This verifies our hypothesis on the artifact discrepancy. Since our task is to improve generalization performance, that is, test results on completely unrelated data sets, a slight decrease in performance on FS and NT is acceptable.

4.3. The Impact of Perturbations

We presents the impact of different perturbations in Figure 5. We finetune the Xception on different categories of perturbed frames, respectively.

From figure 5, we have the following observations. For DF, all methods have good performance except Outer-Scaling. For FF, the Rotation we proposed has reached the best response, indicating that FF artifacts show more edge information.For FS, the two methods that the texture of other images to perturb the original image have the best response, indicating that it is meaningful to introduce various textures. This also explains that the blending boundary constructed by rotating does not perform as well as replacing the image. For NT, Inner-GaussBlur, Inner-Scaling, and Rotation have a high response. Compared with other types of data sets, it is difficult for Celeb-DF to get a good response with a single perturbation method.

The impact of GaussBlur and Scaling are similar. When the face's interior is disturbed, it responds very well to DF, FS, and NT, while it is deficient to FS. However, when the modified area is the background, the result is the opposite. The effect will be better on FS. It verifies that our model does not merely detect specific texture features but captures the difference between internal and external textures. Since the data set is heavily compressed, a lot of information is lost. The results on the five test data demonstrate that, different perturbation would benifit different types of deepfakes.

Table 2
Ablation study (AUC (%)).

Method	FaceForensics++					Celeb-DF
	DF	FF	FS	NT	ALL	
Xcep-ADDG	**99.99**	**99.40**	98.38	97.47	**98.81**	77.60
DADD-ADDG	99.94	99.21	**98.50**	97.50	98.79	81.42
DADD-ADDG ($\ell_{2,1}$)	99.92	99.21	97.72	**97.90**	98.69	**82.93**

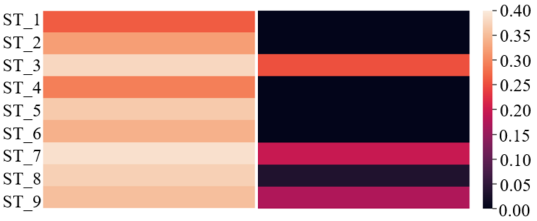

Figure 4: Visual result of feature selection implemented by $\ell_{2,1}$ regularization (λ=0.1). The left indicates that no $\ell_{2,1}$, and the right indicates that $\ell_{2,1}$ is used.

4.4. The Impact of DADD

Table 2 demonstrates the results of the methods training on the data generated by the ADDG. It is clear that the results on the four categories of FaceForensics++ are close. But the results on Celeb-DF are different. Our proposed multi-task learning framework's performance is 3.82% higher than that when using the Xception network only.

We also report the test results of each sub-task in the Figure 6. Compared with Figure 5, it is obvious that all the sub-tasks achieves a better performance. This means the multi-task scheme have improved the information in the common features. For example, the Rotation perturbation in Figure 5 is about 70% for FS and Celeb-DF, while its corresponding sub-task ST_7 in Figure 6 achieves 99% and 80% respectively. This reveals that DADD introduces significant improvements.

4.5. The Impact of $\ell_{2,1}$ Regularization

Table 2 demonstrates the impact of $\ell_{2,1}$ regularization. It improve the final result by 1.51% on cross-data Celeb-DF. This indicates that the feature selection benefit the model performances. We also test the feature selection hyper-parameter λ, and log the impact of λ in Figure 7. When $\lambda = 0.1$, the model achieves the best performance. Lower λ improve the performance in a small ratio, while

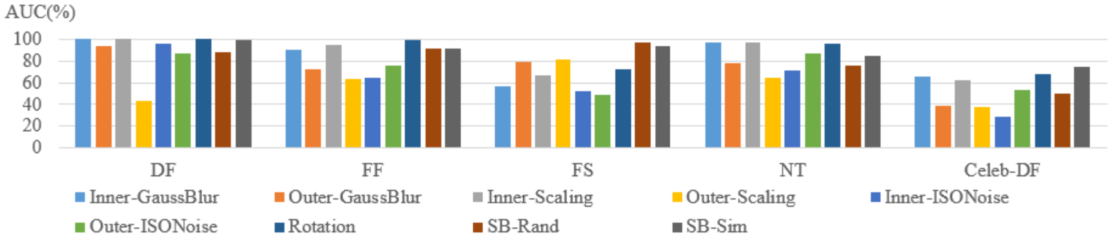

Figure 5: The results of Xceptions trained on different perturbations.

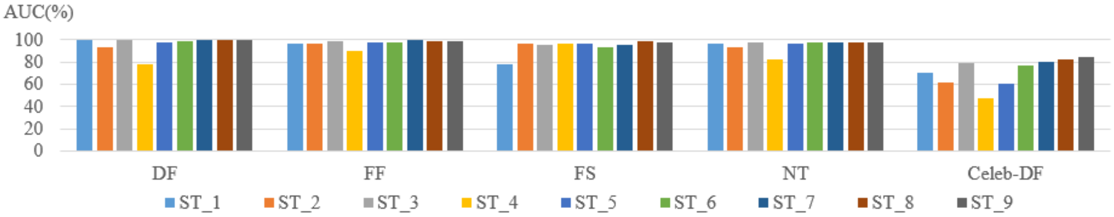

Figure 6: The predicted results of sub-tasks.

higher λ causes a sharp performance drop.

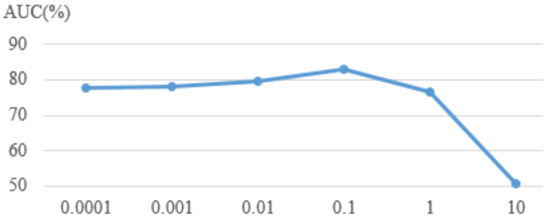

Figure 7: The average performance of different λ on Celeb-DF (AUC%).

We also visualize the layer with regularization in Figure 4. The features from ST_3, ST_7, and ST_9 contribute most. The corresponding perturbations are Inner-Scaling, Rotation, and Sim-Swap. This means they could be the delegates in the final prediction. Note that this doesn't mean only these three sub-tasks are necessary. Their performances are based on the shared features, which is learned from all the sub-tasks.

5. CONCLUSION

In this paper, we made a hypothesis that the discrepant artifacts caused by the frame manipulations are the key differences between pristine videos and deepfakes. To address the discrepancy, we proposed a self-supervised approach composed of the artifact-discrepant data generator and deepfake artifact discrepancy detector, to learn the discrepancy with pristine videos only. We conducted extensive experiments to demonstrate the effectiveness of our proposed approaches.

Deepfake detection is a special domain that aims at challenging beyond the common senses. Since the deepfakes become more realistic, the models have to pay more attention to high-frequency signals and the inherent video fingerprints. This work tried to associate the deepfake artifacts with some common noises, as a powerful tool to understand the unseen artifacts. In future, we plan to leverage this tool to explore the impact of the widely used manipulation methods. Moreover, taking this work as a reference, we are interested in extracting the key artifacts from deepfakes directly.

Acknowledgments

Special thanks are given to the SSDL2021's organizing committee and I am also very grateful to the reviewers for their valuable comments on this paper. This research was supported by Open Funding Project of the State Key Laboratory of Communication Content Cognition (No.20K03). The completion of this paper can not be separated from the Intelligent Media Analysis Group (IMAG) to help. The author would like to also thank Cheng Zhuang, Jiangnan Dai and Shaocong Yang in the IMAG for their valuable discussions.

References

[1] Y. Li, M.-C. Chang, S. Lyu, In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking, arXiv preprint arXiv:1806.02877 (2018).

[2] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP, IEEE, 2019, pp. 8261–8265.

[3] H. Qi, Q. Guo, F. Juefei-Xu, X. Xie, L. Ma, W. Feng, Y. Liu, J. Zhao, Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 4318–4327.

[4] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.

[5] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1–11.

[7] M. Du, S. Pentyala, Y. Li, X. Hu, Towards generalizable forgery detection with locality-aware autoencoder, arXiv preprint arXiv:1909.05999 (2019).

[8] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A new dataset for deepfake forensics, arXiv preprint arXiv:1909.12962 (2019).

[9] P. Chen, J. Liu, T. Liang, G. Zhou, H. Gao, J. Dai, J. Han, Fsspotter: Spotting face-swapped video by spatial and temporal clues, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.

[10] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 46–52.

[11] A. Swaminathan, M. Wu, K. R. Liu, Digital image forensics via intrinsic fingerprints, IEEE transactions on information forensics and security 3 (2008) 101–117.

[12] N. Yu, L. S. Davis, M. Fritz, Attributing fake images to gans: Learning and analyzing gan fingerprints, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7556–7566.

[13] Q. Ye, J. Yang, F. Liu, C. Zhao, N. Ye, T. Yin, L1-norm distance linear discriminant analysis based on an effective iterative algorithm, IEEE Transactions on Circuits and Systems for Video Technology 28 (2016) 114–129.

[14] Q. Ye, H. Zhao, Z. Li, X. Yang, S. Gao, T. Yin, N. Ye, L1-norm distance minimization-based fast robust twin support vector k-plane clustering, IEEE transactions on neural networks and learning systems 29 (2017) 4494–4503.

[15] Z. Yu, W. Peng, X. Li, X. Hong, G. Zhao, Remote heart rate measurement from highly compressed facial videos: an end-to-end deep learning solution with video enhancement, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 151–160.

[16] U. A. Ciftci, I. Demir, L. Yin, Fakecatcher: Detection of synthetic portrait videos using biological signals, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[17] P. Zhou, X. Han, V. I. Morariu, L. S. Davis, Two-stream neural networks for tampered face detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2017, pp. 1831–1839.

[18] H. H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2307–2311.

[19] D. Güera, E. J. Delp, Deepfake video detection using recurrent neural networks, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2018, pp. 1–6.

[20] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) 3 (2019).

[21] X. Xuan, B. Peng, W. Wang, J. Dong, On the generalization of gan image forensics, in: Chinese Conference on Biometric Recognition, Springer, 2019, pp. 134–141.

[22] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, L. Verdoliva, Forensictransfer: Weakly-supervised domain adaptation for forgery detection, arXiv preprint arXiv:1812.02510 (2018).

[23] Y. Nirkin, L. Wolf, Y. Keller, T. Hassner, Deepfake detection based on the discrepancy between the face and its context, arXiv preprint arXiv:2008.12262 (2020).

[24] H. Tang, Z. Li, Z. Peng, J. Tang, Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 610–618.

[25] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, W. Xia, Learning to recognize patch-wise consistency for deepfake detection, arXiv preprint arXiv:2012.09311 (2020).

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

[27] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, X. Zhou, l2, 1-norm regularized discriminative feature selection for unsupervised learning, in: IJCAI International Joint Conference on Artificial Intelligence, AAAI Press/International Joint Conferences on Artificial Intelligence, 2011, pp. 1589–1594.

[28] L. Fu, Z. Li, Q. Ye, H. Yin, Q. Liu, X. Chen, X. Fan, W. Yang, G. Yang, Learning robust discriminant subspace based on joint l2, p-and l2, s-norm distance metrics, IEEE Transactions on Neural Networks and Learning Systems (2020).

[29] Q. Ye, Z. Li, L. Fu, Z. Zhang, W. Yang, G. Yang, Nonpeaked discriminant analysis for data representation, IEEE transactions on neural networks and learning systems 30 (2019) 3818–3832.