2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Contrastive Knowledge Transfer for Deepfake Detection with Limited Data

Dongze Li[1,2], Wenqi Zhuo[1,2], Wei Wang[2*] and Jing Dong[2]
[1]School of Artificial Intelligence, University of Chinese Academy of Sciences
[2]Center for Research on Intelligent Perception and Computing, CASIA
{dongze.li,wenqi.zhuo}@cripac.ia.ac.cn, {wwang, jdong}@nlpr.ia.ac.cn

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956333

***Abstract*—Nowadays forensics methods have shown remarkable progress in detecting maliciously crafted fake images. However, without exception, the training process of deepfake detection models requires a large number of facial images. These models are usually unsuitable for real world applications because of their overlarge size and inferiority in speed. Thus, performing data-efficient deepfake detection is of great importance. In this paper, we propose a contrastive distillation method that maximizes the lower bound of mutual information between the teacher and the student to further improve student's accuracy in a data-limited setting. We observe that models performing deepfake detection, different from other image classification tasks, have shown high robustness when there is a drop in data amount. The proposed knowledge transfer approach is of superior performance compared with vanilla few samples training baseline and other SOTA knowledge transfer methods. We believe we are the first to perform few-sample knowledge distillation on deepfake detection.**

I. Introduction

The rapid development of deep-learning-based face editing has brought rich amusement applications in real life, yet misuse of these technologies can lead to serious moral and legal issues. The manipulated images or videos, the so-called DeepFakes, and detection on them have drawn wide attention in recent years. The mainstream deepfake detection methods can be divided into two categories, one category relies on specific clues left by generation models, and the other is data-driven and utilizes deep neural networks(DNNs) trained on real and fake face images (video frames). The high degree of similarity between real faces and fake faces makes deepfake detection a fine-grained binary classification problem, and deepfake detection models are tended to learn local manipulated traces instead of visual representation with rich semantic information like traditional image classification tasks do despite their different emphasis.

It is widely known that DNNs suffer from the following problems in realistic application scenarios. First of all, these models are large and cumbersome and are hard to be deployed onto speed sensitive terminal devices. Second, in most cases, large amount of data are required to train a DNN, which requires costly data collection and preprocessing procedure. To resolve the first issue, knowledge distillation (KD) was proposed to transfer knowledge from a strong teacher model to a light yet efficient student model. Few data or even data-free learning schemes are purposed to remedy the second issue.

Like most DNNs, deepfake detection models also have the problems mentioned above, and learning data-efficient forensics models is of great significance. However, only a few works [1], [2] have explored applying model compression to forensics models. What's more, neither few-sample learning nor data-free learning methods are suitable for these models. This is because real and fake facial images have high homogeneity. Their feature statistics stored in batch normalization layers of detection models are close to each other, and useful information will be lost in the model inversion stage of data free learning process, which could synthesize surrogate training set for the student model training. This is quite different from mainstream data-free learning methods [3], [4], [5] which can well transfer knowledge from teacher to student models. Meanwhile, forensics models make their decisions based on tiny local artifacts [6], [7], which are hard to be recovered through model inversion, and distillation with the inverted images will lead to illness solutions and poor performance of the student model.

We resort to contrastive learning to solve the above knowledge distillation problem of deepfake detection task. Specifically, we distill the forensics model with a full-dataset trained teacher and a novel contrastive loss term which maximizes the lower bound of the mutual information between the teacher and the student. A memory bank is also involved to alleviate the limitation brought by a few training samples of the student. With the help of the proposed contrastive distillation process, we can successfully get the student model which has improved against the vanilla few sample training baseline.

Our main contribution can be summarized as follow: We propose a novel contrastive distillation scheme that can improve the accuracy of a model trained with few samples by maximizing the lower bound of mutual information between the teacher and the student. As far as we know, it is the first attempt that applies few sample knowledge distillation on deepfake detection. Sufficient experiments have proved our method's superiority against baseline methods. Students trained with our method show competitive performance.

II. Related Work

Deepfake Detection. Traditional deepfake detection approaches are usually based on specific artifacts left by a certain

[*]Corresponding author.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.

forgery method. They usually lack versatility and are fragile to changes in forgery methods and data distribution.

Recent machine-learning and deep-learning methods are capable to handle more complex forgery approaches. Some of the deepfake detection methods focus on different hints left by tampered images and videos, such as abnormal person appearances, inconsistent contexts and behaviors [8], [9], [10], [11], temporal/spatial inconsistency [12] and signal level artifacts [13], [14]. Other forensics methods are data-driven [15], [16], [17] and are not paying attention to a specific trace. [18] uses SVMs and Random Forests to classify forged facial images, which is the first work to use machine-learning methods on image forensics. [19] proposes a two-stream network for face manipulation detection. MesoNet proposed by Yamagishi [20] uses a network with low layer numbers, focusing on the mesoscopic properties of images, to detect manipulated images.

Rossler [21] shows that a Xception model outperforms than other models on the tampered image detection task. In practice, EfficientNet [22] and MobileNets [23], [24] have also shown good performance with more effective inference speed.

Few Sample Learning. Learning with few samples has been extensively studied under the concept of one-shot or few-shot learning. Some literature approaches leverage generative models to estimate the whole data distribution with few samples or improve the quality of GAN generated images for training [25], [3], [26], [27]. Other works [28], [29], [30] study the problem from the view of meta learning, which follows a “learning to learn” paradigm to learn a strong meta-learner first and quickly adapt the learner for the following target tasks.

Knowledge Distillation. First proposed by Hinton [31], knowledge distillation is formed as the training process where a light student model mimics the output or the intermediate activations of one or more strong teacher models. The student is lighter and more convenient to deploy, and also shows better generalization ability. The earliest distillation method is conducted by minimizing the Kullback–Leibler divergence between the teacher and the student together with the student’s classification loss. Recent studies on knowledge distillation [32], [33] mostly focus on finding more effective features and more representative metrics that help the student contact the teacher better, and to better take advantage of the given data through learning from the data’s inner properties. Some unsupervised learning strategies such as contrastive learning are also involved in knowledge distillation [34]. Although recent years have seen remarkable progress in knowledge distillation, not much research has been conducted to apply knowledge distillation to image forensics under a few sample setting.

Contrastive Learning. Its main idea is to treat each training sample as a different category and to learn a metric space where representations of positive pairs stay close and representations of negative pairs are pushed apart. Recent researches [35], [36], [37]concentrate on learning better representations with carefully designed loss functions. In this work, we naturally introduce contrastive learning into our distillation process.

III. Method

A. Overview

Our method is based on the following phenomenon. Unlike general image classification models, which suffer from serious performance drop when number of samples in their training set declines heavily, deepfake detection models trained with a relatively small dataset do show some ability on making correct predictions with limited data, but there still exists a gap in accuracy between these models and models obtained with full dataset.

To tackle this problem, we resort to knowledge distillation which can transfer knowledge from one or more teacher models to a light student model. The student model can learn more generalized knowledge during the training process. To further make full use of the few samples we have and to help the student model to learn, we introduce a contrastive distillation loss term, which brings negative samples to support the student model to gain more information from the teacher and learn more distincted representations of the real images and the fake ones. The full pipeline can be seen in Fig. 1.

B. Problem Definition

For a given sample $x \sim p_{\text{data}}(x)$, we firstly define variables $s = f_s(x)$ and $t = f_t(x)$ for the student’s and teacher’s output representations respectively. We denote the joint probability of (s, t) as $p(t, s)$, and their marginal probabilities as p_s and p_t respectively.

Since the distillation aims to match the output probabilities of the teacher and the student models, KL divergence, also called relative entropy, is often used as the measure of how one probability distribution Q is different from a second, reference probability distribution.

$$L_{distill} = KL\left(p_t(t) \| p_s(s)\right). \tag{1}$$

Instead of KL divergence, we use mutual information (MI) as the distance of two distributions, which can be viewed as the amount of information contained in a random variable S about another random variable T, or the reduction of uncertainty in the random variable S due to the knowledge of the other random variable T.

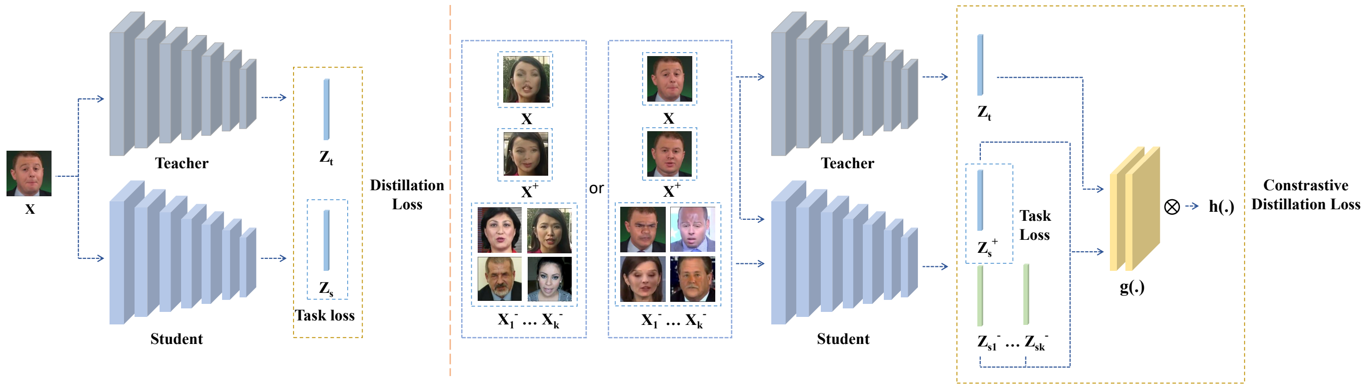
$$MI(t, s) = \sum_{s,t} p(t, s) \log\left(\frac{p(t, s)}{p(t)p(s)}\right) \tag{2}$$

Hence, for the student model optimization, our aim is sampling enough xs for training in the knowledge transfer manner.

C. Data Efficiency of Deepfake Detection

Image forensics models are usually considered fragile and easily misclassified. Surprisingly, they show high robustness on data volume, that is, when the number of training sample in the forensics dataset declines in a feasible range, the performance of the trained forensics model won’t degrade heavily. This is because samples belonging to both classes

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.



Vanilla Knowledege Distillation

Our Contrastive Knowledge Distillation Scheme

Fig. 1. From left to right, vanilla knowledge distillation versus our contrastive knowledge distillation pipeline. Different from the traditional knowledge distillation, our method brings real and fake images as positive samples in turn, and helps the student learn from the full-dataset trained teacher and few images available. Z_s^+ and Z_t are the positive representations extracted by the student and the teacher respectively. $Z_{s1} \ldots Z_{sk}$ are the representations of the negative samples extracted by the student model. $g(.)$ is a trainable linear mapping function.

have high homogeneity, both in image space and feature space. Furthermore, the forensics models concentrate more on inconsistent traces on tampered images left by generative models instead of some hierarchical features that are usually learned by general image classification models (e.g. ImageNet classification models). The accuracy of the forensics model trained on datasets that contain a different number of samples and features visualizations on forensics models are shown in Fig. 2, which supports our hypothesis. Based on the observations above, we can conclude that, to train a usable forensics model, only a small number of samples are enough, but if you want to train a model with enough accuracy, a large amount of data is necessary. Then a contrastive distillation scheme on image forensics models is proposed, which helps the model learn better representations under the supervision of the teacher which is trained on enough data, though maximizing mutual information between teacher and student.

However, for knowledge distillation with few samples, it is difficult to accurately match the probability distribution of S with that of T because of insufficient data. Fortunately, real faces in forensics tasks have high homogeneity but diversity in deepfake manners which makes us exploit a contrastive representation learning process to efficiently train the student model under a few sample setting.

The introduction of diverse negative samples for distillation can help the student model to better distinguish manipulated faces while fixing the shortcoming of incomplete distribution of training data. Note that image forensics is a binary classification problem, where real and fake samples naturally belong to the opposite class for contrastive learning. For a positive sample (real face) x^+, we collect its K corresponding negative samples (fake faces) $x^-_{\{1\ldots K\}}$ from training data. Then, we feed these samples to the teacher and student models accordingly to obtain the output representation pairs denoted as $\{(f_t(x^+), f_s(x^+)), ((f_t(x^+), f_s(x_1^-)), \cdots, ((f_t(x^+), f_s(x_K^-))\}$.

D. Contrastive Knowledge Transfer

We define a distribution q as Eq. (3), with a symbol $C = 1$ means a pair is drawn from the joint distribution, and $C = 0$ for the marginal distribution.

$$\begin{aligned} q(t, s \mid C = 1) &= p(t, s) \\ q(t, s \mid C = 0) &= p(t)p(s). \end{aligned} \tag{3}$$

Since there exists one positive pair (both real images for teacher and student models) and K negative pairs (real images for teacher and fake images for students), during the learning process, the priors of C is

$$\begin{aligned} q(C = 1) &= \frac{1}{K+1} \\ q(C = 0) &= \frac{K}{K+1}. \end{aligned} \tag{4}$$

The posterior of $C = 1$ can be derived as

$$\begin{aligned} \log q(C = 1 \mid t, s) &= \log \frac{p(t,s)}{p(t,s) + Kp(t)p(s)} \\ &= -\log\left(1 + K\frac{p(t)p(s)}{p(t,s)}\right) \\ &\leq -\log(K) + \log\frac{p(t,s)}{p(t)p(s)}, \end{aligned} \tag{5}$$

Since the mutual information between the teacher's and student's output representation can be written as Eq. (2), we can get the lower bound of $MI(t, s)$ after taking expectation on the both side of the inequality, that is

$$MI(T, S) \geq \log(k) + \mathbb{E}_{q(t,s|C=1)} \log q(C = 1 \mid t, s). \tag{6}$$

Therefore, we can relax our objective function to maximum this lower bound. As we do not know the true distribution

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.

data distribution $q(C = 1 \mid t, s)$, we try to estimate it by fitting a model $h(\cdot)$ to sample from the data distribution. For brevity, we denote the embeddings output by the teacher and the student as z_t and z_s. We use an NCE loss term as our h which can be written as

$$h(t, s) = \frac{e^{g^T(z_t) \cdot g^S(z_s)/\tau}}{e^{g^T(z_t) \cdot g^S(z_s)/\tau} + \frac{K}{M}}, \tag{7}$$

where M is the cardinality of the dataset and $g(\cdot)$ is a trainable linear transform function which maps $z_t = f_t(\cdot)$ and $z_s = f_s(\cdot)$ into the same space and performs normalization.

Therefore, the contrastive optimization objective can be written as

$$\begin{aligned} L_{distill} = & \mathbb{E}_{q(t,s|C=1)}[\log h(z_t, z_s)] \\ & + K \cdot \mathbb{E}_{q(t,s|C=0)}[\log(1 - h(z_t, z_s))]. \end{aligned} \tag{8}$$

Combining the task loss function, which is a simple cross entropy and our loss. Our final loss function can be written as

$$L_{final} = \lambda L_{distill} + L_{task}, \tag{9}$$

where λ is the weight hyperparameter.

Ensemble Knowledge Transfer In order to improve our student model on different types of forgery methods, we introduce multiple teachers to assist the student's learning process. Consider we have p models $f_{t1}, f_{t2}, f_{t3}...f_{tp}$ trained on the same subset, and the distillation loss term on ith model can been written as $L^i_{distill}$. Our ensemble distillation loss term can be formulated as

$$L_{ens} = L_{task} + \lambda \sum_{i=1}^{p} L^i_{distill}. \tag{10}$$

Memory bank Image resolution in image forensics are usually large(typically 299*299). During our training process, in order to avoid overlarge batch size, we implement a memory bank. In each iteration, after images are fed to the network, their representations will be stored in the memory bank for providing sufficient negative samples for the next training process.

Our full contrastive distillation process can be seen in algorithm 1.

Algorithm 1 Contrastive distillation for deepfake detection

```
Input: pre-trained teacher model $f_t$, randomly initialize student model $f_s$, estimation model $h$, the number of negatives samples $K$, training set $D$, memory bank $M$
Output: trained student model $f_s$
for each epoch do
    Reload $M$;
    for each iteration step do
        1. Sample training data from $D$ and get positive and negative pairs
        $\{(X, X^+), ((X, X_1^-), ((X, X_2^-)..., ((X, X_K^-)\}$;
        2. Forward and get representations
        $\{(z_{t_1}^+, z_{s_1}^+), (z_{t_1}^+, z_{s_1}^-), (z_{t_1}^+, z_{s_2}^-)..., (z_{t_1}^+, z_{s_K}^-)\}$;
        3. Calculate distillation loss Eq. (9);
        4. Backward and update student $f_s$;
        5. Push student representations into memory bank $M$;
    end for
end for
return $f_s$
```

IV. Experiments

We use vanilla training, shorten as VT, simple knowledge distillation(KD) [31], relational distillation [32](RKD) as our comparison methods. We abbreviate our contrastive distillation method as CKD. We show that on different datasets, given limited number of samples, our method shows priority by a large margin.

A. *Preliminary*

Datasets. We choose Face Forensics++ and Celeb-DF as our dataset. **Face Forensics++** [7] consists of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes [38], Face2Face [39], FaceSwap [40] and NeuralTextures [41]. **Celeb-DF** [42] is a challenging dataset which contains 5,639 high-quality Deep-Fake videos of celebrities generated using improved synthesis process. Face areas in each video in the dataset are aligned and cropped, then resized into the resolution of 299×299.

Hyperparameters. We set the hyperparameter temperature τ as 0.05 and λ as 0.1, batchsize is set to 24, and the negative sample number is set to 32 in all experiments except when the number of training samples is 10, in this situation batchsize is 10 and k also equals to 10.

B. *Validation of Data Quantity Efficiency*

We validate our data robustness hypothesis on Face Forensics++ dataset and Celeb-DF dataset. We perform vanilla training on three mainstream deepfake detectors: Xception, EfficientMet and MobileNet, on two forensics datasets: FF++ and Celeb-DF. We exponentially increase the number of images in the training set, meanwhile keeping samples belonging to different classes balanced and recording the accuracy of the model. The result can be seen in Fig. 2. We can observe that despite the rapid decline in training sample number, the decline in accuracy presents a relatively gentle trend.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.

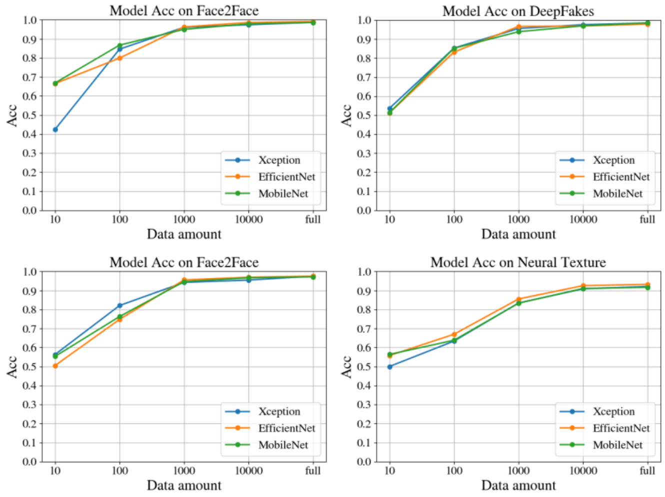


Fig. 2. Accuracy of different models on Face forensics++ dataset, vanilla training. Models trained with relatively small number of data have shown acceptable accuracy.

C. Result on Face Forensics++

The experiments are first carried out on several subdatasets in Face Forensics++. Each subdataset contains natural face images and tampered face images generated by a certain image manipulation method and is divided into non-overlap training set and testing set respectively. First, we train our teacher models with the full training set part and randomly initialize our student models. Then, a small proportion of images in the training set are set aside for training the student, our contrastive distillation scheme is applied in this stage. Then we test the accuracy of the teacher and the student on the same testing set. The number of real and fake images is equal in both the training set and the testing set.

We report the student accuracy on FF++ obtained with different numbers of training samples coming from four sub-datasets, see in Table I. The first row of the table reports the performance of teachers with the full dataset available. While the other parts of the table have shown the accuracy of student models obtained with a different training strategy with the assistance from the teachers.

Students obtained with knowledge distillation have shown significant improvement in accuracy. And students with vanilla training have shown an acceptable accuracy, which further certificates the robustness of deepfake detection models to training data number. When the volume of training set is in a reasonable range, with the supervision of the teacher model and our contrastive distillation scheme, our students have shown much higher performance than the baseline methods. It is worth noting that relational distillation [32] has appeared to be unstable, and the result students trained with relation loss added are partly better than baseline and partly not.

TABLE I

RESULT ON FACE FORENSICS++ DATASET, ACC. WE SHOW THE PERFORMANCE OF OUR STUDENT MODELS IN DETECTING FAKE IMAGES GENERATED BY SEVERAL FORGERY METHODS. THE BEST RESULTS ARE BOLDED. WITH THE NOVEL CONTRASTIVE DISTILLATION SCHEME, OUR METHOD SURPASSES BASELINE METHODS BY A LARGE MARGIN EVEN THOUGH FEW TRAINING DATA ARE AVAILABLE.

dataset volume	dataset	method	teacher-student		
			Xception Xception	Xception EfficientNet	EfficientNet MobileNet
Full	FS	VT	0.987	0.991	0.986
	F2F		0.976	0.976	0.972
	DF		0.985	0.979	0.985
	NT		0.966	0.970	0.956
1000	FS	VT	0.961	0.963	0.950
		KD	0.964	0.964	**0.953**
		RKD	0.959	0.944	0.952
		CKD	**0.979**	**0.971**	0.951
	F2F	VT	0.943	0.956	0.947
		KD	0.952	**0.962**	0.948
		RKD	0.923	0.910	0.921
		CKD	**0.964**	0.949	**0.968**
	DF	VT	0.956	0.967	0.939
		KD	0.963	0.967	0.942
		RKD	0.957	0.952	0.930
		CKD	**0.966**	**0.970**	**0.956**
	NT	VT	0.835	0.856	0.834
		KD	**0.870**	0.862	0.849
		RKD	0.850	0.834	0.839
		CKD	0.869	**0.875**	**0.859**
100	FS	VT	0.846	0.799	0.867
		KD	**0.872**	0.868	0.875
		RKD	0.850	0.828	0.849
		CKD	0.851	**0.874**	**0.890**
	F2F	VT	0.821	0.748	0.763
		KD	0.828	0.850	0.784
		RKD	0.82	0.705	0.770
		CKD	**0.839**	0.740	**0.785**
	DF	VT	0.853	0.832	0.852
		KD	0.859	0.865	0.839
		RKD	0.834	0.825	0.794
		CKD	**0.879**	**0.88**	**0.859**
	NT	VT	0.634	0.67	0.639
		KD	0.644	0.625	0.631
		RKD	0.644	0.628	0.625
		CKD	**0.678**	**0.636**	**0.665**
10	FS	VT	0.424	0.665	0.668
		KD	0.667	**0.667**	**0.725**
		RKD	0.667	0.450	0.667
		CKD	**0.667**	0.625	0.680
	F2F	VT	0.563	0.505	0.553
		KD	0.540	0.522	**0.559**
		RKD	0.625	0.438	0.495
		CKD	**0.604**	**0.625**	0.530
	DF	VT	0.536	0.512	0.513
		KD	**0.667**	**0.565**	**0.588**
		RKD	0.569	0.563	0.510
		CKD	0.540	0.535	0.512
	NT	VT	0.500	0.558	0.564
		KD	0.517	0.576	0.544
		RKD	0.554	0.505	0.566
		CKD	**0.587**	**0.510**	**0.568**

However, with only 10 samples available, the student shows the trend of collapse (nearly random guessing) and little accuracy is improved with any of the knowledge distillation method, which suggests the number of data is too small for the student model to learn the full distribution of the whole forensics dataset.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.

TABLE II
RESULT ON CELEB-DF DATASET, ACC. WE SHOW OUR RESULT ON CELEB-DF, A DATASET WITH MORE SOPHISTICATED FORGED IMAGES. HIGHER PERFORMANCE IMPROVEMENTS CAN BE OBSERVED.

dataset volume	method	teacher-student		
		Xception Xception	Xception EfficientNet	Efficient Mobilenet
full	VT	0.970	0.972	0.962
1000	VT	0.825	0.843	0.820
	KD	0.830	0.843	0.840
	RKD	0.825	0.857	0.828
	CKD	**0.866**	**0.902**	**0.852**
100	VT	0.702	0.748	0.740
	KD	0.713	0.756	0.775
	RKD	0.735	**0.804**	0.751
	CKD	**0.760**	0.798	**0.775**
10	VT	0.490	0.54	0.512
	KD	0.558	0.578	0.536
	RKD	0.573	0.562	0.601
	CKD	**0.648**	**0.612**	**0.640**

TABLE III
RESULT OF ENSEMBLE KNOWLEDGE TRANSFER ON CELEB-DF DATASET, ACC. WE SHOW THE PERFORMANCE OF A LIGHT MODEL, MOBILENET UNDER THE SUPERVISION OF ONE OR TWO TEACHERS WITH DIFFERENT ARCHITECTURES.

dataset volume	teacher-student	
	EfficientNet → MobileNet	Xception EfficientNet → MobileNet
1000	0.852	**0.865**
100	0.775	**0.794**

To better show our method's interpretability, we have also carried out a visualization experiment. Heatmaps generated with GradCAM++[43] are shown. It can be seen that the forensic models trained with our method are paying more attention to the tampered area of the fake images.

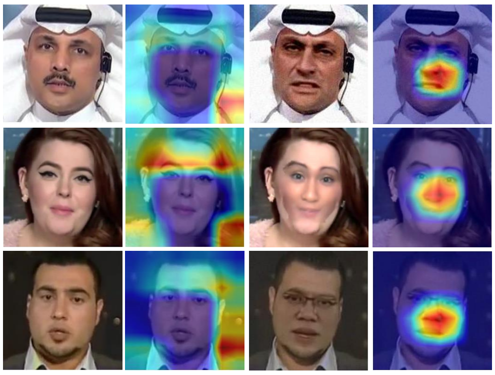


Fig. 3. Heatmaps of real and fake images on a Xception model trained with our method on Face Forensic++ dataset.

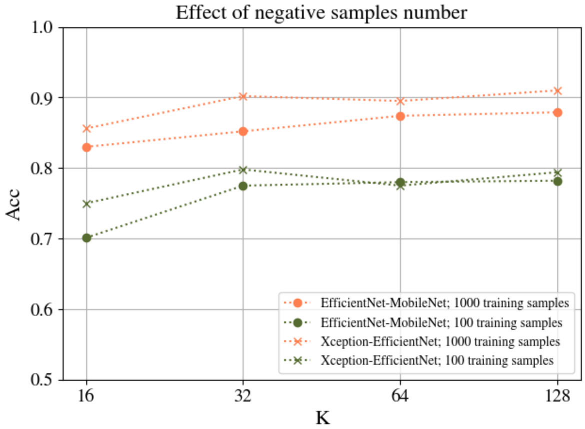


Fig. 4. Effect of negative sample number under different model architectures and different numbers of training samples.

D. Result on Celeb-DF

We further conduct our experiments on the more challenging Celeb-DF dataset. Models trained and distilled on few samples of Celeb-DF have shown relatively lower accuracy compared with those trained on a single subset of FaceForensics++ because of the more sophisticated synthetic methods. However our method is still effective on this dataset and can obtain a series of performance improvements.

E. Ensemble Knowledge Distillation

We conduct ensemble knowledge transfer experiments on Celeb-DF dataset, in which several teachers with different architectures have been introduced for training. Since the results when training sample number equals to 10 appear to be extremely poor and unstable, we only list the result of 100 and 1000 samples. Models trained with multiple teachers have shown obvious better performance then those trained with single teacher. The result can be seen in table III.

F. Effect of K

We study the effect of the number of negative samples, and the training number of samples is set to 1000 and 100 for the same reason of section IV-E. The results can be seen in Fig. 4. Blindly increasing the number of negative samples cannot bring significant performance improvement, and the accuracy when K=128 and K=32 have little gap, so we optimally set K to 32 in each experiment.

V. CONCLUSION

Model compression technologies are rarely explored in the realm of deepfake detection. To fill this gap, we show the inner robustness of deepfake detection models to the decline in the training set volume and propose a novel contrastive distillation scheme to improve the accuracy of a data-efficient light student model. Our experiments on several mainstream deepfake detection models have shown promising results, and suggest that more efforts should be devoted to exploring new training strategies for more efficient deepfake detection.

VI. ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Program of China under Grant No. 2020AAA0140003 and the National Natural Science Foundation of China (NSFC) under Grant 61972395.

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.

References

[1] M. Kim, S. Tariq, and S. S. Woo, "Fretal: Generalizing deepfake detection using knowledge distillation and representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1001–1012. 1

[2] ——, "Cored: Generalizing fake media detection with continual representation using distillation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 337–346. 1

[3] H. Chen, Y. Wang, C. Xu, Z. Yang, C. Liu, B. Shi, C. Xu, C. Xu, and Q. Tian, "Data-free learning of student networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3514–3522. 1, 2

[4] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to distill: Data-free knowledge transfer via deepinversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8715–8724. 1

[5] G. Fang, J. Song, X. Wang, C. Shen, X. Wang, and M. Song, "Contrastive model inversion for data-free knowledge distillation," *arXiv preprint arXiv:2105.08584*, 2021. 1

[6] P. Samangouei, M. Kabkab, and R. D.-G. Chellappa, "Protecting classifiers against adversarial attacks using generative models. arxiv 2018," *arXiv preprint arXiv:1805.06605*. 1

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019. 1, 4

[8] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7. 2

[9] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020, pp. 1–6. 2

[10] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 108–15 117. 2

[11] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[12] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," *arXiv preprint arXiv:2009.07480*, 2020. 2

[13] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018. 2

[14] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and 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. 2

[15] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 023–15 033. 2

[16] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2185–2194. 2

[17] Y. Zhou and S.-N. Lim, "Joint audio-visual deepfake detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 800–14 809. 2

[18] Y. Zhang, L. Zheng, and V. L. Thing, "Automated face swapping and its detection," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2017, pp. 15–19. 2

[19] P. Zhou, X. Han, V. I. Morariu, and 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. 2

[20] D. Afchar, V. Nozick, J. Yamagishi, and I. 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. 2

[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1–11. 2

[22] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019. 2

[23] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. 2

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520. 2

[25] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006. 2

[26] V. Besnier, H. Jain, A. Bursuc, M. Cord, and P. Pérez, "This dataset does not exist: training models from generated images," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5. 2

[27] L. Luo, M. Sandler, Z. Lin, A. Zhmoginov, and A. Howard, "Large-scale generative data-free distillation," *arXiv preprint arXiv:2012.05578*, 2020. 2

[28] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135. 2

[29] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018. 2

[30] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 403–412. 2

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 2, 4

[32] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967–3976. 2, 4, 5

[33] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374. 2

[34] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019. 2

[35] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 2

[36] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738. 2

[37] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 2

[38] deepfakes, "Deepfakes," github.com/deepfakes/faceswap, 2018, accessed: 2022-1-10. 4

[39] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395. 4

[40] MarekKowalski, "Faceswap," github.com/MarekKowalski/FaceSwap/, 2018, accessed: 2022-1-10. 4

[41] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019. 4

[42] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3207–3216. 4

[43] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847. 6

Authorized licensed use limited to: INSTITUTE OF AUTOMATION CAS. Downloaded on May 31,2023 at 03:33:45 UTC from IEEE Xplore. Restrictions apply.