Confidence-Calibrated Face Image Forgery Detection with Contrastive Representation Distillation

Puning Yang[1,2][0000−0002−6333−8805], Huaibo Huang[1,2][0000−0001−5866−2283], Zhiyong Wang[3][0000−0002−8043−0312], Aijing Yu[1,2][0000−0002−4782−9858], and Ran He[1,2][0000−0002−3807−991X]

[1] Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, University of Sydney, Australia
{puning.yang, huaibo.huang, aijing.yu}@cripac.ia.ac.cn, zhiyong.wang@sydney.edu.au, rhe@nlpr.ia.ac.cn

Abstract. Face forgery detection has been increasingly investigated due to the great success of various deepfake techniques. While most existing face forgery detection methods have achieved excellent results on the test split of the same dataset or the same type of manipulations, they often do not work well on unseen datasets or unseen manipulations due to the issue of model generalization. Therefore, in this paper, we propose a novel contrastive distillation calibration (CDC) framework, which distills the contrastive representations with confidence calibraion to address this generalization issue. Different from previous methods that equally treat the two forgery types, Face Swapping and Face Reenactment, we devise a dual-teacher module where the knowledge is separately learned for each forgery type. A contrastive representation learning strategy is further presented to enhance the representations of diverse forgery artifacts. To prevent the proposed model from being overconfident, we propose a novel Kullback-Leibler divergence loss with dynamic weights to moderate the dual-teacher's outputs. In addition, we introduce label smoothing to calibrate the model confidence with the target outputs. Extensive experiments on three popular datasets show that our proposed method achieves the state-of-the-art performance for cross-dataset face forgery detection.

Keywords: Deepfake Detection · Confidence Calibration · Knowledge Distillation.

1 Introduction

Recent years have witnessed the rapid development of various deepfake techniques, such as face swapping and face reenactment [27, 46–48]. As a result, face

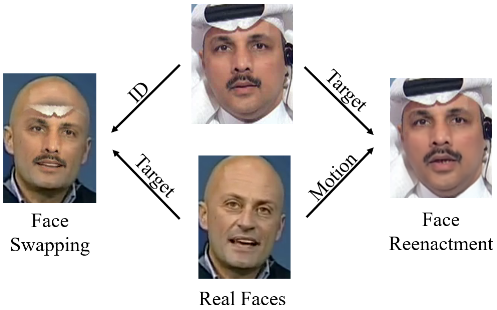

Fig. 1. Forgery samples from FaceForensic++ (FF++) [40]. Face Swapping wants to change the identity and keep the motion of the target face. On the contrary, Face Reenactment wants to preserve the identity and change the motion of the target face. Different forgery artifacts are generated during the different forgery processes. Based on this observation, we propose a feature-augmented contrastive representation approach.

forgery detection has been increasingly investigated to address the societal concerns on identity fraud, and many face forgery detection methods have been developed with promising progress[15, 55, 56].

Existing methods generally learn representations from spatial, temporal, and frequency domains to distinguish forged faces from genuine ones. Early studies focused on detecting face forgery in a single domain[38, 37]. Although these methods achieved excellent performance on seen datasets and manipulations, they lack generalization to unseen datasets and manipulations. Then multi-domain methods[35, 33] were proposed to improve generalization by fusing the features of multiple domains. However, the redundant representations introduced by multi-domain features can easily lead to over-fitting. As a result, additional constraints on inter-domain independence are required, which eventually increases computational costs. Recently, deep learning based methods[1, 40] have been investigated to learn features common to different types of forgeries. However, none of the existing techniques consider the forgery artifact differences between different types of forgery. As illustrated in Fig. 1, two types of forgeries, Face Swapping and Face Reenactment, have distinct characteristics due to different objectives, which need to be exploited separately.

Therefore, in this paper, we propose a novel contrastive representation distillation model to address the issue of model generalization. Specifically, we devise a dual-teacher module in which each teacher is trained with one of two forgery types. That is, a Face Swap teacher and a Face Reenactment teacher are trained separately to obtain discriminative features for individual forgeries. Then, the knowledge from both teachers will be jointly distilled to a student model for improved model generalization.

To further improve model generalization, we refine the task of face forgery detection from binary classification (i.e., fake vs real) to fine-grained confidence scores in the range of $[0, 1]$. To this end, we propose to calibrate the model confidence with the scores. Therefore, we further devise a new loss function, namely Confidence Calibration Loss, to calibrate the output of our framework. We exploit label smoothing to quantify the target output at finer levels. The distribution of targets is changed into a uniform distribution between the ground-truth labels and the labels processed by label smoothing regularization. The new targets will guide the model to make more fine-grained predictions to alleviate the issue of being overconfident. In addition, according to the outputs from each teacher, two dynamic confidence weights are assigned to each input sample. A loss with dynamic weights expresses the prediction from two teachers to the current sample more accurately. Therefore, we name our proposed framework as *Contrastive Distillation Calibration* (CDC). Overall, the key contributions of this paper are summarized as follows:

- We propose a novel contrastive distillation calibration framework to enhance the generalization of face forgery detection by designing a dual-teacher module for knowledge distillation and utilizing contrastive representation learning without additional disentangling.
- We propose to calibrate the model confidence with the targets for the first time for face forgery detection and devise a new Kullback-Leibler divergence based loss function with label smoothing strategy.
- We perform comprehensive experiments on three widely used benchmark datasets: FaceForensics++[40], Celeb-DF[31], and DFDC[8] to demonstrate that our proposed method outperforms the state-of-the-art face forgery detection methods on unseen datasets.

Code is publicly available at: `https://github.com/Puning97/CDC_face_forgery_detection`

2 Related Work

2.1 Face forgery Detection

Most of the existing face forgery detection methods identify forgery traces from the perspective of spatial[28, 53, 37, 40], frequency[33, 38], and temporal[35, 56, 29] domains. For example, Face x-ray[28] mainly paid attention to the mixing step existing in most face forgery cases and achieved state-of-the-art performance from the generalization perspective on raw videos. F^3-Net[38] exploited Discrete Cosine Transform (DCT) coefficients to extract the frequency features and achieved state-of-the-art performance on highly compressed videos. Various methods have also been proposed to exploit specific temporal incoherence in the temporal domain, such as eye blinking[29], lips motion[15], or expression[35].

However, these methods generally focus on learning low-level features from given datasets with a specific type of forgery, which could not be generalized across different datasets and different types of forgeries. In our work, we propose a dual-teacher module to enhance the learning of generalized representations with contrastive learning.

2.2 Knowledge Distillation

Knowledge distillation refers to a learning process of distilling knowledge from a big teacher model to a small student model. It has a history of more than ten years. Busira *et al.* [3] presented a method to compress a set of models into a single model without significant accuracy loss. Ba and Caruana [2] extended this idea to deep learning by using the logits of the teacher model. Hinton *et al.* [21] revived this idea under the name of knowledge distillation that distills class probabilities by minimizing the Kullback-Leibler (KL)-divergence between the softmax outputs of the teacher and student. In addition, the hidden representation from the teacher has been proved to hold additional knowledge that can contribute to improving the student's performance. Especially in computer vision, some recent knowledge distillation methods[39, 24, 54, 42] were proposed to minimize the mean squared error (MSE) between the representation-level knowledge of two models. They addressed how to better extract more useful knowledge from the teacher model and transfer it to the student. In the case of vision tasks, the most common methods[14, 20, 10] of knowledge distillation focus on combining the ground truth and the teacher's predictions as the overall targets to train students.

On the one hand, we continue the idea of most knowledge distillation models: distilling the knowledge of the teacher model into the student model. On the other hand, existing distillation methods generally aim to transfer features from a big model to a small one. Such an operation often leads to an issue that the student lacks prominent local advantages and the global representation ability is weaker than their teachers. Therefore, we first train two small teachers with local salient feature representation, and then distill these two small teachers' knowledge into a big student model. Theoretically, the large student preserves the representations of both teachers and improve model generalization through the transfer process.

2.3 Confidence Calibration

Modifying model improves its robustness and generalization [17, 18]. Confidence calibration is effective in enhancing the generalization of a model and improving its reliability to be deployed in realistic scenarios[43, 13]. As there exists an overfitting issue for deep neural networks, Guo *et al.*[13] explained that existing neural networks often make overconfident predictions and proposed the concept of confidence calibration. Hein *et al.*[19] suggested that neural networks using the ReLU activation function are essentially piecewise linear functions, thus explaining why out-of-distribution data can easily cause softmax classifiers to generate highly confident but incorrect outputs: piecewise linear functions imply that the methods which operate on the output of classifiers cannot recognize an input as out-of-distribution inputs. Confidence calibration improves the generalization of a model from this perspective.

Many confidence calibration methods have been proposed in recent years, including temperature scaling[32, 22], mixup[49], label smoothing[44], Monte Carlo

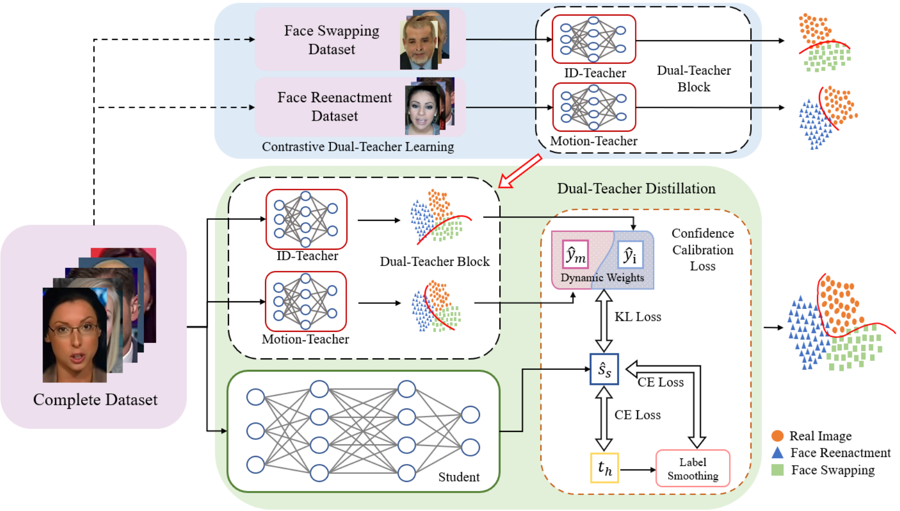

Fig. 2. Overview of our proposed CDC framework, we firstly perform the Contrastive Dual-teacher Learning to train the two teachers. Then, we undertake contrastive representation-based distillation with confidence calibration. Note that abbreviations have been used for some symbols. For instance, $\hat{y}_m = \hat{y}_{motion}$, $t_h = t_{hard}$, $\hat{y}_i = \hat{y}_{id}$, $\hat{s}_s = \hat{s}_{stu}$.

Dropout[12], and Deep Ensembles[26]. Label smoothing is a method of fusing the ground truth with a uniform distribution. It enforces a model to generate a smoother probability distribution to fit the soft targets, which is utilized in our distillation framework. In addition to label smoothing, we also propose to exploit dynamic confidence weights to better leverage the knowledge of the two teachers.

3 Methodology

As shown in Fig. 2, our proposed forgery detection method consists of three key components: a dual-teacher block, a student block, and a confidence calibration block. Firstly, we divide the FF++ dataset into two sub-datasets: Face Swapping Dataset (FSD) containing face swapped images and their corresponding genuine images and Face Reenactment Dataset (FRD) containing motion reenacted images and their corresponding genuine images, and obtain an id-teacher and a motion-teacher by training our backbone on FSD and FRD, respectively. Secondly, the knowledge of both teachers in the dual-teacher block is transferred to the student by training it on a multi-category forgery dataset. Finally, to achieve fine-grained classification of multiple types of forgeries, we devise confidence calibration-based loss functions with dynamic weights and introduce label smoothing to better supervise network training.

Teachers	Face Swapping	Face Reenactment
ID-Teacher	99.53	58.12
Motion-Teacher	76.58	99.25

Table 1. Cross-evaluation results of the Dual-teacher block in terms of AUC (%) on FSD and FRD.

3.1 Feature Representation

Existing forgery methods include two modes: face swapping and face reenactment. Face swapping aims to exchange identity information while keeping the motion information maximally. On the contrary, face reenactment aims to exchange the motion information and keep the identity information. Inspired by the principle of contrastive learning, we divide the FF++ dataset into two sub-datasets: Face Swapping Dataset, which includes face swapped forgeries and corresponding genuine images, and Face Reenactment Dataset, which provides face reenacted forgeries and corresponding genuine images. Two backbones are trained on these two sub-datasets, respectively. Now we obtain two teachers: ID-Teacher and Motion-Teacher. A cross-evaluation experiment was undertaken with results shown in Tab. 1, to verify the independence and validity between the ID-Teacher and the Motion-Teacher. The difference in cross-category evaluation indicates existing face reenactment forgery methods keep the identity information well. However, face swapping methods do not keep the motion information well.

We introduce the details of our backbone, which consists of two parts: the EfficientNet-B3[45] and the Feature Transformer. These two parts are trained in an end-to-end manner for teacher training. Overall, given a suspect image $I \in \mathbb{R}^{H\times W\times 3}$, the first stage is the EfficientNet-B3[45] pre-trained from TIMM[52]. It extracts identity or motion feature $F = \Phi_{EfficientNet-B3}(I)$, $F \in \mathbb{R}^{H'\times W'\times C}$.

The second stage is Feature Transformer based on Vision Transformer[9]. With the EfficientNet-B3, we obtain the features $F \in \mathbb{R}^{H'\times W'\times C}$. Note that a global feature F can be represented as a sequence of local features $F_l \in \mathbb{R}^{H'\times W'}, l \in 1, 2, ..., C$, which is a 2D sequence of token embeddings in the standard Vision Transformer[9]. Note that C denotes the input sequence length, H and W are as same as the input patch size p. Following the settings in ViT[9], we flatten the patches and map them to D dimensions with a trainable linear projection (Eq.1). We add a learnable class embedding to the sequence of embedded patches ($z_0^0 = F_{class}$), learnable 1D position embeddings (E_{pos}) also have been included to retain positional information.

$$z_0 = [F_{class}; F_1E; F_2E; \cdots ; F_CE] + E_{pos}, \tag{1}$$

Where F_n denotes the n-th part in feature F, $E \in \mathbb{R}^{(H'\cdot W')\times D}$, $E_{pos} \in \mathbb{R}^{(C+1)\times D}$.

The core block of Feature Transformer is L standard Transformer Encoder blocks and each block consists of a multi-head self-attention(MSA) (Eq.2) block and an MLP block (Eq.3). Layernorm (LN)[50] is applied before every block and residual connections after every block. The MLP contains two layers with a GELU activation function:

$$z'_\ell = \mathrm{MSA}(\mathrm{LN}(z_{\ell-1})) + z_{\ell-1}, \ell = 1 \cdots L, \tag{2}$$

$$z_\ell = \mathrm{MLP}(\mathrm{LN}(z'_\ell)) + z'_\ell, \ell = 1 \cdots L. \tag{3}$$

The output of the Transformer Encoder (z_L^0) is the representation of the image. We can apply an MLP head for the final prediction:

$$y = \mathrm{MLP}(\mathrm{LN}(z_L^0)). \tag{4}$$

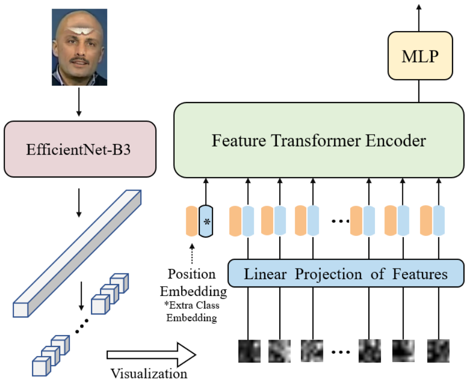

Fig. 3. Illustration of the backbone architecture which consists of EfficientNet-B3 and a variant of Vision Transformer. We split the features extracted from EfficientNet-B3 along the channel dimension and feed the sequences of features to a standard Transformer Encoder.

3.2 Dual-teacher Knowledge Distillation (DKD)

Now we have two teachers to form the dual-teacher block. Specifically, the ID-Teacher maps face swapped samples I_{id} to the output y_{id}, and the Motion-Teacher maps face reenacted samples I_{motion} to the output y_{motion}. Our goal is to teach a student that inherits advantages from both teachers by distilling the knowledge from the dual-teacher block. Considering that the scales of the data between teachers and students are different, we fine-tuned the backbone structure. The difference between teachers and student is the number of layers and heads in the Feature Transformer encoder. A given image I_{train} is fed to three branches: the ID-Teacher, the Motion-Teacher, and the Student. On the one hand, we obtain the predictions $\hat{s}$ from the Student and the hard targets

t_{hard} from the groundtruth of the training dataset. On the other hand, the ID-Teacher and the Motion-Teacher respectively predict a score of the input, then we get $\hat{y}_{id}$ and $\hat{y}_{motion}$. We use the binary cross-entropy (BCE) loss to supervise the prediction on hard targets (Eq.5). The learning process between teachers and student is supervised by the Kullback-Leibler (KL) divergence loss (Eq.6).

$$\mathcal{L}_{hard} = \frac{1}{N}\sum_{1}^{N}\mathrm{BCE}(\hat{\mathrm{s}}_{\mathrm{stu}}, \hat{\mathrm{t}}_{\mathrm{hard}}), \tag{5}$$

$$\mathcal{L}_{teacher} = \frac{1}{N}\sum_{1}^{N}(\mathrm{KL}(\hat{s}_{stu}, \hat{y}_{id}) + \mathrm{KL}(\hat{s}_{stu}, \hat{y}_{motion})). \tag{6}$$

The overall loss function of our distillation model is as follows:

$$\mathcal{L}_{distill} = \lambda\mathcal{L}_{hard} + \lambda\mathcal{L}_{teacher}. \tag{7}$$

3.3 Confidence Calibration Loss (CCL)

As shown in Fig. 4, our model also suffers from overconfidence. It is necessary to calibrate the confidence of our framework. We present two sub-functions to achieve this goal: Dynamic Confidence Weights and Label Smoothing Regularization.

Dynamic Confidence Weights (DCW) Given a training sample, we usually get different results from the two teachers, which means different levels of confidence. Therefore we propose the Dynamic Confidence Weight strategy, which is based on the outputs from the teachers. For instance, one teacher's output of a training sample is $\hat{y}_s$, the probability on the other category is $1 - \hat{y}_s$. We define the absolute value (Eq.8) between these two probability scores as the model confidence index, which represents the teacher's confidence in its prediction.

$$\lambda_s = |1 - (\hat{y}_s) - \hat{y}_s| = |2\hat{y}_s - 1|. \tag{8}$$

For each sample, we get two dynamic confidence weights λ_{id} and λ_{motion}:

$$\lambda_{id} = |2\hat{y}_{id} - 1|, \lambda_{motion} = |2\hat{y}_{motion} - 1|. \tag{9}$$

we can calculate the loss of samples:

$$\mathcal{L}_{id} = \frac{1}{N}\sum_{1}^{N}(\lambda_{id}KL(\hat{s}_{stu}, \hat{y}_{id})), \tag{10}$$

$$\mathcal{L}_{motion} = \frac{1}{N}\sum_{1}^{N}(\lambda_{motion}\mathrm{KL}(\hat{s}_{stu}, \hat{y}_{motion})), \tag{11}$$

and the final loss function can be written as:

$$\mathcal{L}'_{teacher} = \mathcal{L}_{id} + \mathcal{L}_{motion}. \tag{12}$$

Label Smoothing Regularization(LSR) To further improve the generalization ability, we exploit the label smoothing method in our task. Based on our observations, forged images often contain some genuine content. For instance, both face swapped samples and face reenacted samples have common real contents: the background in the images. Besides, face swapped samples also have part of the same motion information as the genuine samples. Coincidentally, face reenacted samples have the same identity information as the genuine samples. Thus, it is reasonable to set the targets from the combination between the ground-truth labels and its converted outputs. Considering a smoothing parameter ϵ, a sample of the ground-truth label y, we replace the label distribution:

$$P_i = \begin{cases} 1, if(i = y) \\ 0, if(i \neq y) \end{cases} \Rightarrow P_i = \begin{cases} (1-\epsilon), if(i = y) \\ \epsilon, if(i \neq y) \end{cases} . \tag{13}$$

Now we get a new label distribution, which is a mixture of the original ground-truth distribution and the converted distribution with weights $1 - \epsilon$ and ϵ. We practically interpret LSR with the cross entropy:

$$\mathrm{LS}(\hat{p}, t) = \begin{cases} (1-\epsilon) * (-\sum_{i=0}^{1} t_i \log \hat{p}_i), if(i = y) \\ \epsilon * (-\sum_{i=0}^{1} t_i \log \hat{p}_i), if(i \neq y) \end{cases} . \tag{14}$$

The final label smoothing loss is:

$$\mathcal{L}_{smoothing} = \frac{1}{N} \sum_{1}^{N} \mathrm{LS}(\hat{\mathrm{s}}_{\mathrm{stu}}.\mathrm{t}_{\mathrm{hard}}). \tag{15}$$

Finally, our total loss function is as follows:

$$\mathcal{L}_{total} = (1 - \lambda_t)\mathcal{L}_{hard} + \lambda_t \mathcal{L}'_{teacher} + \lambda_l \mathcal{L}_{smoothing}. \tag{16}$$

4 Experimental Results and Discussions

We evaluated the performance of our proposed CDC (i.e., DKD + CCL) against multiple state-of-the-art methods on three publicly available datasets. We show that our model achieves convincing performance under the in-dataset setting. To demonstrate the robust generalization ability of our model, we conducted the cross-dataset evaluation by training the model with only FF++[40] datasets and testing on unseen datasets. Ablation studies explore the contribution of each component in our framework, such as the impact of DKD and CCL.

4.1 Experimental Settings

Datasets. Following recent related works on face forgery detection [33, 55, 1, 30], we conducted our experiments on the three benchmark public deepfake datasets: FaceForensics++ (FF++)[40], Celeb-DF[31], and Deepfake Detection Challenge

(DFDC)[8]. FaceForensics++ (FF++) is the most widely used dataset in many deepfake detection approaches. It contains 1,000 original real videos from the Internet, and each real video corresponds to 4 forgery ones, which are manipulated by Deepfakes[11], NeuralTextures[47], FaceSwap[25], and Face2Face[48], respectively. Celeb-DF consists of high-quality forged celebrity videos using an advanced synthesis process. Deepfake Detection Challenge (DFDC) public test set was released for the Deepfake Detection Challenge, which contains many low-quality videos and makes it exceptionally challenging.

Evaluation Metrics. We utilized the Accuracy rate (ACC) and the Area Under Receiver Characteristic Curve (AUC) as our evaluation metrics. (1) ACC. The accuracy rate is the most popular metric in the classification task. It is also applied to evaluate the performance of face forgery detection, and we used ACC as one of the evaluation metrics in the experiment. (2) AUC. Following the Celeb-DF[31] and DFDC[8], we used AUC as the other evaluation metric to evaluate the performance in the cross-dataset evaluation.

Pre-processing and Training Setting. For all video frames, we used Retinaface[7] to detect faces and saved the aligned facial images as inputs with a size of 224×224. We augmented our training data using Albumentations[4]. We set hyper-parameters $\lambda_t = 0.15$, $\lambda_l = 0.1$ in the Equation 16. Optimizer was set to Adam[23] for end-to-end training of the complete model with a learning rate of $1e-4$ and decay of $1e-6$. We trained our models with a batch size of 32 for 100 epochs. All our results were obtained on four NVIDIA RTX3090 GPUs.

4.2 In-dataset Evaluation

Although our framework focuses on the generalization ability for the face forgery detection task, it still achieves competitive results in the in-dataset evaluation with FF++[40] and DFDC[8]. Given a dataset, our model is trained on both genuine and deepfake data from train split, and its performance is evaluated with the corresponding test split. Different from existing works, we compare the performance in two sub-datasets: the Face Swapping Dataset and the Face Reenactment Dataset. As shown in Tab. 2, 3, and Fig. 4, our framework is on par with existing state-of-the-art methods. Selim [41] achieved a better performance in terms of AUC than our framework because it was devised for the setting of the

Method	Face Swapping		Face Reenactment	
	ACC	AUC	ACC	AUC
Face X-ray[28]	98.97	99.18	**98.73**	**98.99**
I3D[5]	97.85	98.32	94.65	95.17
MIL[51]	97.54	97.21	98.19	98.27
Ours	**99.45**	**99.52**	98.12	98.56

Table 2. Detection performance on unseen manipulations with our framework and others compared on the FF++ dataset.

known DFDC dataset and our model was not fine-tuned for the known dataset evaluation. Besides, model confidence index distribution demonstrates that the model's confidence is also better calibrated without a clear loss of detection accuracy.

4.3 Cross-dataset Evaluation

In real-world scenarios, the target of the forgery detection task is often the outcome of images generated by a new model with an unknown source. Successfully detecting unseen images indicates the robustness of the model. Cross-dataset evaluation can reflect the generalization ability of the model well. Our experiments were designed to train our framework on the FF++ dataset and then test it on the Celeb-DF dataset to verify the model's generalization ability. As shown in Fig. 4 and Tab. 3, our model outperforms the state-of-the-art methods. Besides, the benefit of calibrating model confidence in the training phase has been reflected in a more reasonable confidence index distribution and a higher AUC in the cross-dataset evaluation.

Method	DFDC (AUC (%))
Capsule[37]	53.3
Multi-task[36]	53.6
HeadPose[53]	55.9
Two-stream[57]	61.4
VA-MLP[34]	61.9
VA-LogReg	66.2
MesoInception4	73.2
Meso4[1]	75.3
Xception-raw[40]	49.9
Xception-c40	69.7
Xception-c23	72.2
FWA[30]	72.7
DSP-FWA[30]	75.5
Emotion[35]	84.4
Selim[41]	**98.6**
Ours	97.9

Methods	FF++	Celeb-DF
Two-stream[57]	70.1	53.8
Multi-task[36]	76.3	54.3
HeadPose[53]	47.3	54.6
Meso4[1]	84.7	54.8
MesoInception4	83.0	53.6
VA-MLP[34]	66.4	55.0
VA-LogReg	78.0	55.1
FWA[30]	80.1	56.9
Capsule[37]	96.6	57.5
Xception-raw[40]	99.7	48.2
Xception-c23	99.7	65.3
Xception-c40	95.5	65.5
DSP-FWA[30]	93.0	64.6
F^3-Net[38]	98.1	65.2
Multi-attentional[55]	**99.8**	67.4
Two-branch[33]	93.2	73.4
Face X-ray[28]	99.2	74.8
Ours	99.1	**75.1**

Table 3. Detection performance on known dataset (DFDC, on the left) and unseen dataset (Celeb-DF, on the right) with our framework and others compared in terms of AUC (%). Our method's performance is comparable to that of the best model Selim result[48] in the DFDC competition and obtains the state-of- the-art performance in the cross-dataset evaluation. Results of some other methods are cited directly from [33].

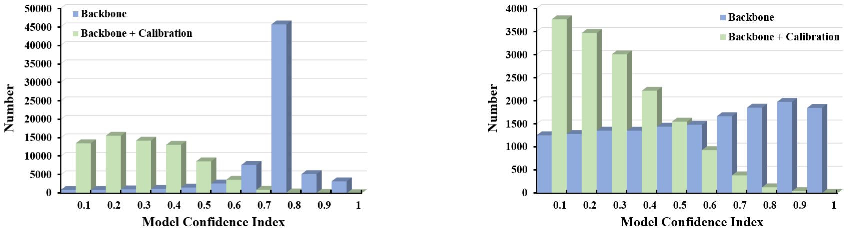

Fig. 4. Model Confidence Index(Eq. 8) on the FF++ dataset (left) and the Celeb-DF dataset (right), and we can conclude that CCL can effectively mitigate the model's overconfidence problem. On the unseen data, unlike the baseline lack of confidence, the model using CCL can confidently predict the authenticity of an image.

4.4 Ablation Study

We performed comprehensive studies on the FF++[40] and the Celeb-DF[31] dataset to validate our design of the overall framework. In summary, compared with the ablation study that removes any components, the proposed framework shows higher detection accuracy in different face forgery models and general authenticity detection. Specifically, We have analyzed the performance of each component separately from the two major aspects of DKD and CCL. The results are as shown below:

Effect of DKD. We first validate the backbone of teachers and students. For example, for the Motion Teacher, we trained five variants of our framework: 1) ResNet-50; 2) Xception; 3) EfficientNet-B7; 4) EfficientNet-B4; 5) EfficientNet-B3.

Teacher Backbone	FF++	Celeb-DF	Params
ResNet-50[16]	98.62	66.83	24M
Xception[6]	98.85	67.42	23M
EfficientNet-B7[45]	99.64	67.59	88M
EfficientNet-B4[45]	99.42	67.55	19M
EfficientNet-B3[45]	99.24	67.49	12M

Table 4. Abalation Studies for backbone variants . Frame-level AUC(%) is reported.

Tab. 4 has shown the results of different backbones. We can conclude that different backbones have little influence on the detection results under the in-dataset and cross-dataset settings. Considering the computation cost, we finally chose EfficientNet-B3 as our backbone.

To validate the effectiveness of our Feature Transformer, we performed an ablation study on the framework. We trained four layer variants and four head variants of our framework. The results are presented in Tab. 5.

Teacher Model	FF++	Celeb-DF
B3+FTL × 12	99.32	67.27
B3+FTL × 3	99.12	67.58
B3+FTL × 2	98.99	67.62
B3+FTL × 1	98.89	67.72
B3+FTL × 1+HEAD × 12	98.89	67.72
B3+FTL × 1+HEAD × 16	98.95	67.70
B3+FTL × 1+HEAD × 8	98.84	68.02
B3+FTL × 1+HEAD × 6	98.78	68.98

Table 5. Ablation study on the number of layers and heads of the Feature Transformer Encoder in our teacher framework. Frame-level AUC(%) is reported.

We notice that 1) Feature Transformer can improve the generalization capability; 2) too many layers of the standard encoder in Feature Transformer cannot further improve the teacher's performance; and 3) an appropriate number of heads can achieve the best result.

With two teachers, we automatically choose the same structure for the student. However, the student with the same structure as teachers was mediocre. We speculated that the current structure is too small to accommodate the hypothetical space of two teachers simultaneously. Thus, we adjusted the structural parameters of the student model. As shown in Tab. 6, the student with two layers of the Feature Transformer encoder and twelve heads in each layer achieves the best performance.

Effect of CCL. After verifying and carefully adjusting the structure of teachers and students, we confirmed the necessity of the various components of the calibration section and adjusted them.

As shown in Tab. 7, we firstly validated the effectiveness of dynamic confidence weights and label smoothing. The result shows that while the model generalization ability is improved, the in-dataset evaluation is almost unaffected.

Student Model (Model Only)	FF++	Celeb-DF
B3+FTL × 1+HEAD × 6	98.97	68.78
B3+FTL × 1+HEAD × 8	98.92	69.84
B3+FTL × 1+HEAD × 10	98.88	69.82
B3+FTL × 1+HEAD × 12	98.94	70.68
B3+FTL × 1+HEAD × 14	99.01	70.32
B3+FTL × 2+HEAD × 12	99.13	71.62
B3+FTL × 4+HEAD × 12	99.17	68.73

Table 6. Ablation study on the number of heads in each layer of the Feature Transformer Encoder in our student framework. Frame-level AUC(%) is reported.

Calibration Components			Datasets	
Backbone	DCW	LSR	FF++	Celeb-DF
√			99.13	71.62
√	√		99.17	73.66
√		√	99.20	72.18
√	√	√	99.19	75.12

Table 7. Ablation study of calibration components in our framework, Frame-level AUC(%) is reported.

Secondly, we adjusted the proportion of each part in the loss function, namely the weight hyperparameters λ_t and λ_l. As shown in Tab. 8, the performance increases as the proportion of the two calibration components increase. However, after their ratio reaches a specific value, the performance decreases again. There are two possible reasons. On the one hand, we speculate that the excessively high proportion of label smoothing causes the difference between targets to become too small. On the other hand, over-proportioned dynamic confidence weights will cause students to learn in the wrong direction when both teachers make mistakes.

λ_t	λ_l	FF++	Celeb-DF
0.1	0.1	98.84	70.98
0.2	0.1	98.95	71.72
0.15	0.1	99.04	72.09
0.15	0.2	99.17	75.12
0.15	0.4	99.25	73.47

Table 8. Ablation study on different hyper-parameters λ_t and λ_l in our framework. Frame-level AUC(%) is reported.

Due to the limited space, more ablation studies are available in supplementary materials.

5 Conclusions

In this paper, we have presented the CDC framework for improved generalization on face forgery detection. The proposed framework consists of three components: contrastive representation learning, dual-teacher distillation, and confidence calibration. Instead of treating forgery detection as a simple binary classification task, we calibrate the labels and model confidence to refine the targets. Through extensive experiments with contrastive representation distillation and confidence calibration, we demonstrated the superiority of our method compared with the state-of-the-art method in terms of AUC.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: WIFS (2018)
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: NIPS (2014)
3. Busira, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: ACM KDD (2006)
4. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information (2020)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: CVPR (2017)
7. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: Retinaface: Single-stage dense face localisation in the wild. arXiv preprint arXiv:1905.00641 (2019)
8. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., Zhang, C.: Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. NIPS (2020)
11. FaceSwapDevs: Deepfakes (2019), `https://github.com/deepfakes/faceswap` Accessed Novemember 7, 2021
12. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML (2016)
13. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
14. Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., Xu, C.: Distilling object detectors via decoupled features. In: CVPR (2021)
15. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR (2021)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
17. He, R., Hu, B.G., Yuan, X.T.: Robust discriminant analysis based on nonparametric maximum entropy. In: Asian Conference on Machine Learning. pp. 120–134. Springer (2009)
18. He, R., Hu, B., Yuan, X., Zheng, W.S.: Principal component analysis based on non-parametric maximum entropy. Neurocomputing **73**(10-12), 1840–1852 (2010)
19. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: CVPR (2019)
20. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: ICCV (2019)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Workshop (2015)
22. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: CVPR (2020)

23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Komodakis, N., Zagoruyko, S.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
25. Kowalski, M.: Faceswap (2018), `https://github.com/MarekKowalski/FaceSwap` Accessed Novemember 7, 2021
26. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS (2017)
27. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: CVPR (2020)
28. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR (2020)
29. Li, Y., Chang, M.C., Lyu, S.: In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In: WIFS Workshop (2018)
30. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. In: CVPR Workshops (2019)
31. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: CVPR (2020)
32. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
33. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: ECCV (2020)
34. Matern, F., Riess, C., Stamminger, M.: Exploiting visual artifacts to expose deepfakes and face manipulations. In: WACV Workshops (2019)
35. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emotions don't lie: An audio-visual deepfake detection method using affective cues. In: ACM MM (2020)
36. Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I.: Multi-task learning for detecting and segmenting manipulated facial images and videos. In: BTAS (2019)
37. Nguyen, H.H., Yamagishi, J., Echizen, I.: Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP (2019)
38. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV (2020)
39. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: ICLR (2015)
40. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV (2019)
41. Seferbekov, S.: `https://github.com/selimsef/dfdc\_deepfake\_challenge` (2020)
42. Srinivas, S., Fleuret, F.: Knowledge transfer with jacobian matching. In: ICML (2018)
43. Stutz, D., Hein, M., Schiele, B.: Confidence-calibrated adversarial training: Generalizing to unseen attacks. In: ICML (2020)
44. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
45. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
46. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. In: ECCV (2020)

47. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM TOG (2019)
48. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR (2016)
49. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: NIPS (2019)
50. Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D.F., Chao, L.S.: Learning deep transformer models for machine translation. arXiv preprint arXiv:1906.01787 (2019)
51. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. Pattern Recognition (2018)
52. Wightman, R.: Pytorch image models. `https://github.com/rwightman/pytorch-image-models` (2019)
53. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP (2019)
54. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: CVPR (2017)
55. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: CVPR (2021)
56. Zheng, Y., Bao, J., Chen, D., Zeng, M., Wen, F.: Exploring temporal coherence for more general video face forgery detection. In: ICCV (2021)
57. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: CVPRW (2017)