Appendix

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

1 Ablation Study

In the submission, we promise that the remaining ablation studies will be included in the supplementary material. Here they are.

Firstly, we evaluated the influence of the feature split of the EfficientNet-B3[4]. We got several types of feature split in the pyramidal features, such as $1536 \times 7 \times 7$, $384 \times 7 \times 7$, and $136 \times 14 \times 14$. As shown in Tab.1, different splits have some influence on the performance of our framework to some extend.

Feature Split	FF++	Celeb-DF
$1536 \times 7 \times 7$	98.78	68.98
$384 \times 7 \times 7$	98.22	67.44
$136 \times 14 \times 14$	97.95	67.15

Table 1. Ablation study of feature splits in our teachers' training. Frame-level AUC(%) is reported.

Then, we tried to design our backbone with dual-feature splits and a cross-attention mechanism. We exploited the CrossViT framework and two feature splits: $384 \times 7 \times 7$ and $136 \times 14 \times 14$. As shown in Tab.2, compared with the ViT backbone in our submission, the CrossViT backbone performed worse. We speculate that the reason for this phenomenon is redundant feature representation for a single sample.

Finally, we proposed to evaluate the influence of label smoothing. As shown in Tab.3, label smoothing can improve the performance of teachers. It seems like a good incident for the student. Generally speaking, better teachers can distill a better student.

However, as shown in Tab.4 and 5, the performance of students guided by teachers who trained with label smoothing was worse than that of teachers who did not train with label smoothing, both on the FF++ and the Celeb-DF.

Backbone	FF++	Celeb-DF
CrossViT	99.02	72.38
ViT(Ours)	99.19	75.12

Table 2. Ablation study of feature splits and backbone in the distillation process. Frame-level AUC(%) is reported.

Teacher	FF++	Celeb-DF
w Label Smoothing	99.04	69.12
w/o Label smoothing	98.78	68.98

Table 3. Ablation study of label smoothing in our teachers' training. Frame-level AUC(%) is reported.

We conclude that teachers should be confident, even overconfident, in their own domain. Teachers have a clear judgment on each sample. Only in this way can teachers‘ knowledge be credible to the student. Otherwise, teachers with less confidence will not effectively calibrate the students' prediction and confidence through the dynamic confidence weights during the distillation process. The teachers' confidence weights will be too small to recognize the difference between diverse samples.

2 Implement Setup

We used the Pytorch[2] toolkit to implement our framework. In addition to the settings indicated in the submission, the other experimental details are as follows: 1) We apply a warm-up strategy for training. Concretely, the learning rate first increases in the first 10 epochs from 0.0001 to 0.001, then cosinely decayed to 0 for the last 90 epochs. 2) The MLP size and hidden size of the Feature Transformer Encoder are set to 2048 and 49, respectively.

3 Algorithms

Firstly, we present the teachers' training algorithm, taking the ID-Teacher as an example. Then, we present the distillation process. Algorithm is shown in 2.

4 Broader Impact

Our framework aims to detect images if they are forged. The datasets used in our work are public, so there are no potential privacy risks and ethical issues. Besides, the forgery detection task aims to prevent social security issues caused by forged media content. It is a completely positive research direction, and there is no possibility of endangering public safety.

Teacher	Stu. LSR	Stu. w/o LSR
w Label Smoothing	98.87	98.65
w/o Label smoothing	99.19	99.13

Table 4. Ablation study of label smoothing in our distillation process, compared on the FF++[3] dataset. Frame-level AUC(%) is reported.

Teacher	Stu. LSR	Stu. w/o LSR
w Label Smoothing	71.45	70.97
w/o Label smoothing	75.12	74.55

Table 5. Ablation study of label smoothing in our distillation process, compared on the Celeb-DF[1] dataset. Frame-level AUC(%) is reported.

5 QA

1.some comparisons are present in the right of tab. 3 (Celeb-DF), are not present in the left (DFDC) ?

In the Celeb-DF settings, some methods did not provide the in-dataset test results of DFDC, and we carefully decided not to use our replay results to ensure the fairness of the experimental results.

2.Why measure the distance between y and its complement in equation 8 ?

The y value is the probability that the model considers the current sample to be a real face. Correspondingly, 1-y is the probability that the model considers the current sample to be a fake face. The distance between the two above is the model's confidence in its own judgment.

3.There is no explanation on which subsets from FF++ go into Face Swapping Dataset and which go into Face Reenactment dataset ?

The Deepfakes and the Faceswap go into Face Swapping Dataset. The NeuralTexture and the Face2Face go into Face Reenactment dataset.

4.How to get the result in the left of tab. 3 (DFDC) ?

We have get the dual-teacher model from the pretrained process on the FF++ dataset. Then, the teacher models above will guide the untrained student model to do in-dataset experiments on the DFDC.

5.Why do you even need a transformer?May be using a simple EfficientNet would suffice. Why create such complex models with little reasoning behind ?

Actually, we tried the model which only uses a simple EfficientNet. But the model was performed worse than adding a transformer behind the EfficientNet. Theoretically, using convolutional models and transformers at the same time can better take into account high-order semantic features and low-order texture features, so we design such a backbone structure.

Algorithm 1: Teacher-training

Data: one type of forged images I and corresponding label y, EfficientNet-b3 model E, Feature Transformer model T
Result: prediction score $\hat{y}$

for *i in I* **do**
  $F_{H\times W\times C} = E(i)$;
  $F_{H\times W\times C} \rightarrow f^{k}_{H\times W} \quad (k \in 1, 2, ..., C)$;
  $\hat{y}_i = T(f^{k}_{H\times W}) \quad (k \in 1, 2, ..., C)$;
  $Loss_i = CrossEntropyLoss(\hat{y}_i, y_i)$;
end

Algorithm 2: Distillation with Calibration

Data: forged samples I and corresponding label y, ID-Teacher T_{id}, Motion-Teacher T_{motion}, Student S
Result: prediction score $\hat{y}$

for *i in I* **do**
  $y_i^{id} = T_{id}(i)$;
  $y_i^{motion} = T_{motion}(i)$;
  $\lambda_i^{id} = 2y_i^{id} - 1$;
  $\lambda_i^{motion} = 2y_i^{motion} - 1$;
  $\hat{y}_i = S(i)$;
  $Loss_i^{hard} = CrossEntropyLoss(\hat{y}_i, y_i)$;
  $Loss_i^{id} = KLLoss(\hat{y}_i, y_i^{id})$;
  $Loss_i^{motion} = KLLoss(\hat{y}_i, y_i^{motion})$;
  $Loss_i^{conf} = \lambda_i^{id} Loss_i^{id} + \lambda_i^{motion} Loss_i^{motion}$;
  $Loss_i^{smooth} = LabelSmoothing(\hat{y}_i, y_i)$;
  $Loss_i = \lambda_t Loss_i^{hard} + (1 - \lambda_t) Loss_i^{conf} + \lambda_l Loss_i^{smooth}$
end

References

1. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: CVPR (2020)
2. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NIPS (2019)
3. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV (2019)
4. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)